Instituto Universitário de Lisboa

Department of Information Science and Technology

Unfolding the influencing factors and dynamics of overall hotel scores

Miguel Tavares Botelho

Dissertation submitted as partial fulfillment of the requirements for the degree of

Master in Computer Science and Business Management

Co-Supervisor

Dr. Rui Lopes, Assistant Professor

ISCTE-IUL

Co-Supervisor

Dr. Sérgio Moro, Assistant Professor

ISCTE-IUL

August 2019

"Ler é massada, estudar é nada"

Fernando Pessoa

Acknowledgements

This thesis was only possible to conclude with the help of several people to whom I express my sincere thanks.

First of all I would like to thank my supervisors, Professor Rui Lopes and Professor Sérgio Moro, for receiving me as your supervisee, for all the excitement in all our meetings that motivated me, for all the constructive feedback, that made me learn a lot and for the support that made me continue the elaboration of the thesis.

Also, I would like to thank all family and friends who have kept me focused and with the thesis always on my mind.

To Ana who listened to all my complaints and always tried to cheer me up.

To Kaka for being my life coach and making this process of making the thesis much lighter.

I would like to thank my parents. To my mother, not only for the thesis but for telling me off and making me to toe the line. And to my father, also for telling me off and for being better than Google. Thanks for all the inspiration and for everything.

Abstract

The hospitality and tourism industry was boosted by the help of hotel review sites, which consists in an increasing demand on the part of tourists. We extracted more than thirty thousand reviews from Tripadvisor to understand the variations in customers' perceptions of high/low end and chain/independent hotels and on which aspects this variation is most evident.

We used sentiment analysis to assign a score to the aspects of each review. We compared machine learning algorithms, namely, random forest, decision tree and decision tree with adaBoost, to predict the overall score. Then, we used the Gini index to understand the aspects that most influence the overall score.

Finally, we compared the reviews with temporal windows overtime with Jaccard index to characterize the dynamics of customer satisfaction focusing on three aspects: "Service", "Location" and "Sleep". Correlating the responses of the hotel to the users' reviews, we wanted to demonstrate the impact in the customers' perception of the hotel quality.

The best performances were achieved by the decision trees which indicated that "Service" is the most influential aspect for satisfaction, while "Location" and "Sleep" were the aspects considered less important. By identifying the moments of drastic changes, we verified that "Service" is also the most related to the overall score.

These analyses allow hotel management to track the trends of tourists' assessment in each category. Generally speaking, a focus on the "Service" should be done. However, an analysis, for a particular hotel, of the dynamics of the overall score to compare with its category would be advantageous.

Keywords: Hotel customers' satisfaction, Tripadvisor reviews, Data mining, web scraping, sentiment analysis.

Resumo

A indústria da hospitalidade e turismo foi impulsionada pela ajuda de sites de avaliações de hotéis, que leva a uma exigencia crescente por parte dos turistas. Extraímos mais de trinta mil avaliações do Tripadvisor para entender as variações nas percepções dos clientes de hotéis de alta/baixa gama e cadeia/independentes e quais os aspectos essa variação é mais evidente.

Usámos *sentiment analysis* para atribuir uma pontuação aos aspectos de cada avaliação. Comparámos algoritmos de aprendizagem automática, nomeadamente, *random forest, decision tree* e *decision tree with adaBoost*, para prever a pontuação geral. Depois, usámos o índice de Gini para entender os aspectos que mais influenciam a pontuação geral.

Por fim, comparámos avaliações com as janelas temporais ao longo do tempo com o índice de Jaccard para caracterizar a dinâmica de satisfação do cliente com foco em três aspectos: "Service", "Location" e "Sleep". Ao correlacionar as respostas do hotel com as avaliações, queriamos demonstrar o impacto na percepção dos clientes sobre a qualidade dos hoteis.

Os melhores desempenhos foram alcançados pelo *decision tree* que indicou que "Service" é o aspecto mais influente para satisfação, enquanto que "Location" e "Sleep" foram os aspectos considerados menos importantes. Ao identificar os momentos de mudanças drásticas, constatámos que "Service" também é o mais relacionado à pontuação geral.

Estas análises permitem que a gestão dos hoteis acompanhe as tendências da avaliação dos turistas em cada categoria. De um modo geral, um foco no serviço deve ser feito. No entanto, uma análise, para um hotel particular, da dinâmica da pontuação geral para comparar com sua categoria seria vantajosa.

Palavras-chave: Satisfação dos clientes de hoteis, críticas do Tripadvisor, *Data mining*, *Web scraping*, *Sentiment analysis*.

Contents

List of Figures

xiv

List of Tables

Acronyms

ANN	**A**rtificial **N**eural **N**etworks
DT	**D**ecision **T**rees
DTB	**D**ecision **T**rees with ada**B**oost
ES	**E**xplicit **S**cores
eWOM	**e**letronic **W**ord **O**f **M**outh
FS	**F**eature **S**election
GDP	**G**ross **D**omestic **P**roduct
IS	**I**mplicit **S**cores
LDA	**L**atent **D**irichlet **A**llocation
RF	**R**andom **F**orest
SVM	**S**upport **V**ector **M**achine
SA	**S**entiment **A**nalysis
SMOTE	**S**yntheticent **M**inority **O**ver-sampling **TE**chnique
VADER	**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner

Chapter 1

Introduction

Tourists' demands are increasing with the ease of getting information from the Web 2.0. Nowadays, we can find any information we want on the Internet. In one hand tourists can find information of every hotel to make a better choice. In the other hand, hotels managers have a lot of information of what tourists are searching for. Consequently, a lot of information needs to be processed and summarized.

The purpose of this work is to identify the possible differences on the influencing factors and dynamics of satisfaction in the hospitality industry in four types of hotels: a cross between high-end and low-end hotels versus chain and independent hotels. Reviews from London hotels were collected from Tripadvisor in order to analyze customers satisfaction, assigning sentiment score to factors for each review and then a feature selection analysis was used to reveal the influence of each factor on the overall score. An analysis of similarity over time was used to understand the dynamics of the influencing factors in order to understand its causes and possibly contribute to improvements in similar situations.

We collected data from the hotels in London. London provides a wide range of hotels and reviews in English, consequently it is easier to analyze the comments in sentiment analysis tools. In addition, when examples of reviews are presented, it is better for readers to perceive their content. The data was pre-processed and divided into a quadrant, low-end (1, 2 stars) or high-end (4,5 stars) hotels and independent units or hotel chains. Then, in order to determine the influence of each factor on the overall score, sentiment analysis and feature selection techniques were used in the collected reviews. Finally, a time series analysis was made to study the dynamics of influencing factors of satisfaction. This work enables a better understanding of hotel costumers.

This chapter has four sections. The Context section explores the background of the hospitality and tourism industry, namely, the situation of hospitality industry in big metropolis and the growth of the Web 2.0. The Motivation section displays the reasons for studying guest satisfaction with hotels. It starts by reporting some facts regarding gross domestic product (GDP) and employment and then it addresses the satisfaction levels in London properties. General approach and contributions section explains the way we dealt with this problem and its contributions. The thesis outline section presents the structure of this thesis.

1.1 Context

Lodging supply and demand have been growing for a long time in big metropolis (Qu et al., 2002; Tsai et al., 2006; Vanhove, 2017). Furthermore, the hospitality and tourism industry has become one of the most important sectors in world economy (Vanhove, 2017). Conducting a search on Tripadvisor, there are more than 2700 lodging properties in London, more than 2000 lodging properties in Paris and more than 6000 lodging properties in Rome. This vast choice hampers tourist decision making ability because it is virtually impossible for someone to choose one among 6000. This shows the importance of automated decision support systems such as recommendation systems capable of simplifying the individual tourist choice process. On the other hand, due to the increase in global economic power, there is also a greater demand on the part of the consumer, implying that tourists are increasingly more demanding and selective on hotel quality.

Nowadays, tourists do not choose a hotel just for the price but rather for a wide range of factors such as: service quality or room quality, which allows them to make more accurate decisions. While not too long ago, we could find this set of factors, for choosing a hotel, through word-of-mouth. With the increase of technology, the Internet is increasingly a source of information, being able to easily acquire information about hotels through their websites and travel blogs like Airbnb, Tripadvisor and others. Travel blogs are currently the most used source for the choice of a property (Mauri and Minazzi, 2013). These sites contain relevant information about each hotel, either in the presentation page of each hotel, or in reviews and ratings by its guests. However, with the large number of hotels multiplied by the large number of reviews of each hotel it is impossible to make a decision based on all reviews of all hotels. Each tourist "assigns" intrinsically importance to each factor, making the chosen hotel a good decision

for him or her. From the hotel management's perspective, problems also arise with the quantity of information related to the hotel (thousands/millions of users of these sites) as well as with the capability of the implicit communication between consumers, enabled by the Internet, to influence their opinions, i.e., a review can have impact on the perception of hotel quality so it is really important to be aware of customers satisfaction. Understanding these affairs would empower hotel management with the tools to change hotel policies and services with the purpose of increasing customer satisfaction and consequently the number of customers.

1.2 Motivation

Despite of the global crisis there has been a significant recovery of the hospitality and tourism industry, with an increase on jobs and gross domestic product (GDP). The World Travel and Tourism Council reports that the total contribution (direct, indirect and induced) of travel and tourism industry is 10.4% of global GDP in 2017. In 2017, 9.9% of global employment came from this industry. Plus, 20% of the global net jobs created in last decade have been within the Travel and Tourism sector (Travel and Council, 2018b).

The United Kingdom travel and tourism industry represents 10.5% of its GDP. More specifically, 10.5% corresponds to around 237.3 billion euros. Also, in 2017 the total contribution to employment was 11.6% (around 4 million jobs). These numbers are expected to continue to rise. By 2028, the expected contribution for United Kingdom GDP is 11.63%, an increase of 11.8%. In 2017, this made the United Kingdom the fifth country in which the tourism most contributes to the GDP and the twelfth with respect to employment. However, the expected growth of 2.0% of total contribution to the GDP in 2018 ranks United Kingdom the 159th in the list of all countries. It means that there is a lot of space to grow (Travel and Council, 2018b).

London is known as one of the best destinations to visit. Several statistics put London in the top destination for 2018 (e.g. Tripadvisor, 2018a; USNews, 2018). Despite of London having one of the highest international visitor spending, it is not on the top 10 cities in terms of direct travel and tourism contribution for GDP in 2017 (Shanghai is the first with 30 billion euros while London is the 12th with 14.38 billion euros) nor in the hospitality' fastest growing cities in 2016-2017 (Travel and Council, 2018a). This highlights that United Kingdom has room to

improve the hospitality industry, more specifically, London, which calls for an improvement of tourists satisfaction, an increase of sales and a lowering of costs.

Customer satisfaction plays an important role in explaining financial performance of hotels in existing marketing and tourism research (Assaf et al., 2015), the hotels can take advantage of the reviews so as to better satisfy the tourists. Scraping the overall rating of 2710 properties in London found in Tripadvisor and despite of the average of 4.00 out of 5, customer satisfaction can be improved all over the world. Furthermore, the standard deviation of 20% leads to some dissatisfied customers that can turn into satisfied too.

The sheer number of evaluations makes it difficult for hotel managers to know how to best interpret customers reviews. In order to improve their hotels and attract more tourists, a summary of the satisfaction factors should be done. Furthermore, this satisfaction summary allows tourism sites, as Tripadvisor, to improve their ratable aspects, that is, add or remove aspects that tourists want (do not want) to rate.

This work tries to answer which factors most influence the customer's satisfaction, which will support hotel management decisions aimed at improving customer satisfaction.

1.3 General approach and contributions

One of the major technological goals of this thesis is to identify the aspects or features of hotel guest evaluations, either among the several numerical evaluation dimensions or contained in the textual reviews, that better predict customer satisfaction. The thesis flowchart is shown in figure 1.1 and we resume it in the next paragraphs.

We started by extracting data from Tripadvisor and then identifying the features of the reviews we wanted to study. We dealt with the features that Tripadvisor' reviewers can score (the explicit features) plus four features frequently studied in literature, "Food", "Guests", "Tourism" and "Decoration". Then we used sentiment analysis in the text to assign a score of the added features in order to predict the overall score through machine learning algorithms with these features.

To predict the overall score, we had to handle many issues. First, we handled the missing data issue, in which we tried several strategies to complete the missing values. Second, we compared an oversampling strategy to adjust the classes

distribution of our data set with the normal prediction. And third, the features to include issue. That is, once the sentiment analysis did not add much value to the predictions, we had to try several approaches that are described further.

Then, we used feature selection techniques to identify the ones most related with the hotel overall rating. It was concluded that the feature "Service" is the one with most importance.

The other important goal was to study the dynamic behavior of ratings in time. To this end, we applied the Jaccard index to the overall score and to three aspects, "Service", "Sleep" and "Location", and analyzed the changes over time. We also, did a study on hotel managers responses to find out whether the response ratio and the interval between the review the response are a cause of the overall score changes.

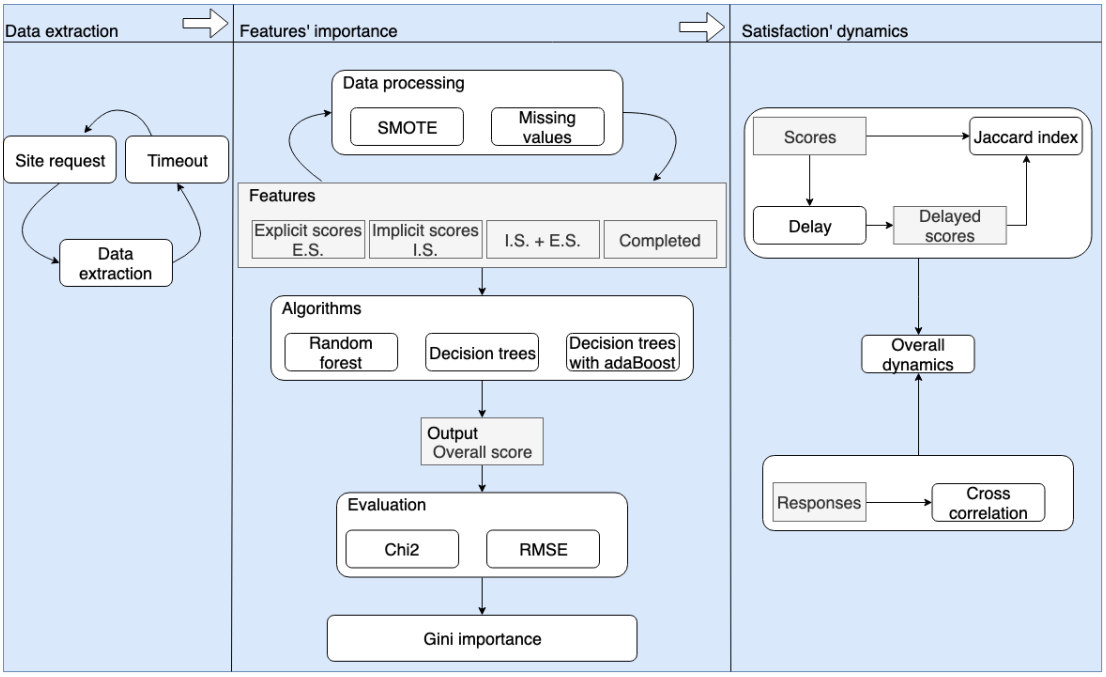

FIGURE 1.1: Thesis flowchart.

By the end of this work, it should be clear the general customer importance assigned to each aspect. Firstly, this work aims to assign a feeling score to each aspect by review and then conclude the aspects that most influence the general satisfaction of customers. Next, it is focused on the time point changes to conclude about the constraints of these changes. In order to better understand satisfaction of London tourists, the results are organized according to four categories: hotel chains vs. independent units, and low-end vs. high-end hotels. We expect to realize differences that justify this categorization.

As shown in chapter 2, several studies have been pointing out the factors that most influence satisfaction. However, none of them with this methodology. We used sentiment analysis to predict sentiment for each aspect but the studies mainly focus on how positive or negative are the aspects, they do not correlate with the overall satisfaction rating. Furthermore, one of the unique characteristics of the work presented here is the temporal aspect of satisfaction, trying to discover events that may be related to satisfaction with Jaccard index. which it is really important to keep up with the tourists' trends. From the reviewed literature, no one has attempted to conclude on the textual and quantitative aspects that most influence the satisfaction through more advanced feature selection algorithms neither using Jaccard index to identify the sentiment shifting in hotels over time.

1.4 Thesis outline

Besides this chapter, Introduction, this work has three more chapters and it is structured as follows:

Chapter 2 presents the related literature and the research gap. The literature review is divided into four parts: Online reviews in hospitality, sentiment analysis of hotel reviews, features importance and satisfaction dynamics. In the online reviews part, we aim to describe the literature of the major subjects that can be addressed to online reviews, as the causes that lead people to write reviews and understating its impact. The sentiment analysis and features importance parts aim to review what has been done in these topics to use the more suitable analysis in this work. Satisfaction dynamics part describes the literature on the way hotel guest satisfaction changes over time.

Chapter 3 starts with Data source section, in which we describe how we collected the data and some characterization of it. After that, we present a section that has a detailed description of the methods we used to analyze the date we collected. Then, we describe the two analyzes that we made to show the importance of the aspects to tourists and their dynamics. Regarding the importance of the aspects, we started by describing the sentiment analysis we made and then the features importance. Regarding the dynamics of the overall score, we describe the analysis between the overall score and three aspects, "Service", "Sleep", and "Location".

Finally, chapter 4, presents the conclusions we draw and points out possible future work.

Chapter 2

Literature Review

With the emergence of the hospitality and tourism industry, there is a growing number of research in this area. In this chapter, literature review on four topics are presented: Online reviews, whereupon it is discussed its importance; Sentiment analysis (SA) of online reviews; hotel managers' responses to reviews studies and Tripadvisor's features importance in which will be taken in consideration the already available multi-criteria ratings and text-mining ratings. Finally, the last subsection focuses on the identified research gap that this work aims to fill.

2.1 Online reviews in hospitality

Online reviews are a type of electronic word-of-mouth (eWOM), Hennig-Thurau et al. (2004) define eWOM as 'any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet'. Thus, it can be concluded that online reviews act as type of eWOM. Therefore, the literature referring to eWOM may be suitable to understand online reviews.

Concerning the impacts of eWOM, the most popular researches focus on issues regarding the pricing of hotels, companies' online reputations, interactions with online users, and the generation of customer loyalty (Loureiro and Kastenholz, 2011; Yacouel and Fleischer, 2012). Most works on these issues use text mining of the online reviews as their tool to draw conclusions.

The literature on hospitality and tourism has been growing. Mariani et al. (2018) and Cantallops and Salvi (2014) mention that there are two major subjects that can be addressed while studying online reviews and hospitality: review-generating factors (the causes that lead people to write reviews) and impacts of

eWOM (understanding the impacts of online reviews). The first aspect is not addressed in this literature review, as this thesis focuses on the second aspect only.

Concerning the second aspect, the main topics related to the impact of the feedback provided in online reviews are online buying, satisfaction and management and sentiment analysis (Schuckert et al., 2015).

Online buying focuses on three aspects: purchase intention, where it is studied the relationship between the sentiment of online reviews and hotel choice intention; customer loyalty intention; and the effects of online reviews in the hotel characteristics such as size or brand; price and sales (Schuckert et al., 2015). Several authors have reported on the importance of the feedback contained in online reviews to future item purchasing by other customers. Ye et al. (2011) concluded that online reviews are the most valuable source of information for choosing an accommodation. Likewise, Cantallops and Salvi (2014), Chen and Xie (2008) and Litvin et al. (2008) mention, somehow, that reviews are an important source for consumers' decision making. In fact, the mere availability of the reviews to the possible customers increases the likelihood of those hotels being included in the decision process (Vermeulen and Seegers, 2009). Of course, that this influence has some repercussion on sales of hotel rooms, the study Ye et al. (2009) points out that a 10% improvement in online review ratings can increase sales by 4.4%.

When it comes to satisfaction, online reviews are one of the best ways to realize the hotel quality of service so there are several studies on this subject. Some are to study the combination of words that are related to sentiment in hospitality reviews (e.g. Levy et al., 2013). Others conclude about the wide range of factors influencing the perceived satisfaction as price, amenities or online responses. In a recent study from 2017, Kim and Park (2017), claim that online review ratings are currently a more significant predictor of hotel performance than traditional customer satisfaction surveys. O'Connor (2010) studied the satisfaction among reviewers thorough text mining reviews of 100 hotels in London with an average of 75 reviews per hotel. The author made an analysis of word frequency, grouped the words by theme and compared those issues with the overall rating concluding that certain words such as "staff", "breakfast" and "clean" are more frequent in positive ratings while "dirt" or "bed" occur more often in negative ratings.

Finally, it all comes together with sentiment analysis. As the name suggests, this type of studies concern the sentiment of online reviews and feature extraction. In the next section, it is discussed several SA techniques and their possible

applications in the context of hotel reviews.

Online reviews have two main roles: provide information about the product or the service and provide recommendations either to buy the product or not, influencing consumers' decision making (Park et al., 2007). Duan et al. (2008) states that the informational role of online reviews is more often used than the recommendation role for decision making.

Due to the massive number of reviews, Hu and Liu (2004) state that summarizing online reviews is an important issue. In their study, they applied text mining to reviews, extracting product features, identifying whether the opinions are positive or not, and summarizing the reviews. Later, Titov and McDonald (2008b) proposed a framework for extracting the ratable aspects of objects from online reviews without human supervision. Concerning hospitality, Rossetti et al. (2016) analyzed online reviews with topic models in order to provide decision support to tourists based on its previous reviews. Following these studies, Calheiros et al. (2017) evaluated the sentiment of topics related to hospitality issues applying the Latent Dirichlet Allocation (LDA) model to reviews of a single hotel. The authors proposed a scalable sentiment analysis process and suggest a fully automated system for feature research.

Summarizing, feature extraction or sentiment classification are some examples of applications through text mining the reviews. Hence, sentiment analysis has an important role on those issues. Next section will present several SA techniques and its possible applications in the context of hotel reviews.

2.2 Sentiment analysis of hotel reviews

Sentiment analysis (SA) and opinion mining are broadly used as synonyms. Some authors differentiate sentiment analysis from opinion mining with respect to few details but in this thesis the two expressions will have the same meaning. Medhat et al. (2014) defines sentiment analysis as "the computational study of people's opinions, attitudes and emotions toward an entity" (p. 1093). They add that these processes are interchangeable and express a mutual meaning. Sentiment analysis of reviews can be useful for recommendation systems or for analyzing the satisfaction among costumers.

To determine the sentiment of the opinion holder about an entity, Medhat et al. (2014) present a categorization of sentiment analysis techniques. As in Figure 2.1 SA can be divided into two main approaches, machine learning and lexicon based.

In turn, machine learning approaches are also divided into two main approaches, unsupervised learning and supervised learning. In case of supervised learning, sentiment values can be taken as the polarity (negative or positive) of text, which is a classification problem, or as strength of text, which is a regression problem. Lexicon based approaches can be divided into dictionary-based approaches and corpus-based approaches. A sentiment lexicon is a list of words or phrases that express sentiment. A dictionary-based approach relies on the idea of having a small set of sentiment words and then add their synonyms and antonyms, from a dictionary. If any new word is added its synonyms and antonyms must also be added. A corpus-based approach discovers other sentiment words and their orientations from a domain corpus.

The problem of assigning a sentiment to text can be addressed at three main levels of granularity: document level, sentence level and aspect level (Liu, 2012). Due to the lack of consideration of granularity levels used by researchers, this thesis does not consider other not so relevant levels such as phrases or expressions level.

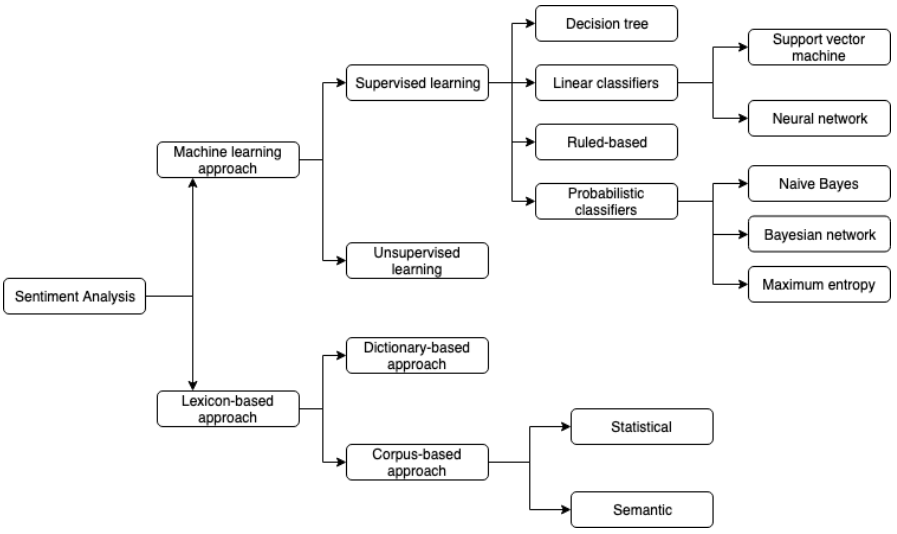

FIGURE 2.1: Sentiment analysis approaches adapted from Medhat et al. (2014).

The general purpose of document level is to classify the sentiment of the whole text. It is used when there is only one topic to be judged. At the document and sentence levels, most techniques employ supervised learning, inputting the features (generally the frequency of each word in the text) and a label (the sentiment of the text) for the training stage and then, in the test stage, the input is the features and the output is the label. Unsupervised algorithms for sentiment classification (i.e., classifying the polarity) are not used so often but there are a few examples of

clustering techniques as Xianghua et al. (2013) did in their study (Medhat et al., 2014). In the early years of research on sentiment analysis, Pang et al. (2002) tried to classify movie reviews in positive or negative through an analysis of each review with SVM, Maximum Entropy and Naive Bayes classifiers. SVM performed best with an average accuracy of 79,34%. Turney (2002) also tried to classify reviews at the document level but with an unsupervised learning approach that performed quite well too with an average accuracy of 74%. Later, Taboada et al. (2011) studied lexicon-based methods for sentiment analysis. The authors considered a lexicon dictionary to an unsupervised learning algorithm, Semantic Orientation caLculator, and achieved an overall accuracy of 78.74%. Pang and Lee (2005) and Goldberg and Zhu (2006) studied review rating prediction at the document level with the best performance of around 60% and 59.3% of accuracy on a four-class respectively. The second improved the first study by modelling rating prediction as a graph-based semi-supervised learning problem. Later, the study of Lei and Qian (2015) also predicted hotel rating reviews. They built three sentiment dictionaries: A sentiment words dictionary, a sentiment degree dictionary with 4 levels, and a negation dictionary. Then assigned values to the words of the dictionary and then evaluated the reviews through an equation. Compared their method with three methods and obtained better results in all datasets.

Sentence level analysis provides sentiment for each sentence, resulting in a more in-depth analysis. The first step is to classify whether a sentence expresses subjective information or factual (objective) information. The second is to assign a sentiment to subjective sentences. The first step is tackled with classification algorithms. In the second step, supervised learning algorithms are not so easy to use because the sentiment of each sentence (the label) is not always available for training. Hence, lexicon based and unsupervised learning approaches will be further exemplified. By classifying words in a sentence and then the sentence itself through the average of log-likelihood ratio of the classified words, Yu and Hatzivassiloglou (2003) achieved very high performance, up to 91% accuracy. Yet, most of sentence classification is through lexicon-based approaches. Examples of it are the works of Hu and Liu (2004), Kim and Hovy (2004) and Nigam and Hurst (2004). Also, Liu and Seneff (2009) proposed an approach for extracting adverb-adjective-noun phrases to predict ratings for each review based on clause structure, obtained by parsing sentences into a hierarchical representation. In addition, Kasper and Vela (2011) developed a system that classifies reviews. They divided the reviews into sentences, created a dictionary of hotel domain, detected

the topic of each sentence and then classified the divided sentences achieving an accuracy of 67% including all sentences (with neutral). This work did not use aspect (or feature) level techniques, but the authors recognized that considering the sentences with more than one topic (e.g. "I loved the breakfast and the room cleanliness") and neutral sentences (e.g. "We stayed for 3 days") would be beneficial.

Aspect level, as the name suggests, aims to classify sentiment of aspects in the text by identifying the entities and their aspects and assigning the sentiment of each aspect. To assign the sentiment of each aspect, one first needs to identify the aspects that will be considered. Therefore, it is usually necessary to extract aspects of the text. The main strategies to classify the aspects are supervised learning and lexicon-based approaches. For the supervised learning, the main approach is to use a dependency parser, which weights each feature based on the position of the feature relative to the target aspect in the parse tree (Wei and Gulla, 2010). Lexicon-based approaches first build a lexicon and then fit as input for some unsupervised algorithms. Their performance was compared with two other methods achieving better results than the others. The above-mentioned authors Hu and Liu (2004) and Kim and Hovy (2004) used lexicon based approaches applied to sentences but they could have used them with aspects as well. Snyder and Barzilay (2007) proposed the Good Grief algorithm to predict ratings for each aspect, modeling the dependencies among aspects. Wang et al. (2010) were the first authors that tried to predict latent aspect rating instead of already explicitly provided aspects in the training data. They used the lexicon-based approach to aspect extraction and predict hotel review's rating through a latent aspect model. More recently, following Taboada et al. (2011) approach to build a lexicon dictionary, Qiu et al. (2018) developed a predictive framework for calculating ratings for non-rated reviews. They extracted the aspects of the review and their contexts (term pairs) and proposed a model based on a Conditional Random Field model. Furthermore, they developed a cumulative logit model that uses aspects and their sentiments in a review to predict the ratings of the review. Their framework outperformed state-of-the-art predicting models (use SentiWordNet (Baccianella et al., 2010) to build feature vectors for reviews and employ cumulative logit model to make the prediction; an SVM-based multiclass classifier (Pang and Lee, 2005); a deep learning based model, called UWRL+ (Tang et al., 2015) and; a convolution neural network based model) (Kalchbrenner et al., 2014).

It is very hard to compare every algorithm to conclude which one is better for sentiment analysis of hotel reviews. Nonetheless there are comparisons between some machine learning algorithms. Some comparisons of relevant papers will be described here. As already mentioned, Pang et al. (2002) compared SVM and Naive Bayes and Maximum Entropy algorithms to perform sentiment analysis, the three algorithms performed quite well. Moraes et al. (2013) made a comparison between Support Vector Machines and Artificial Neural Networks (ANN) applied to document-level and concluded that ANN produce superior results to SVM's. In a recent study, the authors Singh et al. (2017) compared four algorithms: Naive Bayes, J48 (a decision tree classifier), BFTree (also a decision tree classifier) and oneR in three datasets. The authors state that the most frequent classifiers are SVM and Naive Bayes, they also concluded that these classifiers outperform others in terms of accuracy and optimization. It was concluded in their study that oneR had better results. Antonio et al. (2018) proposed a model for hotel review rating prediction, one of the features used was a lexicon dictionary and compared five algorithms (Bayesian linear regression, boosted decision tree, decision forest, linear regression and neural networks), Bayesian linear regression, decision forest and neural networks presented promising results. At the aspects level, to predict sentiment classification the main lines of research are through machine learning techniques instead of lexicon-based while joint aspect detection and SA (both, predicting ratings and classification) are through unsupervised learning methods as Latent Dirichlet Allocation (Schouten and Frasincar, 2016).

2.3 Features' importance

Given the rating of the already available aspects of Tripadvisor and textual aspects, it matters to know the influence of each one in the overall rating of the review. To do so, a literature review about feature selection was carried out. Feature selection aims to select a subset of features by ranking the feature's relevance and select them based on that score to better explain the dependent variable. In the case of hotel evaluation, the dependent variable is the reviews overall rating.

Feature selection techniques can be mainly divided into wrappers, filters, and embedded methods (Guyon and Elisseeff, 2003). Barraza et al. (2019) synthesized these three types as expressed in the table extracted from their study 2.1.

The filter methods select the features based on the correlation of each feature with the outcome variable. Since the ratings of the features of Tripadvisor are

TABLE 2.1: Synthesis of feature selection adapted from Barraza et al. (2019).

Method	Description	Examples of techniques
Filter	Variable ranking techniques as the principle criteria for variable selection by ordering	Correlation criteria Mutual information
Wrapper	Use the predictor as a black box and the predictor performance as the objective function to evaluate the variable subset	Sequential selection algorithms Heuristic search algorithms
Embedded	Reduce the computation time taken up for reclassifying different subsets which is done in wrapper methods by incorporating the feature selection as part of the training process	SVM-RFE (recursive feature elimination) Random Forest

categorical variables, examples of this type of method include the chi-square correlation, mutual information or F-score. Choi and Chu (2001) is an example of a study that uses a filter method to obtain the correlation between variables. In a total of 402 questionnaires about the overall satisfaction and attributes of the hotels, they used multiple regression analysis to investigate the influence of hotel attributes on the overall score. They concluded that "staff service quality", "Room qualities" and "Value" were the most influencing factors of the overall satisfaction level.

The strategy of wrapper methods consists on using a subset of features on a model and comparing their performance at predicting the dependent variable by removing or adding features to it. It is broadly divided into sequential methods and heuristic search algorithms. Usually, a classifier, a feature subset evaluation criterion and a searching technique are used for wrapper methods (Zarshenas and Suzuki, 2016). Genetic Algorithms or K Nearest Neighbors are examples of classifiers. Classification accuracy is an example of the evaluation criteria which is the proportion of true results among the total number of cases. In a very recent study, Mafarja et al. (2019) proposed a feature selection method using binary versions of grasshopper optimization algorithm Saremi et al. (2017), BGOA-M, benchmarked on 25 datasets, then they compared their approach with 11 feature selection algorithms. The results were quite good, achieving an average accuracy of 0.9118 and outperforming the compared methods (the most used five wrappers algorithms and six common filter-based methods (Mafarja et al., 2019)).

However, the wrapper methods have two main disadvantages, they are computationally very expensive and when the observations are insufficient the overfitting risk increases (Maldonado and Weber, 2009). To overcome these issues, embedded methods have their own built-in feature selection methods, i.e., the feature selection is part of the training process. Support vector machine recursive feature elimination, an example of embedded feature selection, behaves by eliminating the

least weighted feature in the construction of the SVM model.

Additionally, there are several hybrid approaches too. With the purpose of reducing the dimension of gene expression data Lu et al. (2017) created an algorithm based on mutual information and a genetic algorithm. They achieved its lowest classification accuracy of around 80%.
From the research elaborated on feature selection and satisfaction aspects among hotel customers, only simple filter methods were found such as chi-square for restaurant data (Wu and Liang, 2009) or Spearman correlation for hotel data (Barsky, 1992).

2.4 Satisfaction dynamics

Due to contextual changes and modifications of salient goals, one would expect that the overall opinion about a service or product or the importance one attaches to each assessment dimension would change over time. Thus, the research on the dynamics of hotel guest satisfaction and of hotel ratings should be quite extensive. However, maybe because of its specificity, there is a significant lack of research on those and related topics.

The study on the way hotel guest satisfaction and hotel ratings change with time would allow identifying revealing patterns, such as whether there is a gradual change or a sudden change or whether the cause is internal, as a service upgrade, or external, as economic changes. It is also important the nature of those temporal differences: different behaviors in different time periods, feedback shift from concrete aspects to global impression and temporal changes regarding the importance of partial ratings. Likewise, long lasting effects of satisfaction are often studied due the importance of customer loyalty.

Bjørkelund et al. (2012) made a temporal sentiment analysis and discussed the possible reasons for changes in opinion. They considered opinions about the following features: breakfast, location, staff, service, cleanliness, and Internet. They assigned sentences to the aspects and classified the reviews using SentiWordNet (Baccianella et al., 2010), a lexical resource for SA. Next, they identified changes over the average monthly sentiment score and focused on a single hotel to find out why sentiment has changed. Through the reviews, realizing that the overall satisfaction was declining, they found out that customers have experienced insects or rats in the hotel. On one hand, this sudden change is related with one-off events that occurred occasionally in a hotel and consequently the satisfaction at

that time was negative. However, there can be a gradual change of opinion, this kind of change is usually related to sloppiness of the hotel' management, i.e., when the stakeholders do not seek for a continuous improvement of the hotel, several factors as the service, are often left behind. These gradual changes are difficult to detect in a short period of time but the later they are detected the worse.

The later study was an example of an internal cause in which the authors were able to identify this shift on customer's feelings. Fukuhara et al. (2007) proposed a simple method for analyzing temporal trends of sentiments. The authors sought to find out if it was possible to discover changes in customers' feelings in external causes. They analyzed those trends, assigning sentiment to phrases from articles and news and comparing it overtime when an earthquake occurred near the hotel whose reviews were collected, concluding that it is possible to check differences of sentiments overtime. Counting the frequency of the words they concluded that sentiment of reviews was way more negative. This event is an example of an extreme external event, but would it be discovered if a mild change happened? Furthermore, would it be discovered if a gradual external change happened? The literature to answer these questions is lacking, however an increase in the national minimum wage may be considered a positive, external and gradual event that could be noticed the customer satisfaction through reviews would be increasing (or decreasing).

The nature of the temporal differences can be expressed in different behaviors in different time periods, feedback shift from concrete aspects to global impression or temporal changes regarding the importance of partial ratings.

The periods of time can be infinitely divided, one can divide into decades, years, months or even days. To analyze the temporal dynamics of hotel reviews, Wu et al. (2010) created a system, OpinionSeer, that shows the evolution of review's sentiment over time. They collected reviews from 2005 until 2010 and then identified a possible temporal opinion pattern, namely that more complaints were presented in April, May, and December. However, they did not seek a reason for this pattern.

Two temporal periods can be generally distinguished that may show differences in satisfaction and its aspects: the tourism season and that off-season. In a more recent study, Soldić Frleta and Smolčić Jurdana (2018) detected the differences in satisfaction levels during those temporal periods. The analysis of a survey of 1249 respondents, during 2016 in Opatija and Rijeka (Croatia), they revealed that tourists during the peak season expressed a statistically significant higher

level of satisfaction than those visiting the hotel in the pre- and post- season. The authors also analyzed five satisfaction dimensions (transportation and information, facilities and value for money, environment, quality and safety and hospitality) overtime. Through Games–Howell for *post hoc* testing analysis, they concluded that there is no statistically significant difference in overall satisfaction and in satisfaction dimensions except for "transportation and information", between pre- and post- season tourists. This means that "season" is a good predictor for tourism satisfaction.

It is worth mentioning that there are a few studies about the significant role of time in how customer feedback shifts from focusing on concrete details to more abstract details as time passes. On one hand, Pizzi et al. (2015) conducted a study that concludes that not only the customers tend to shift their feedback from concrete aspect to more global aspects the more they travel but also if a failure occurs, an appropriate recovery response has a positive impact on customer evaluations. On the other hand, Bernini and Cagnone (2014) did not detect differences regarding the importance given to aspects over time when evaluating a hotel. In a questionnaire they applied to tourists of Rimini, Italy, from 2004 to 2006, they used the LISREL approach (Joreskog and Sorbom, 1988) to analyze three aspects: Local environment, leisure services and accommodation. The temporal analysis was made year by year and it was concluded that the importance given to each aspect did not change. However, this analysis can be explored through other variants (as importance of aspects through shorter periods).

In addition, there are many aspects that lead to shifting satisfaction. Some circumstances may lead to a change of feeling regarding aspects and consequently to the overall satisfaction. As an example, from the above study Bjørkelund et al. (2012) if the rates are associated with the "Cleanliness" aspect they sure did decrease the overall satisfaction.

To increase positive feelings towards a hotel, these events overtime are important to detect. Early research found that there is a positive relationship between changes in customer satisfaction and changes in the performance of the hotel over time. The same research also revealed that satisfaction leads to brand loyalty overtime (Bernhardt et al., 2000; Bojanic, 1996). It would be beneficial that these changes could be automatically detected by comparing temporal periods through similarity metrics as Jaccard index (described in 3 chapter) or Salton's cosine similarity. However, no research of this topic was found in the reviewed literature.

2.5 Research gap

Much of research related to online reviews in hospitality focuses in pricing of hotels, companies' online reputations, interactions with online users, the generation of customer loyalty and more recently, many authors focus in the detection of fake reviews. All these topics are mainly analyzed with text-mining. However, research is lacking on the way the text-mining is done.

From the reviewed literature, many authors tried to predict the overall guest satisfaction (e.g. Mattila and O'Neill, 2003), associate a polarity to a sentence of a review (e.g. Kasper and Vela, 2011) or extracting the importance of the aspects to overall hotel scores (e.g. Choi and Chu, 2001). Additionally, these analyzes are often used in other contexts, such as restaurant reviews as Titov and McDonald (2008a) did by proposing a model that extracts and scores aspects from reviews. However, we did not find any example in literature that combines it all.

Additionally, we extracted data of 25 hotels in London and segmented the hotels in four categories that we did not find any literature that anyone has attempted to study this.

Research is lacking in the dynamics of the overall score. Analyzing the similarity of the overall scores with the Jaccard index allows us to detect patterns and find sudden changes. From the hotel managers' perspective, predicting these behavioral patterns and detecting sudden changes, is significantly useful for keeping up with tourist trends and consequently, let tourists more satisfied. No literature was found in this topic. However, dynamics of the management responses posted on hotel reviews is a much more researched topic. There are many research on the impact of the review responses on sales or customer satisfaction. We did not find any studies analyzing the responses time and the responses ratio.

Chapter 3

Materials, methods and results

This chapter is structured in three sections. The data source section aims to provide a detailed characterization of the data collected from Tripadvisor. The modeling and evaluation contents present a background of the methods we used. The data analysis section aims, fundamentally, to determine the features most influential on the overall score.

3.1 Data source

In order to analyze and answer the issues previously formulated, a data collection from the site Tripadvisor.com was made. In this section we start by describing the way we collected the data and then we describe the data by providing some general statistics about it.

3.1.1 Data collection

By searching on the Internet, we can find several sites with hotel reviews from where we can collect data, such as Tripadvisor, Booking, or Expedia. Tripadvisor was chosen by being a well-known site and the largest online network of travel consumers (O'Connor, 2010; Peng et al., 2018; Xiang and Gretzel, 2010). In Tripadvisor, besides the comment, a traveler can rate an overall score and a wide range of aspects about a hotel.

The data collection includes relevant information from the presentation page of the hotels followed by all the reviews from each hotel. This collection was restricted to London hotels for reasons mentioned in the previous chapters: despite of London being one of the top destinations in 2018 there is still room for improvements in tourists' satisfaction. In addition, there are more people speaking English which

makes it easier to find English reviews in London hotels and all tools of SA used in the reviewed literature are focused on English. To choose the hotels, the considered criterion was whether they fit into the created categories or not. Next, the first 25 hotels appearing in the Tripadvisor were selected trying to reach the four categories in the same number of hotels.

Although to the best of our knowledge we could not find any literature differentiating high-end hotels from low-end hotels, we have decided to consider the four and five star hotels as high-end hotels while the two and three star hotels as low-end hotels. One star hotels were not included since there is no filter to one star hotels in Tripadvisor. The distinction between a chain hotel and an independent unit is not always clear. A chain hotel makes part of a group of hotels owned by one company. Ingram (1996) affirms that a chain must be more than 2 hotels, however, it is not always possible to understand whether or not the hotels belong to a group of 3 or more hotels or not. The collected hotels and its categorization can be found in table A.1 in appendix A.

The technique used for extracting data from the site is generally known as web-scraping. The web-scraping was made through python scripts which can be mainly divided into three parts: getting the hotel name, the number of reviews and its overall score; getting the URL for each review; and getting the reviews from the URLs.

The first script collected relevant information from the presentation page of the hotels. First, the URLs of the selected hotels were manually collected and then the information from each hotel was automatically extracted by requesting the site, using the package "Requests: HTTP for Humans"(Reitz, 2018) to access it and using XPATH through the package "lxml"(LXML, 2018) to specify the location of the data. The extraction of URLs and reviews was made similarly. Once the Tripadvisor presents only five reviews per page, and given that they followed an easily identifiable pattern, it was possible to automatically generate URLs of all the five review pages. Finally, the extraction of each review was made. This extraction consumes a lot of time and there are mainly three problems in web-scraping (at least for beginners in web-scraping). The request to the site does not last forever so it is important to prepare the web-scraping program to handle the timeouts. Also, the structure of web pages is constantly changing which requires the constant change of the XPATH specifications. Finally, there is some data that is only loaded in the site when a button is clicked (as a "load more" button).

Whenever this happened, it was necessary to find alternative pages containing the whole review text.

Figure 3.1 is an example of a review extracted from Tripadvisor. In order to find out which features are most important in predicting the overall score, the scores of explicit aspects (explicit aspects are the ones that the users can score in Tripadvisor), the textual comment (or implicit aspects) and the summary were extracted. Our study regarding the most important factors conditioning the overall score was not limited to the study of the relation between the explicit aspects and the review text, on one side and the overall score, on the other side. We have also studied variations of the patterns of the client assessments, for which we needed temporal information. To find out the temporal moments of changes in the ratings and its causes, the response from hotel management to the customer review, its date and the date the review was posted were extracted. Finally, contextual attributes were extracted too: the username, user contributions, helpful votes and the URL from the review. The user does not need to classify every existing features so it is normal to have empty values in those attributes. Figure 3.2 shows what was extracted from the review of the figure 3.1 after parsing it to JSON format.

This process took some months due to the three types of problems we mentioned (request timeouts, web page structure changes and information not directly contained on the accessed page).

3.1.2 Data characterization

Once we have the data, with this extraction, 32 815 reviews were obtained from 25 hotels. From Tripadvisor, there are around 1 080 hotels and 1 423 000 reviews in London up to date (December 2018) (Tripadvisor, 2018b), which means that this sample corresponds to 2% of the hotels and reviews. Furthermore, table 3.1 shows that there are 16 897 reviews about the hotel chains and high-end category, 2 575 about the hotel chains and low-end category, 11 507 reviews about independent units and high-end and 1 836 about independent units and low-end. A total of 28 404 reviews about high-end hotels, 4 411 about low-end hotels, 19 472 about chain hotels and 13 343 about independent units. These numbers may show a significant difference between high-end hotels and low-end hotels, but it does not mean that the study cannot be done, it only means that the confidence in some results may be lower than in others.

Again, the Tripadvisor provides us the number of hotels per star rating: 143 two star hotels, 357 three star hotels, 294 four star hotels and 127 five star hotels.

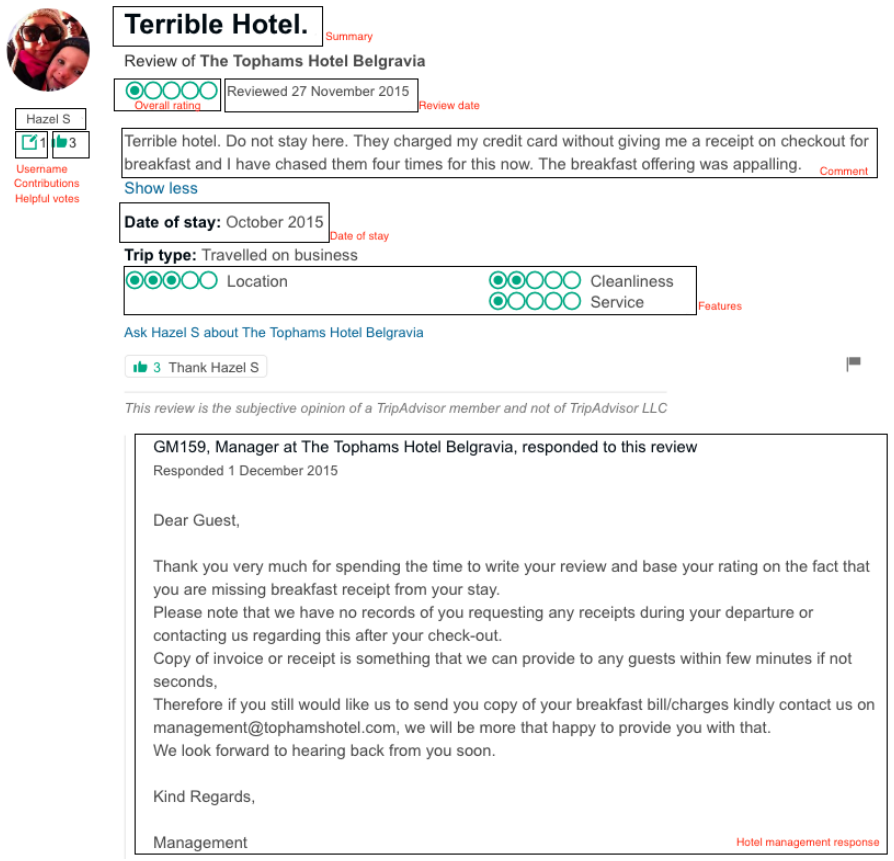

FIGURE 3.1: Example of a review from Tripadvisor.


```
{
    "username": "Hazel S",
    "rating": 1.0,
    "review_date": "Reviewed 27 November 2015 ",
    "features_score": {
        "Location" :3.0,
        "Cleanliness" :2.0,
        "Service" :1.0
    },
    "user_contributions": "1",
    "review": "Terrible hotel. Do not stay here. They charged my credit card without giving me a receipt on
checkout for breakfast and I have chased them four times for this now. The breakfast offering was appalling.",
    "review_response_date": "Responded 1 December 2015",
    "stayed": "October 2015, travelled on business",
    "summary": "Terrible Hotel.",
    "review_response": [
        "Dear Guest,",
        "Thank you very much for spending the time to write your review and base your rating on the fact that you
are missing breakfast receipt from your stay.",
        "Please note that we have no records of you requesting any receipts during your departure or contacting
us regarding this after your check-out.",
        "Copy of invoice or receipt is something that we can provide to any guests within few minutes if not
seconds,",
        "Therefore if you still would like us to send you copy of your breakfast bill/charges kindly contact us
on management@tophamshotel.com, we will be more that happy to provide you with that.",
        "We look forward to hearing back from you soon.",
        "Kind Regards,",
        "Management"
    ],
    "urlHotel": "http://tripadvisor.co.uk/ShowUserReviews-g186338-d209138-r329326397-The_Tophams_Hotel_Belgravia-
London_England.html\r",
    "user_helpful_votes": "3"
},
```


FIGURE 3.2: The corresponding JSON from the review of the figure 3.1.

This means that, not considering one star hotels, our data has 3.8% of high-end hotels and 1.8% low-end hotels in London. Unfortunately, Tripadvisor does not provide a filter to differentiate between chain hotels and independent hotels so the relative numbers of the existing chain and independent hotels in London could not be calculated. We did this distinction by hand, reading about the hotels in the

Internet.

The table 3.1 also reports the number of responses of the managers to the customer reviews and their rate. As mentioned earlier, these statistics about the responses are important for studying the customer satisfaction because hotel responses can be a possible cause of the dynamics of the scores. The table shows that low-end hotels have a significant lower response rate then high-end hotels. Of the set of all reviews of low-end chain hotels only 47.96% gave rise to hotel responses. In low-end independent hotels, only 31.04% of the reviews were responded by the hotel management. In high-end chain hotels, only 86.09% of the reviews were responded. Finally, in high-end independent hotels, only 73.26% were responded.

TABLE 3.1: Distribution of the reviews and hotels for each category.

		Hotels	Reviews	Responses	Responses Ratio
High-end	Chain	9	16 897	14 548	86.09
	Independent	7	11 507	8 430	73.26
Low-end	Chain	5	2 575	1 235	47.96
	Independent	4	1 836	570	31.04

The table A.2 in appendix A presents the number of ratings assigned to each feature by category in absolute values. To a better understating of this numbers, the table 3.2 presents the ratio between the number of reviews for each category (shown in table 3.1) and the number of reviews in which each feature was scored. This table highlights that the "Service" feature is the most scored feature while the "Sleep Quality" feature is the least used feature. Curiously, low-end chain hotels have the highest rate of scored features while high-end chain hotels have the lowest value.

TABLE 3.2: Rate of assigned features and the sum of reviews

		Service	Value	Cleanliness	Sleep	Location	Rooms
High-end	Chain	52.7	34.8	34.3	32.3	34.7	34.0
	Independent	57.0	45.0	44.2	35.8	44.1	39.7
Low-end	Chain	57.0	54.8	55.3	38.5	54.4	54.4
	Independent	52.7	44.1	43.2	27.0	42.8	40.1

To obtain predictions that are not biased, it is important to have an even distribution of data for each value of the overall score (20% from 1 to 5). However, there is no such thing as perfection and this distribution is no exception. Table 3.3 shows how overall score is distributed in the four categories and in the total of the reviews. While in low-end hotels, overall scores are better distributed, with its highest rate of 34.86%, high-end hotels are not. Most of the reviews on high-end

hotels score the hotel with 5. For example, with around 70% of reviews with five stars may imply overfitting the data and any algorithm which predicts five stars in every review would have an accuracy of around 70%. In the next section, decision trees, decision trees with adaBoost and random forest try to predict the overall score without balancing the data and next, we present a strategy that balances our data so the algorithms do not overfit.

TABLE 3.3: Distribution of overall score by category and total of reviews.

		1	2	3	4	5
High-end	Chain	1.86	2.84	8.33	26.50	60.47
	Independent	1.05	1.21	3.81	11.58	82.35
Low-end	Chain	30.21	16.74	23.73	21.55	7.77
	Independent	34.86	15.80	18.52	19.50	11.33
Total		5.65	4.08	8.52	20.49	61.26

We also found useful to analyze the frequency of the words in the reviews. The most frequent words may be a sign of what customers think, what they expected and what they do care about. First, we removed pronouns, determinants and propositions. We also replaced the plural words with the singular ones. Some different word expressions (e.g., “check in” and “check-in”) were converted to the same word (e.g., “checkin”). The different currency symbols used around the world were also replaced with the same symbol (to $ symbol). After applying these filters, only around one million words remained. From these words, the ten most frequent words are shown in figure 3.3 and sum a total of around 120 000, 12% of the total filtered words.

Some words are expected to have high frequencies, such as, "room", "hotel" or "stay". They are the top three more frequent words in the reviews. Additionally, there are some words that we can associate to certain aspects from Tripadvisor, such as "staff", or "service" meaning that they could be added to the set of explicit features rated by the customers. For example, the words "service" and "staff" are clearly associated with the aspect "Service". By being two of the most frequent words this may mean that "Service" is one of the aspects that customers most care about.

To study the causes of the dynamics of overall score, an analysis of the number of reviews over time for each category is shown in figure 3.4. This figure aims to unfold whether the number of reviews in a time interval is relevant or not i.e., in a time interval, if there are few reviews, one review will have a large impact in the overall score while if there are a lot of reviews in that time interval, one

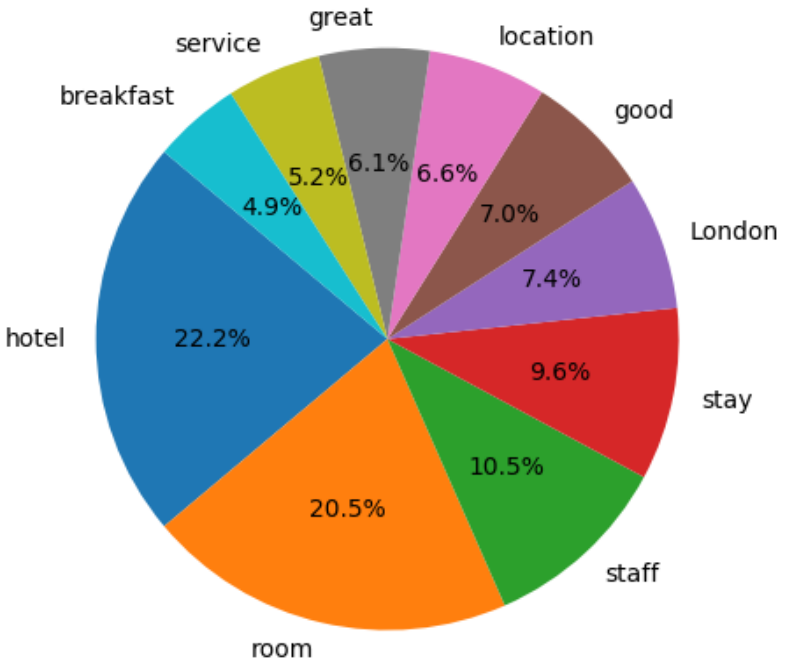

FIGURE 3.3: Words frequency in all reviews.

review will not have a relevant impact. Out of interest, figure 3.4 also shows that the number of reviews in low-end hotels, increased from 2012 to 2015, maybe due the more accentuated European crisis in that period, while the number of reviews in high-end hotels only from middle 2014 up to 2019, the number of reviews in high-end hotels increased.

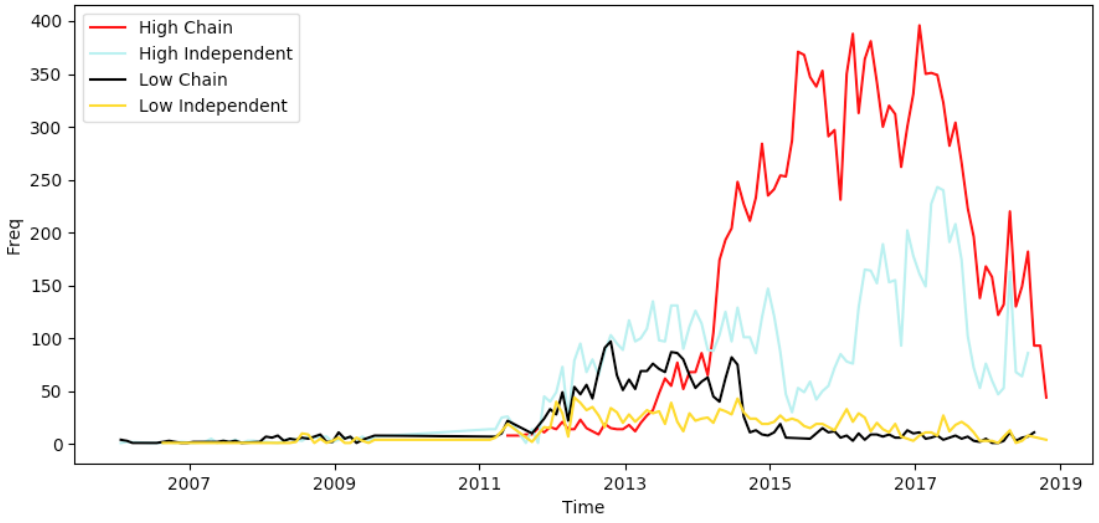

FIGURE 3.4: Number of reviews of each category by month

3.2 Modeling and evaluation contents

In this section we present a detailed description of the methods we used to analyze the data we collected. We start by describing three machine learning algorithms that we used to predict the overall score. We also describe the used metrics to evaluate these algorithms. Then we describe the similarity measure that we applied to find out the dynamics of the overall score and at last we characterize the cross-correlation functions between two signals.

3.2.1 Machine learning algorithms

When someone wants to predict a label such as the five possible scores of the Tripadvisor, it is a classification problem. On the other hand, when the label is a continuous value, it is a regression problem. However, it is possible to convert a regression problem to a classification problem in some cases. Generally, the approach to apply a supervised learning algorithm, is to extract features from the data, normalizing the values for each feature and trying with some algorithms to predict the values. To get the algorithms' performance there are several metrics we can apply. To a binary classification problem, we have well known metrics, as accuracy or f1-score. While, in a multi-class or regression problem those metrics may not be suitable, the main metrics that are often used in literature, are the root mean square error and the mean absolute error.

In this work we explored different supervised learning methods: Random forest (RF), decision trees (DT), and decision trees with the ensemble learning routine adaBoost (DTB) to find out which one best predicts the overall score. In our case, we tried DT and DTB as a regression problem to predict the multiple labels. Then, once we considered the overall score from Tripadvisor a numerical variable, we focused on Root Mean Square Error (RMSE), Chi-squared and confusion matrix to evaluate the algorithm's performance. We could have evaluated with mean absolute error (MAE) too, but it measures the average magnitude of the error as well as root mean square error so the difference between RMSE and MAE is that the RMSE gives a relatively high weight to large errors.

Decision trees

Decision trees are a non-parametric supervised learning method that predicts the value of a target variable by learning simple decision rules inferred from the data features. When fitting the training data, a decision tree splits the training data into smaller subsets, in such way that each subset is labeled as homogeneous

as possible. To train the tree is to find the tree that minimizes the impurity of the tree, that is, the average impurity of the leaf nodes. A node in which all the samples have the same label is a pure node (impurity=0).

There are different impurity measures, namely misclassification error, entropy and Gini index. The one used in this work is Gini index, which is calculated as follows:

$$i(m) = -\sum_{k=1}^{K} P(k|m)(1 - P(k|m)) \tag{3.1}$$

Where *P(k/m)* is the *a posteriori* distribution of the labels associated to each tree node m, and K is the number of classes.

When training a tree, it is computed how much each feature decreases the weighted impurity in a tree. This decrease of impurity, for each feature, is averaged for the forest, and the features are ranked according to that. The sum of all the features' scores is one, and the feature with highest score is considered the most relevant feature.

One drawback of using this algorithm to interpret the relevance of features is that if two relevant features are highly correlated with each other, as soon as one of them is considered relevant, the importance of the other decreases a lot because the impurity decrease it would cause has already been performed by the other feature.

Ensemble learning is a model that makes predictions based on different algorithms. The ensemble can be made by bagging: training a bunch of individual models in a parallel way. Each model is trained by a random subset of the data; or boosting: training a bunch of individual models in a sequential way. Each individual model learns from mistakes made by the previous model. Random forest is an example of an ensemble model using bagging while decision trees with adaBoost is an example of a boosting ensemble model (Opitz and Maclin, 1999).

AdaBoost

AdaBoost, short for "Adaptive Boosting" is an ensemble learning routine to improve algorithm's performance. It aims to convert a set of weak classifiers or regressors (classifiers/regressors with low accuracy) into a strong one (Freund and Schapire, 1997), through a linear combination of the results. The predicted value of adaBoost, *F(x)*, from M weak x is as follow:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{3.2}$$

Where h_m stands for the classification/regression produced by weak learners m and γ_m is the corresponding weight. To make things work, in training phase, adaBoost adjusts the weights and sets the trees for each regressor. We applied adaBoost regressor proposed by Drucker (1997) to predict the labels.

Random forest

A Random Forest is an ensemble of decision trees, trained with bagging, i.e., the sampling of training subsets for each tree is performed randomly with replacement. Besides, differently from the bagging trees method, the number of features considered to select the best feature is only a subset of the original set of features. For classification random forests, the number of features considered at each node is typically the squared root of the number of total features.

3.2.2 Evaluation metrics

Root Mean Square Error

The Root mean square error represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences so the lower the value is, the better. RMSE can be represented as it follows, where N stands for the size o the sample:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \tag{3.3}$$

Chi-square

Chi-square test is not a common metric to evaluate algorithm's performance. It is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories and it can be expressed in this way:

$$X^2 = \sum_{i=1}^{N} \frac{(Predicted_i - Actual_i)^2}{Predicted_i} \tag{3.4}$$

For interpretation purposes we divide chi-square by the number of samples and abbreviate it to chi2.

Confusion matrix

Confusion matrix as the name suggests, is a matrix that shows the performance of an algorithm. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class. The matrix can evaluate binary class problems and multi-class problems. In case of predicting

the output classes, ideally, the values should be located in the diagonal of the matrix. In our case, once we predicted the labels with regressors, each prediction was assigned to the most appropriated class.

3.2.3 Similarity and signal correlations

Jaccard index

Jaccard index is a method to measure similarity between two sets. It measures the intersection over union of two sets, x and y, and it is defined as:

$$J(x, y) = \frac{|x \cap y|}{|x \cup y|} \tag{3.5}$$

There are some variants of the Jaccard index that include weights, distances or probabilities. In our study we want the Jaccard index to include a difference between scores and not only a binary inclusion. So the Jaccard similarity coefficient (also known as Ruzicka similarity) is defined as:

$$J_w(x, y) = \frac{\sum_i min(x_i, y_i)}{\sum_i max(x_i, y_i)} \tag{3.6}$$

Cross-correlation

To know how much resemblance exists between two signals, we use cross-correlation. Because we have discrete values the cross-correlation z of time signals x and y is expressed as:

$$z[k] = \sum_{i=0}^{||x||-1} x_i * y_{i-k+N-1} \tag{3.7}$$

For k = 0, 1, ..., *//x// + //y//* -2

Where *//x//* is the length of x and N = *max(//x//, //y//)*

3.3 Data analysis

In this section we show the importance of several aspects of Tripadvisor to tourists. In a first analysis, to visually understand the relationship between the multiple explicit features and the overall score, figure 3.5 shows the monthly average rating of the explicit features and the overall score.

The figure 3.5 shows that the scores until 2011 are more unstable than from 2011 onwards. This is because the amount of reviews is increasing annually due

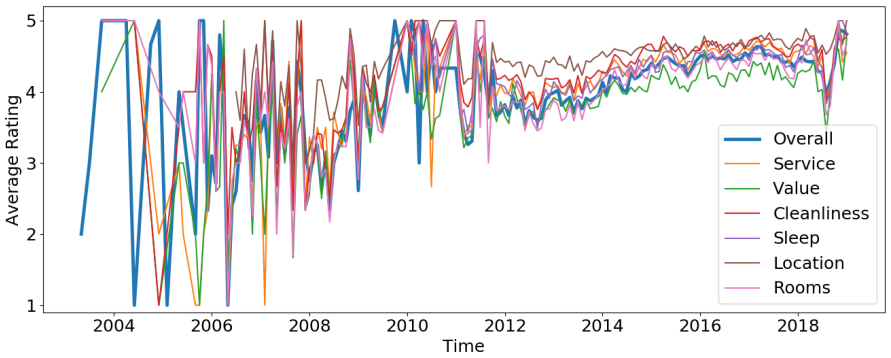

FIGURE 3.5: Monthly average rating of the aspects and overall score.

to the growth of the eWOM (Erkan and Evans, 2016). Additionally, the figure shows that, over time, the scores tend to be higher. We can also see that despite of "Location" being a bit more distant from the overall score, all aspects follow a monthly average rating similar to the overall score.

Therefore, to determine the aspects that influence most the overall score, this first study is divided in two parts: (1) assigning a score to each aspect by review and (2) computing the relation of these aspects to the overall score.

The first part aims at finding out what implicit aspects contained in the text of each review may be influencing the overall score. Beyond the ratable explicit aspects from Tripadvisor, we considered four more aspects: the "Food", "Guests", "Tourism" and "Decoration". Sentiment analysis was used to score the aspects from the text of each review. Once we have the scores from the implicit and the explicit aspects, in the second part, we compared three algorithms, random forest, decision trees and decision trees with the ensemble learning routine adaBoost to find out which one best predicts the overall score. This work concludes that service is the most influencing aspect for all categories.

3.3.1 Sentiment Analysis

From the data collection we already had the six features that users could rate. Namely “Service”, “Value”, “Cleanliness”, “Sleep”, “Location” and “Rooms”. In this first part of the study, instead of extracting aspects with Topic Modelling processes as Calheiros et al. (2017) did with Latent Dirichlet Allocation algorithm, we used aspects mentioned in the literature. We aggregated the aspects until we find a consensual set of aspects. We have performed sentiment analysis considering the six aspects already explicit in Tripadvisor and four additional ones: "Food",

"Guests", "Tourism" and "Decoration", which are often referred in the literature (e.g. Berezina et al., 2016; Xu et al., 2013).

Then we made a script that assigns scores to each of these ten aspects by review. The script reads a review and splits it into sentences through the analysis of the punctuation marks. Then, it attributes a score to each sentence using VADER (Hutto and Gilbert, 2014), a rule based method for sentiment analysis. VADER uses a lexicon of words along with their associated sentiment intensity measures. It receives a sentence and scores it in a scale from -1 to 1. VADER is pointed out to evaluate reviews for considering informal language, acronyms, some emoticons, upper case letters and punctuation (for example, it emphasizes sentences with exclamation marks). To a better comprehension and also to be able to compare them with the those explicitly made available to the customers by the Tripadvisor, we normalized the score of the sentences to the Tripadvisor scale. I.e., the obtained VADER scores were normalized from 1 to 5, where -1 corresponds to 1 and 1 corresponds to 5.

Once every sentence of the review text has a score, the script assigns aspects to each sentence. To do that, it compares whether any word in a sentence is in a lexicon of words which are associated with the respective aspects. For instance, the words associated with the feature "Guests" are "guests", "clients", "hosts", "visitors", "tourists" and "customers". If any sentence of the review has one or more of these words, the review will have the score of the sentence assigned to the feature "Guests". If there is more than one sentence with the same aspect, the mean of the scores is assigned. If a review does not contain any of the aspects, this aspect is tagged with the value "NULL". Each aspect in the lexicon has an average of ten words. It means that each word in each review will be compared to the total of 100 terms of the aspects.

The process of assigning sentiment scores to more than 30 000 reviews would be computationally expensive in terms of time, however it only took around 3 minutes to finish it. Therefore, we decided to improve our lexicon in order to reach more sentences. In the second run, we added several terms to the lexicon, including slang terms and misspelled words. The lexicon increased to a mean of 20 terms for each aspect. Some sentences, as "We stayed there for 3 nights", do not have any aspect so it is impossible to assign sentiment scores to 100% of the sentences. However, the rate improved from 63.1% to 71.1%, while the computational costs did not significantly increase.

As stated on Literature Review, chapter 2 sentiment analysis tools have some limitations. VADER is not an exception. As an example, VADER scored the sentence "Being very critical the breakfast was nearly 100 percent.", taken from a review of the Ritz hotel, with 2.24 (from 1 to 5) while for humans, this sentence is clearly near to 5. Furthermore, some sentences that have more than one aspect have a score that does not correspond to what people intend. The sentence "I did not like the service but the breakfast was fine" has a relatively null score, 3.2 from 1 to 5 score meaning that the two aspects in it, "Service" and "Food", were assigned with 3.2 and not a positive score for the "Service" and negative for the "Food". To have better results, it would be beneficial to detect irony and to have better tools for scoring reviews at the aspect level.

Given the score to the aspects in the review, the next section describes the analysis to the features that influence the overall score.

3.3.2 Features' importance

In order to know which features are influencing the overall score, first we have to know which algorithm best predicts the overall score. We compared three algorithms, random forest (RF), decision trees (DT) and decision trees with the ensemble learning routine adaBoost (DTB).

The three algorithms were tested with three different sets of features for the four categories and the total reviews. In the first test, we included the sentiment scores from the review texts, the implicit scores, and the Tripadvisor available features, the explicit scores (IS and ES). The second test considered only explicit scores (ES). Finally, in the third test, we used only implicit scores from the reviews (IS).

There is an important issue we had to deal with: the values that were not assigned by the user or that were not contained in the review text. In fact, as an example, table A.2 shows that only 33% of the users have assigned the feature "Sleep quality". We compared several approaches to overcome the missing values problem. We could delete the data that was not completely filled but it would let us without enough data so we decided, through an imputation strategy, to fill the null data using different imputation strategies. We tried five ways to complete the missing values: completing them with 0, with 6, with the average of all scores on the missing aspect, the mode of all scores on the missing aspect, and dealing with the features as if they were categorical values, that is, considering the nulls as a

categorical value and from 1 to 5 the other 5 categorical values. Encoding the six categories with one-Hot encoder from scikit-learn (Pedregosa et al., 2011).

To prevent the algorithms from overfitting, we used k-fold cross validation, namely ten-fold cross validation, a number of folds widely chosen in literature (e.g. Smola and Vishwanathan, 2008). Ten-fold cross validation splits the data into ten folds, one fold is for testing while the other nine are used for training. It iterates 10 times, each time the test set changes to a different fold.

To evaluate the performance of the algorithms we adopted two metrics, the root mean square error (RMSE) and one-way chi-square test divided by the number of samples (Chi2). RMSE is the square root of the average of the square of the difference between predicted and actual values. Chi2 is the division between the sum of the division between the square of the difference between actual and predicted values and expected values. Detailed explanations of these two metrics are presented in chapter 1.

TABLE 3.4: RMSE of the overall score prediction with different number of features.

Number of features		1	2	3	4	5	6
High-end	Chain	1.75	1.25	0.68	0.71	0.62	0.63
	Independent	0.99	0.90	0.64	0.99	0.80	0.60
Low-end	Chain	1.69	1.65	0.78	0.66	0.64	0.65
	Independent	0.63	1.64	0.82	0.72	0.70	0.60
Total		1.54	1.16	0.83	1.08	0.77	0.65

We applied these models to the four categories ("low chain", "high chain", "low independent", "high independent") and to the total of reviews. The results are shown in appendix A, tables A.3 to A.7, respectively. Curiously, the tables show that, from the features included in the models, ES performs better than IS and ES. Furthermore, IS features also does not contribute to predict the overall score. This may be justified by the amount of missing values in IS. Therefore, we did an analysis of the performance of random forest by predicting the overall score by varying the number of missing values to show the missing values impact. This impact is shown in table 3.4.

The results on table 3.4 are the result of the RMSE evaluating the random forest. It shows that a lower number of missing values leads to a better performance of the models. Hence, we added a new set of features, we called "Completed". For this new set, only the explicit features of the Tripadvisor, ES, were considered. When an ES feature is missing, its value was completed with the corresponding IS,

if it exists. If both the ES and the IS score is missing, the ES score is completed using the described strategy for missing values, replacing the nulls. For example, if the "service" is missing and the user talked about the service in the review text, the feature "Service" will have the corresponding score from the text; if the user did not talk about the service in the text, the strategy of the missing values is used.

TABLE 3.5: Best results from predicting the overall score.

		RMSE	Chi2	Algorithm	Features	Missing
High-end	Chain	0.72	0.25	DT	Completed	6
	Independent	0.61	0.18	DT	ES	6
Low-end	Chain	0.90	0.44	DT	Completed	0 or Avg
	Independent	1.03	0.55	DT	Completed	Avg
Total		0.88	0.44	DT	Completed	0

We executed sixty experiments combining the five strategies of the missing values, three algorithms and four sets of features for each category. We present the results in appendix, in tables A.3 to A.7. The best performances in each category for the regression/classification tasks are shown in table 3.5. From it, there are several conclusions that we can stress regarding the performance of the algorithms, missing values and the used features.

Decision trees performed better than the other two algorithms in all categories, with its best score of an astonishing RMSE value of 0.61 and a Chi2 value of 0.18 in high-end independent hotels. Typically, random forest and decision trees with adaBoost perform better than the decision tree in the state-of-the-art. The fact that the decision trees has the best performance is unusual. Howsoever, all three algorithms had a great performance.

Regarding the included features, "Completed" features achieved higher performances except for high-end independent hotels category that was achieved by "ES" features. The fact that the "completed" features performed better is coherent with what was expected and reveals the success of the strategy developed to complete the missing values with text information. Finally, concerning the strategies used to deal with missing values, algorithms in high-end categories performed better when the missing value is replaced with 6 while Low-end categories and the total of reviews performed better with 0 or average. These results may be due to overfitting because high-end hotels' scores are highly unbalanced towards maximum values, whereas low-end scores are significantly biased towards low to average values.

To a more detailed analysis of the error, figures 3.6 to 3.10 present the confusion matrices and the heat map of the predicted values from the algorithm with the best performance in each category.

Each row of both, the matrices and the heat maps, represents the instances in an actual class divided by the total of instances in each class while each column represents the instances in a predicted class in relative values (%). The above-mentioned figures are consistent with the values achieved with the two previous metrics and they also allowed us to unveil that the predicted values in high-end hotels are being inflated while in low-end hotels, the predicted values are better distributed.

These results may mean that the methods are overfitting the data. From the three algorithms the most robust to overfitting data are random forest and decision trees with adaBoost while decision trees is not so robust. The results show that the less robust the algorithms are, the better they predict the overall score. Furthermore, as mentioned in the data source section, if the data is not well distributed it can lead to overfitting. Table 3.3 shows that around 70% of the reviews in high-end hotels have an overall score of five stars which means the data is highly unbalanced. Therefore, next, we used a balancing method to artificially create a better distribution of the data, without distorting it. The used method is called Synthetic Minority Over-sampling Technique (SMOTE).

The table A.8 shows the results achieved with the three algorithms, replacing the missing values with 6 and the average and using the "Completed" and "ES" features. The best results from the above-mentioned table are in table 3.6.

Table 3.6: Best results from predicting overall score with SMOTE

		RMSE	Chi2	Algorithm	Features	Missing
High-end	Chain	1.08	0.66	DTB	Completed	6
	Independent	1.24	0.79	DTB	ES	Avg
Low-end	Chain	0.92	0.43	DT	Completed	Avg
	Independent	1.00	0.52	DTB	Completed	6
Total		1.09	0.65	DTB	Completed	6

We can conclude that the algorithm that performed better in most cases changed to decision trees with adaBoost, except for low chain category which did not changed. The kind of features "Completed" kept being the best in most cases except for high independent category.

In low-end hotels, the best values are better than the ones from the table 3.5. However, in general, the algorithms performed worst with the SMOTE technique.

	P=1	P=2	P=3	P=4	P=5
R=1	30.87	4.80	6.53	48.57	9.23
R=2	14.18	10.82	12.91	48.06	14.03
R=3	3.47	1.89	20.41	50.77	23.45
R=4	0.18	0.01	7.21	43.24	49.35
R=5	0.12	0.00	2.00	16.26	81.62

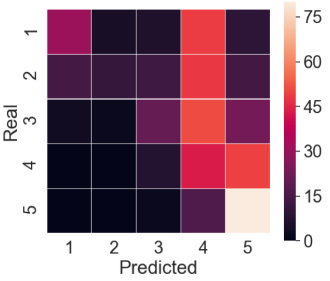

FIGURE 3.6: Confusion matrix and heat map in all reviews.

	P=1	P=2	P=3	P=4	P=5
R=1	14.01	19.11	8.28	22.29	36.31
R=2	5.83	16.88	25.21	17.92	34.17
R=3	0.92	4.40	27.70	26.99	39.99
R=4	0.04	0.09	3.89	20.57	75.41
R=5	0.08	0.03	0.29	9.55	90.05

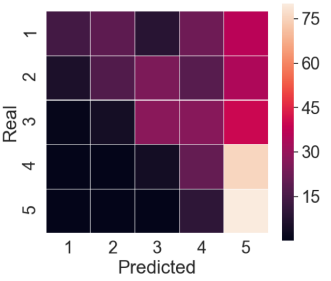

FIGURE 3.7: Confusion matrix and heat map in high chain category.

	P=1	P=2	P=3	P=4	P=5
R=1	7.44	9.92	14.05	8.26	60.33
R=2	1.44	4.32	17.99	19.42	56.83
R=3	0.46	1.37	13.70	29.22	55.25
R=4	0.08	0.00	1.73	12.90	85.30
R=5	0.06	0.12	0.56	0.88	98.39

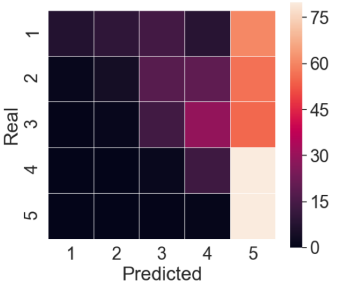

FIGURE 3.8: Confusion matrix and heat map in high independent category.

	P=1	P=2	P=3	P=4	P=5
R=1	37.02	54.88	6.30	1.80	0.00
R=2	17.87	67.05	10.44	4.64	0.00
R=3	2.13	43.54	28.48	25.53	0.33
R=4	1.44	12.97	18.38	60.72	6.49
R=5	1.00	11.50	12.50	32.50	42.50

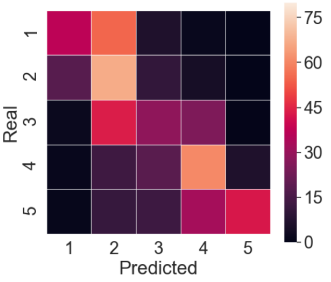

FIGURE 3.9: Confusion matrix and heat map in low chain category.

	P=1	P=2	P=3	P=4	P=5
R=1	37.66	48.91	10.78	2.50	0.16
R=2	7.93	63.45	20.69	6.55	1.38
R=3	2.06	44.71	29.41	20.00	3.82
R=4	0.00	17.60	22.35	44.13	15.92
R=5	0.48	12.98	18.27	27.88	40.38

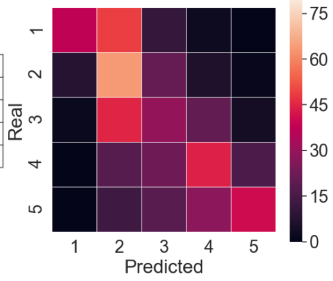

FIGURE 3.10: Confusion matrix and heat map in low independent.

Additionally, tables 3.11 to 3.15 are a more visual confirmation of that. Comparing it with those without balancing the data, when predicting the overall score in high-end hotels balancing the data with SMOTE, instead of the values being inflated, the predicted values are mostly with value "3". In low-end hotels it remains similar to the previous predictions.

We decided to discard SMOTE for three reasons: the values that are better than the ones without SMOTE are not significantly better; there is no evidence that the data is not overfitting as well and; oversampling the data increases the number of training examples, thus increasing the learning time. Consequently, for the following analyzes, we decided to use decision trees because it was the algorithm that achieved best results (see table 3.5).

Given the methods that best predict the overall score, the final step is to find out the importance of each feature. This may be especially important for hotel managers to decide upon which strategies to engage to promote their hotels. With decision trees, the importance of a feature, also known as Gini importance, is computed as the total reduction of the criterion brought by that feature, detailed in 2. Figure 3.16 shows that the "Service" is the most influencing aspect when predicting the overall score, meaning that the tourists in every category assign more importance to the service over the rest of the aspects. It should be noticed that it has the highest "assigned feature" vs number of reviews rate (A.2). "Value" would be an aspect that tourists of low-end hotels would care, however it is eight times less important than the "Service" to predict the overall score. "Cleanliness" is influencing the overall score in low-end hotels more than high-end hotels. This could be explained by high-end hotels being always clean so tourists' expectations are fulfilled, not being surprised either negatively or positively about the cleanliness quality. "Sleep" and "Location" aspects are not influencing overall score. Finally, "Rooms" do influence the overall score, but the influence is less in the two independent hotels categories.

Interpreting figure 3.16, we can highlight that "Service" is the most influencing aspect, when predicting the overall score followed by "Rooms" which is coherent with a literature as in Dolnicar and Otter (2003). This does not prove that "Service" is the aspect that is most important to tourists in general. In fact, "Service" scores could follow an inverse relationship with the overall score and still be the aspect that most influence the overall score.

Figure 3.16 also shows that Gini' importance of the different aspects from "Low chain" hotels is better distributed than the others. While Gini importance in all

	P=1	P=2	P=3	P=4	P=5
R=1	34.54	15.76	44.24	5.11	0.35
R=2	10.24	29.76	49.89	9.49	0.62
R=3	0.54	18.46	60.44	18.92	1.65
R=4	0.07	2.83	53.49	27.55	16.05
R=5	0.22	4.90	30.66	37.89	26.33

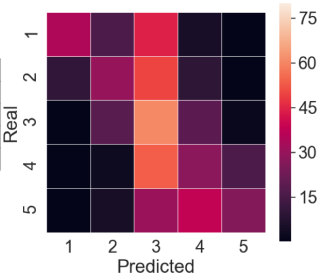

FIGURE 3.11: Confusion matrix and heat map in all reviews with SMOTE.

	P=1	P=2	P=3	P=4	P=5
R=1	26.37	17.21	53.44	2.98	0.01
R=2	11.62	36.43	44.01	7.94	0.01
R=3	1.09	17.82	60.60	19.44	1.06
R=4	0.05	2.40	42.41	40.77	14.38
R=5	0.11	2.70	34.83	19.04	43.32

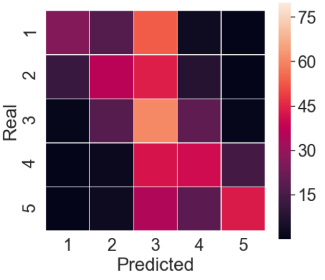

FIGURE 3.12: Confusion matrix and heat map in high chain with SMOTE.

	P=1	P=2	P=3	P=4	P=5
R=1	28.95	19.84	48.36	2.85	0.00
R=2	8.95	25.42	62.79	2.84	0.00
R=3	1.40	14.63	72.37	11.53	0.06
R=4	0.08	3.00	70.63	23.08	3.21
R=5	0.31	0.15	64.74	34.34	0.46

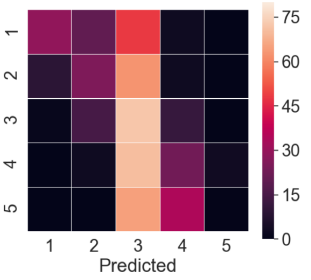

FIGURE 3.13: Confusion matrix and heat map in high independent with SMOTE.

	P=1	P=2	P=3	P=4	P=5
R=1	33.68	57.58	2.96	5.27	0.51
R=2	10.67	77.51	4.63	6.94	0.26
R=3	0.26	44.09	15.94	33.55	6.17
R=4	0.13	14.14	10.41	51.54	23.78
R=5	0.13	8.74	3.98	30.59	56.56

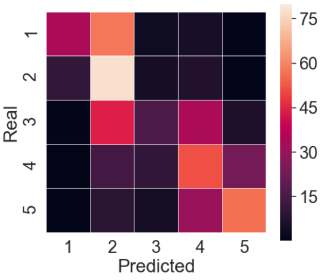

FIGURE 3.14: Confusion matrix and heat map in low chain with SMOTE.

	P=1	P=2	P=3	P=4	P=5
R=1	33.68	57.58	2.96	5.27	0.51
R=2	10.67	77.51	4.63	6.94	0.26
R=3	0.26	44.09	15.94	33.55	6.17
R=4	0.13	14.14	10.41	51.54	23.78
R=5	0.13	8.74	3.98	30.59	56.56

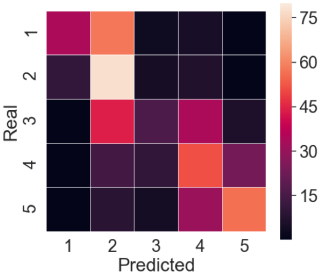

FIGURE 3.15: Confusion matrix and heat map in low independent with SMOTE.

38

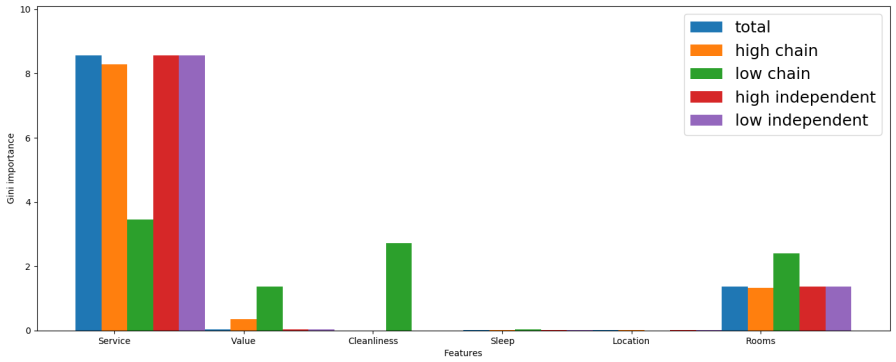

FIGURE 3.16: Features importance by category.

categories (total), "high chain", and independent hotels is around 80% in "Service" and 15% in "Rooms" meaning that the algorithms can predict the overall score mainly by "Service" and "Rooms", the Gini importance in "Low chain" hotels is around 35% in "Service", 15% in "Value", 30% in "Cleanliness" and 25% in "Rooms". In the previous section, table A.2 shows that "low chain" category have the highest rate of scored features (around 45%) which may influence the distribution of the importance of each aspect. This fact motivates the hypothesis that a better distribution of the scores leads to a more homogeneous distribution of the importance of the different aspects.

The importance of the aspects in independent hotels are very similar. Additionally, the importance of the aspects "Sleep" and "Location" did not stand out in all hotel categories. In a forethought, it may mean that tourists do not care about "Sleep" and "Location". However, it may also mean that the quality of those aspects is perceived homogeneously in all reviews.

To investigate these two hypotheses, we present the standard deviation of the aspects' ratings across the reviews for all hotel categories in table 3.7. The standard deviation gives us the amount of variation of a set. The higher the standard deviation is, the greater the dispersion of the set, which means that the perception of the aspects vary a lot from customer to customer. As we can see in the table, despite of the lower standard deviations of "Sleep" and "Location" in low-end hotels, standard deviations of the aspects are similar in all aspects. "Sleep" and "Location" standard deviation' are not outliers so we cannot conclude that the lower importance attributed by the Gini index is a consequence of their homogeneity.

Another possible hypothesis that justifies the lack of importance of those aspects is that the correlation between aspects is high. As stated in the previous section, if two relevant features are highly correlated with each other, as soon as one of them is considered relevant, the importance of the other decreases a lot. To

verify this hypothesis, tables 3.8 show the Pearson's correlation between aspects in all hotel categories. From it, we can see that "Sleep" and "Location" aspects are, indeed, correlated with other aspects, yet, it is not prominent in relation to other correlations.

TABLE 3.7: Standard deviation of the aspects in each category.

		Service	Value	Cleanliness	Sleep	Location	Rooms
High-end	Chain	0.93	1.06	0.85	0.98	0.88	0.97
	Independent	0.82	1.04	0.65	0.68	0.60	0.77
Low-end	Chain	1.17	1.05	1.02	0.88	0.82	1.11
	Independent	1.10	1.08	1.08	0.95	0.86	1.14
Total		1.97	2.13	2.29	2.23	2.29	2.07

TABLE 3.8: Pearson's correlation between aspects in every category.

		Service	Value	Cleanliness	Sleep	Location	Rooms
High-end Chain	Service	1.00	0.41	0.34	0.31	0.21	0.32
	Value	0.41	1.00	0.46	0.42	0.29	0.35
	Cleanliness	0.34	0.46	1.00	0.37	0.28	0.33
	Sleep	0.31	0.42	0.37	1.00	0.25	0.36
	Location	0.21	0.29	0.28	0.25	1.00	0.23
	Rooms	0.32	0.35	0.33	0.36	0.23	1.00
High-end Independent	Service	1.00	0.70	0.65	0.51	0.59	0.60
	Value	0.70	1.00	0.64	0.54	0.60	0.64
	Cleanliness	0.65	0.64	1.00	0.57	0.64	0.67
	Sleep	0.51	0.54	0.57	1.00	0.53	0.64
	Location	0.59	0.60	0.64	0.53	1.00	0.60
	Rooms	0.60	0.64	0.67	0.64	0.60	1.00
Low-end Chain	Service	1.00	0.65	0.61	0.44	0.46	0.58
	Value	0.65	1.00	0.64	0.45	0.43	0.56
	Cleanliness	0.61	0.64	1.00	0.43	0.47	0.60
	Sleep	0.44	0.45	0.43	1.00	0.31	0.47
	Location	0.46	0.43	0.47	0.31	1.00	0.40
	Rooms	0.58	0.56	0.60	0.47	0.40	1.00
Low-end Independent	Service	1.00	0.69	0.68	0.52	0.51	0.67
	Value	0.69	1.00	0.69	0.54	0.56	0.65
	Cleanliness	0.68	0.69	1.00	0.57	0.53	0.74
	Sleep	0.52	0.54	0.57	1.00	0.38	0.59
	Location	0.51	0.56	0.53	0.38	1.00	0.50
	Rooms	0.67	0.65	0.74	0.59	0.50	1.00
Total	Service	1.00	0.39	0.37	0.31	0.33	0.30
	Value	0.39	1.00	0.51	0.38	0.46	0.30
	Cleanliness	0.37	0.51	1.00	0.39	0.46	0.34
	Sleep	0.31	0.38	0.39	1.00	0.35	0.38
	Location	0.33	0.46	0.46	0.35	1.00	0.30
	Rooms	0.30	0.30	0.34	0.38	0.30	1.00

Generally speaking, we stress that the aspect that have more impact in a customer' review is the "Service", followed by "Rooms". In high-end hotels and "low independent" category, "Service" has a Gini importance of around 0.85 and 0.15 in "Rooms". While in "high independent" hotels, the satisfaction among users is better distributed by all aspects. In short, we could not conclude about the lack of importance of "Sleep" and "Location", of course, not all customers from hotels act the same way. Furthermore, people' likes are constantly changing overtime due several causes. For a further analysis we report the dynamics of the overall score and three aspects: the "Service", the one with most importance, "Sleep" and "Location", those with no importance according to the Gini index.

3.3.3 Dynamics of the overall score

This section presents the study of the dynamics of the overall score overtime. The study starts with a comparison of the average values by month of the overall score and three aspects. Then we applied the Jaccard index over time in order to detect changes and trends from the hotel managers and the tourists. Afterwards we studied intrinsic factors to find out the causes of these changes. The intrinsic factors we chose to study are the three aspects, "Service", "Location" and "Sleep"; and the responses of the hotel to the customers reviews. "Service" is expected to have a stronger correlation with the dynamics of the overall score because it was considered the most important feature in the previous section. "Location" and "Sleep", on the other hand, were not considered so important, therefore a cross study of these three aspects is expected to clarify the relationship between them and the overall score dynamics.

Moreover, the responses of the hotel to the customers' reviews are expected to influence the overall score as well. Intuitively, one would consider that the hotel being cooperative and friendly with their customers would influence positively the overall score.

3.3.4 Satisfaction dynamics

In order to draw more confident conclusions about the importance of the features from the figure 3.16, in this section we describe an analysis of the dynamics of the overall score parallel to the dynamics of the "Service", "Sleep" and "Location".

As mentioned previously, figure 3.16 shows a large discrepancy in the importance of the mentioned aspects. The "Service" is about ten times more relevant

than the "Sleep" and "Service". This does not mean that the sleep quality or the location of the hotel do not matter to tourists. As a matter of fact, figure 3.3 shows that "location" is one of the most frequent words in reviews.

Additionally, we did an analysis of the monthly average ratings of the overall score and the scores of "Service", "Sleep" and "Location". Despite of expecting higher averages in high-end hotels, from the Gini importance analysis, one would expect that the overall averages and "Service" averages would be more related with each other than the "Sleep" and "Location" with the overall average. The figures 3.17, 3.18, 3.19, 3.20 display the monthly average of the three aspects simultaneously with overall score by hotel category. The analysis of these figures confirms that, the three curves follow similarly the overall score curve, in such way that the difference of importance achieved by Gini index is not evidenced. To a more detailed analysis, we analyzed the relationship between the three aspects and the overall score by analyzing the distribution of each rating (1-5) of the overall score and the ratings of the three aspects, in percentage, per hotel category. The tables 3.9, 3.10, 3.11, 3.12 show the ratings distribution of the three aspects in high-end chain category, high-end independent category, low-end chain category and low-end independent category, respectively. Ideally, if an aspect has a strong importance to the dynamics of the overall score, we would expect to observe a diagonal matrix.

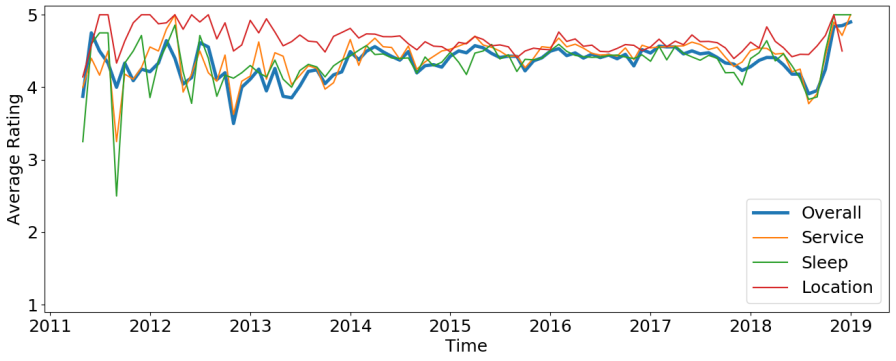

FIGURE 3.17: Overall score and the three aspects in high chain category.

The table 3.9 describes the distribution for the high-end chain hotel category, in which the "Service" has the stronger correlation with the overall score followed by "Sleep", followed by "Location". The table 3.10 shows the distribution for the high-end independent hotel category, in which the "Service" has the stronger correlation with the overall score followed by "Sleep" and then "Location". In fact, "Service" is the only aspect that seems to influence the overall score while the other two appear to have a constant rating, regardless of the overall score. This

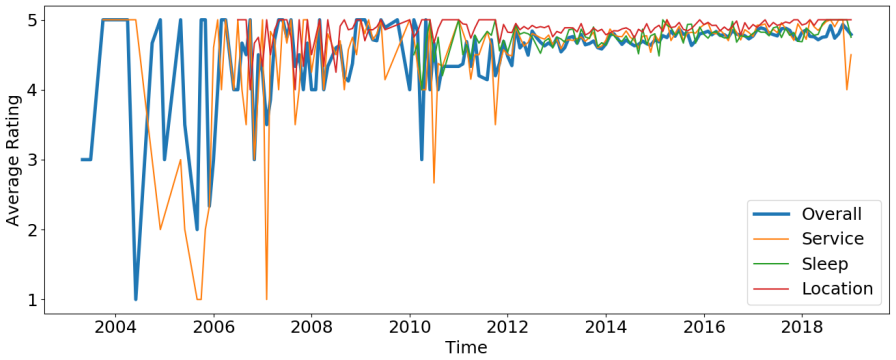

FIGURE 3.18: Overall score and the three aspects in high independent category.

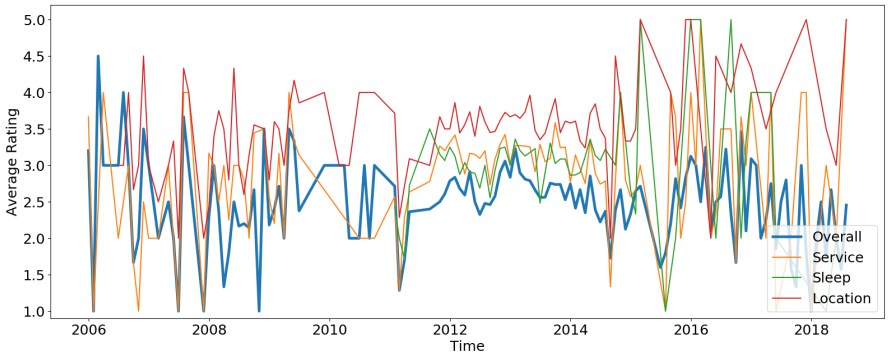

FIGURE 3.19: Overall score and the three aspects in low chain category.

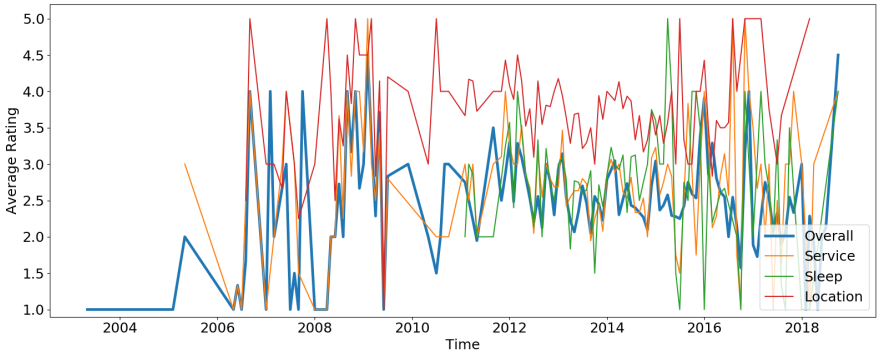

FIGURE 3.20: Overall score and the three aspects in low independent category.

is especially evident for "Location". Table 3.11 describes the distribution for the low-end chain hotel category, in which the "Sleep" has the stronger correlation with the overall score, slightly better than "Service", followed by "Location". Table 3.12 describes the distribution for the low-end independent hotel category, in which the "Service" and "Sleep" have the stronger correlation with the overall score, followed by "Location".

Generally speaking, from this analysis we can see that "Service" is the most important aspect contributing to the overall score in all categories except for low-end chain category. This is coherent with the results obtained with Gini index.

TABLE 3.9: High Chain ratings distribution

		Overall				
		1	2	3	4	5
Service	1	**59.26**	22.13	2.38	0.16	0.23
	2	19.26	**37.94**	17.09	0.61	0.06
	3	16.30	26.09	**44.90**	7.33	0.23
	4	2.96	11.07	26.89	**51.41**	5.85
	5	2.22	2.77	8.74	40.48	**93.62**
Sleep	1	**44.94**	18.24	5.03	0.13	0.20
	2	20.22	27.06	14.68	1.23	0.03
	3	24.72	**30.00**	31.03	10.01	1.33
	4	4.49	18.82	**40.67**	**54.26**	13.76
	5	5.62	5.88	8.60	34.37	**84.67**
Location	1	6.06	1.69	0.78	0.06	0.15
	2	4.04	2.26	2.52	0.55	0.03
	3	29.29	19.21	17.86	6.13	1.17
	4	**33.33**	**52.54**	**43.88**	39.15	14.12
	5	27.27	24.29	34.95	**54.11**	**84.53**

TABLE 3.10: High Independent ratings distribution

		Overall				
		1	2	3	4	5
Service	1	**50.70**	21.25	4.17	0.84	0.56
	2	11.27	**26.25**	12.12	0.84	0.33
	3	8.45	22.50	**32.95**	8.84	0.64
	4	4.23	12.5	22.35	31.70	3.76
	5	25.35	17.50	28.41	**57.78**	**94.72**
Sleep	1	**36.11**	5.41	5.10	0.97	0.13
	2	8.33	8.11	6.37	2.42	0.13
	3	13.89	24.32	24.20	6.78	1.50
	4	16.67	27.03	29.94	31.72	7.93
	5	25.00	**35.14**	**34.39**	**58.11**	**90.30**
Location	1	1.85	0.00	0.00	0.00	0.00
	2	1.85	5.00	0.48	0.00	0.03
	3	16.67	6.67	5.71	2.38	0.44
	4	29.63	33.33	29.52	24.86	5.29
	5	**50.00**	**55.00**	**64.29**	**72.76**	**94.24**

Interestingly, we found with this analysis that "Sleep" plays an important role within the low-end hotels which is not the case for the high-end hotels.

In order to study whether the "Service", "Sleep" and "Location" follow or not the dynamics of the overall score, an analysis of Jaccard index for the overall score and these three aspects was carried out. As mentioned in the previous in the Jaccard index measures the similarity between two sets, and it is defined as the size

TABLE 3.11: Low chain ratings distribution

		Overall				
		1	2	3	4	5
Service	1	**67.11**	12.50	1.66	0.00	0.78
	2	18.57	37.07	5.82	0.54	0.00
	3	11.94	**40.95**	**50.69**	16.03	0.00
	4	1.86	7.76	35.73	**56.25**	15.50
	5	0.53	1.72	6.09	27.17	**83.72**
Sleep	1	**68.04**	28.95	3.17	0.00	1.03
	2	20.09	**40.79**	13.49	0.74	0.00
	3	10.50	27.63	**46.43**	19.49	0.00
	4	0.91	2.63	32.54	**57.35**	16.49
	5	0.46	0.00	4.37	22.43	**82.47**
Location	1	24.38	5.56	0.57	0.00	0.81
	2	14.94	12.50	4.31	0.28	0.81
	3	**39.34**	**43.06**	35.34	13.88	3.23
	4	15.51	31.48	**45.11**	**46.74**	19.35
	5	5.82	7.41	14.66	39.09	**75.81**

TABLE 3.12: Low Independent ratings distribution

		Overall				
		1	2	3	4	5
Service	1	**78.44**	14.81	3.35	0.00	0.99
	2	13.75	**42.59**	19.55	1.94	0.00
	3	7.50	36.42	**50.28**	13.59	2.97
	4	0.31	4.94	22.91	**58.74**	16.83
	5	0.00	1.23	3.91	25.73	**79.21**
Sleep	1	**73.68**	22.73	3.09	0.00	0.00
	2	11.84	**43.18**	13.40	0.87	0.00
	3	11.18	25.00	**49.48**	19.13	4.65
	4	3.29	7.95	25.77	**57.39**	25.38
	5	0.00	1.14	8.25	22.61	**69.77**
Location	1	21.46	0.79	0.00	0.00	0.00
	2	13.03	12.70	4.49	1.81	0.00
	3	**34.48**	28.57	26.28	10.84	0.00
	4	20.31	**38.10**	**34.63**	33.13	10.39
	5	10.73	19.84	34.61	**54.22**	**89.61**

of the intersection divided by the size of the union of the sample sets. However, we used the weighted Jaccard similarity, also explained in the previous section, to take into account the distance between two scores. Our approach consists in comparing the reviews of a certain time interval with the evaluations of another time interval and thus analyzing the dynamic aspects of the scores over time, for example the seasonality.

The two sets being compared in each calculation of the Jaccard index must be the same size. Therefore, the choice of the sizes of the sets to be compared was an issue that deserved careful analysis. One option would be to select a fixed number of evaluations, N and compare the first N evaluations with the following N evaluations and so on, regardless of the time window that this N observations represent.

The other option would be to select a fixed size temporal window. To ensure that the number of evaluations in each time window is the same, we applied bootstrap to the samples, as proposed by Efron (1992). This method consists of adding evaluations to the window that has the smallest number of scores between two windows. This consists of replicating randomly selected scores from the window being bootstrapped (resampling with replacement).

For example, we could have, for a given hotel category, ten evaluations in April, twenty evaluations in May and sixty in June. Using the first option and selecting for example, $N = 30$, we would lose our notion of time. While using the second option, we keep the notion of time and solve the fact that the months have different number of evaluations with bootstrapping. Hence, we chose the second option.

Afterwards, to define the temporal window, we compared each months' evaluations with those of the following month, each quarter and each year. In each case, we oversampled the window with less scores, using bootstrapping, and calculated the Jaccard index of each window and the following one. The results for the monthly windows are presented in figures 3.21 to 3.24, high chain, high independent, low chain, low independent respectively. The results for the quarterly and annually windows are shown in appendix A (Figures A.1 to A.8) because the variations are more evident in shorter temporal windows. Plus, we can detect touristic seasonality better in months than in quarters or years.

Although the figures might look a bit confusing, with four lines represented simultaneously, they show that the three aspects follow the overall curve. We can also see that we can mainly separate two time intervals - 2003 to 2011 and 2012 to 2018. From 2003 to 2011, due the lack of reviews, Jaccard index has large variations, while from 2012 to 2018 it does not. Moreover, in high-end hotels, it tends to have a higher Jaccard index over time, meaning that as the time goes by, the changes in overall score tend to decrease.

To allow a clearer interpretation of the graphics represented in the four figures 3.21 to 3.24 we set a threshold of 0.5 that is also represented in the graphics with a horizontal line. The values of the Jaccard index above the threshold represent small

variations between two consecutive months. While values below the threshold evidence bigger variations which we want to analyze more carefully.

Afterwards, we calculated the times that the Jaccard index of the overall score is below the threshold. For all these occurrences we counted the number of times that the Jaccard index of each of the three aspects ("Service", "Sleep" and "Location") were also below the threshold. The table 3.13 shows the ratio between overall Jaccard index below 0.5 and the aspects Jaccard indexes of the three aspects, with values below 0.5. We expect that the aspects more relevant for the overall score to have higher ratio in the table.

In the table 3.13, high chain category does not have any results because Jaccard index of the overall score never falls below 0.5. This means that this category is more stable with no sudden changes which is what one would expect in high-end chain hotels more than in low-end hotels. On the other hand, there were more sudden changes in low-end hotels. The table 3.13 shows that "Service" follows more closely the overall score changes than the other two aspects do. In fact, more than 80% of the times that the Jaccard index of the overall score falls below the threshold, the Jaccard index of the "Service" also falls below the threshold. "Location", in its turn, also follows these changes, especially in low-end hotels. This could mean that the causes of the overall changes are related mostly to the "Service", secondly to the "Location" and at least the "Sleep". It is interesting to notice that, despite the fact that the hotels do not change their location, the perception of the customers changes overtime. This highlights the fact that all of these aspects are only subjective perceptions of the customers and may not be a result of objective measures. It is important to hotel managers to understand the perceptions of the tourists and to follow their trends.

Table 3.13: Aspects rate whenever the overall Jaccard index is lower than 0.5

		Service	Sleep	Location
High-end	Chain	-	-	-
	Independent	83.33	16.67	16.67
Low-end	Chain	87.50	28.13	75.00
	Independent	85.00	42.50	65.00

From this analysis, we can find out the seasonality by checking the comparisons between months that have lower Jaccard indexes.

We wanted to find out if there is a sudden change in the tourists' behavior in certain months by counting the times the Jaccard index is less than 0.5 in each month. The table 3.14 shows the times that each aspect or overall score in low-end

FIGURE 3.21: Jaccard index overtime in high chain category.

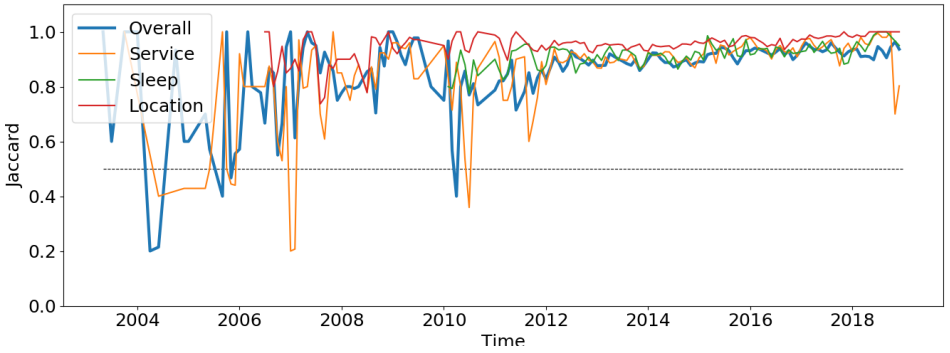

FIGURE 3.22: Jaccard index overtime in high independent category.

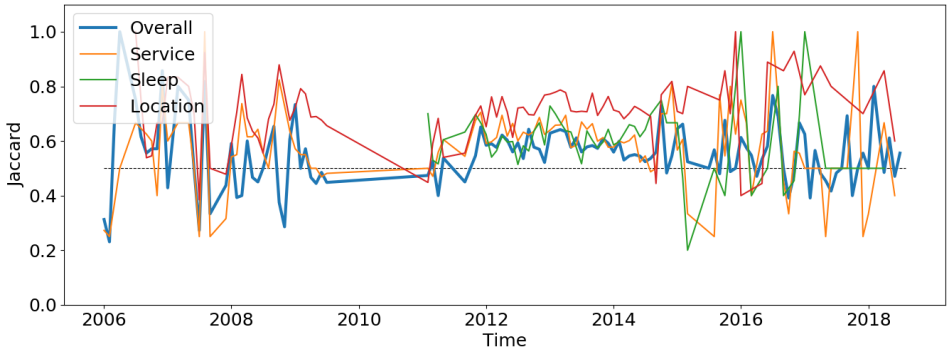

FIGURE 3.23: Jaccard index overtime in low chain category.

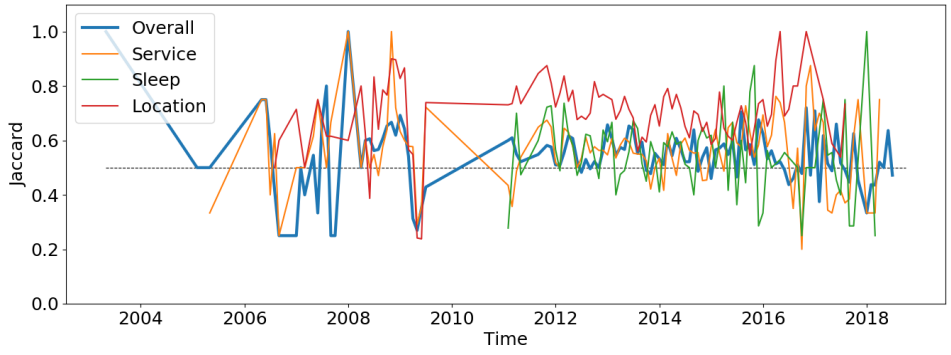

FIGURE 3.24: Jaccard index overtime in low independent category.

hotels is below 0.5. A reminder that high-end chains do not have any value below the threshold, and also, high-end independent category has few values below it so the table does not present high-end hotels. Despite that, we can see that the overall score and the three aspects are not prominent in any month. We cannot analyze any seasonality with this method, however, focusing on the eighth month, August, and in Low-end independent hotels, the Jaccard index from the overall score fell below 0.5 five times, "Service" fell four times, while "Sleep" did not had Jaccard index below 0.5 and "Location" had once. Possibly meaning that the sudden changes of the overall score are related to the "Service".

TABLE 3.14: The times in each month that Jaccard index falls below 0.5.

		1	2	3	4	5	6	7	8	9	10	11	12
	Overall	3	4	3	4	2	3	4	1	2	3	2	2
Low	Service	5	2	3	3	1	4	2	3	4	1	1	2
Chain	Sleep		1	2	1	1	2		1	1	1	1	
	Location	1	1			1		1		2			1
	Overall	3	6	2	3	4	4	3	5	4	3	3	1
Low	Service	3	2	3	2	3	3	1	4	3	2	3	1
Independent	Sleep	4	3	4	1	2	3	3		4	3	3	1
	Location			2		1	1		1				

We also studied the changes overtime in a more particular way, by applying this method to the hotel that we extracted with more reviews, "The Savoy". To this end, we applied the Jaccard index to the overall score and to the three previously mentioned aspects, "Service", "Sleep" and "Location" and compared the trends from the hotel with those of its category, high-end independent hotels. This allows the management team from "The Savoy" detecting the general behavioral patterns from their tourists, to know better their customers and their trends so they can differentiate from its category and have a competitive advantage. Figure 3.25 shows the Jaccard index of the scores of "The Savoy".

Comparing the figures 3.25 and 3.22 to find out if the trends of the high-end independent category resonate with this hotel, we can identify that the aspects in both figures follow the overall score. In most points, both figures have the four Jaccard indexes really above the defined threshold. In figure 3.25 we have four significant falls of the line representing the "Service" as well as in figure 3.22 but, unlike in figure 3.22, the overall score of "The Savoy" does not change significantly. This is an indication that the "Service" is not that related with the overall score in this hotel comparing with its category. However, both figures are significantly

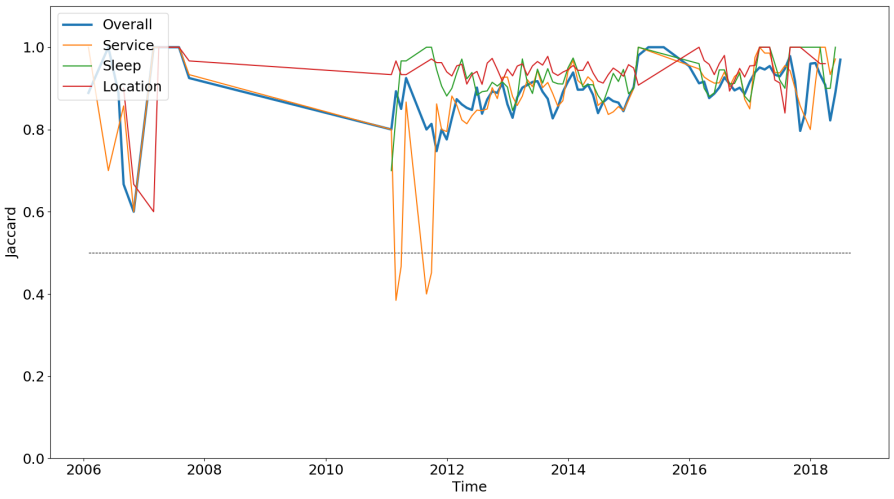

FIGURE 3.25: Jaccard index from the hotel "The Savoy".

similar which makes us to conclude that "The Savoy" hotel follows the trends from the high-end independent category.

From the hotel management' perspective it matters to know which time periods the hotel did not follow the changes from its category, analyze the several components that the hotel offers to their customers that are influencing the overall score and to have better insights of their tourists' demands.

3.3.5 Hotel mangers responses

As another dynamic aspect, we wanted to determine if the responses of the hotel management to the customer reviews had an impact in the overall scores. Namely, we wanted to find if responses ratio and the interval between the review and response have a relation with the overall score and if there is any differences between categories.

Our approach started by determining the time interval between reviews and its responses and expand this analysis to each possible overall score. Regarding the response ratio we started by determining the delay (the time lag) of the possible influence. After determining the delay, we determined the degree to which the response ratio is related to the overall score.

The responses ratios are already exposed in the previous section, table 3.1. We mentioned that there is a significant difference between high-end and low-end in terms of responses ratio. To allow a better interpretation of this table we expanded it to the responses ratio of the possible scores of the overall score and consequently we can detect the relation between these scores and the responses ratio and check if the differences of high-end and low-end are indifferent to the overall scores. The

table 3.15 shows the responses ratio by score and category. We can also see in this table that there is a significant difference between high-end and low-end hotels. Additionally, this table shows that chain hotels tend to respond to higher overall scores while independent ones have a more uniform response ratio distribution.

		1	2	3	4	5
High-end	Chain	0.74	0.68	0.72	0.83	0.91
	Independent	0.70	0.67	0.70	0.68	0.74
Low-end	Chain	0.48	0.42	0.45	0.52	0.62
	Independent	0.30	0.32	0.28	0.30	0.38

TABLE 3.15: Responses ratio by score and category.

After that, we found useful to study the time interval between the reviews and the responses to find out if the categories and the overall score have any influence on response ratio. We present this time interval in boxplots to show the distribution of the time interval in figure 3.26. The boxplots show several interesting facts. They are placed really near zero meaning that, normally, hotel's managers respond to the reviews up to 15 days. We cut the figure due the several outliers in the boxplots that we believe to be tests from the hotels managers, there were responses with 1 500 days of interval. Also, the boxplots show that the distribution of the response time when the overall score is 1 and 3 is higher in all categories except for low-end chain hotels when the response time appears to increase as the overall score increases.

Despite of 3.26 showing that the distribution of the response time is not, in general, affected by the score nor the categories, we could not conclude about the influence of it because we only have the response time to the reviews which were responded. Those that remain not answered are not included in the figure but due the lack of responses, the boxplots would be really near to zero which makes it impossible to analyze.

To obtain the lag, we used a cross-correlation between the response ratio and the average overall score. This correlation is calculated by overlapping signals, temporally shifting one relative to the other, and computing a correlation coefficient for that time shift. This process is repeated for increasingly larger positive and negative time shifts. Each time, one signal is shifted by one sample to the left (negative time lags) or to the right (positive time lags). The cross correlation of the two signals (response ratio and overall score).

It returns an array containing, in each position, the correlation coefficient of the two signals for the corresponding time shift. The correlation coefficient, for

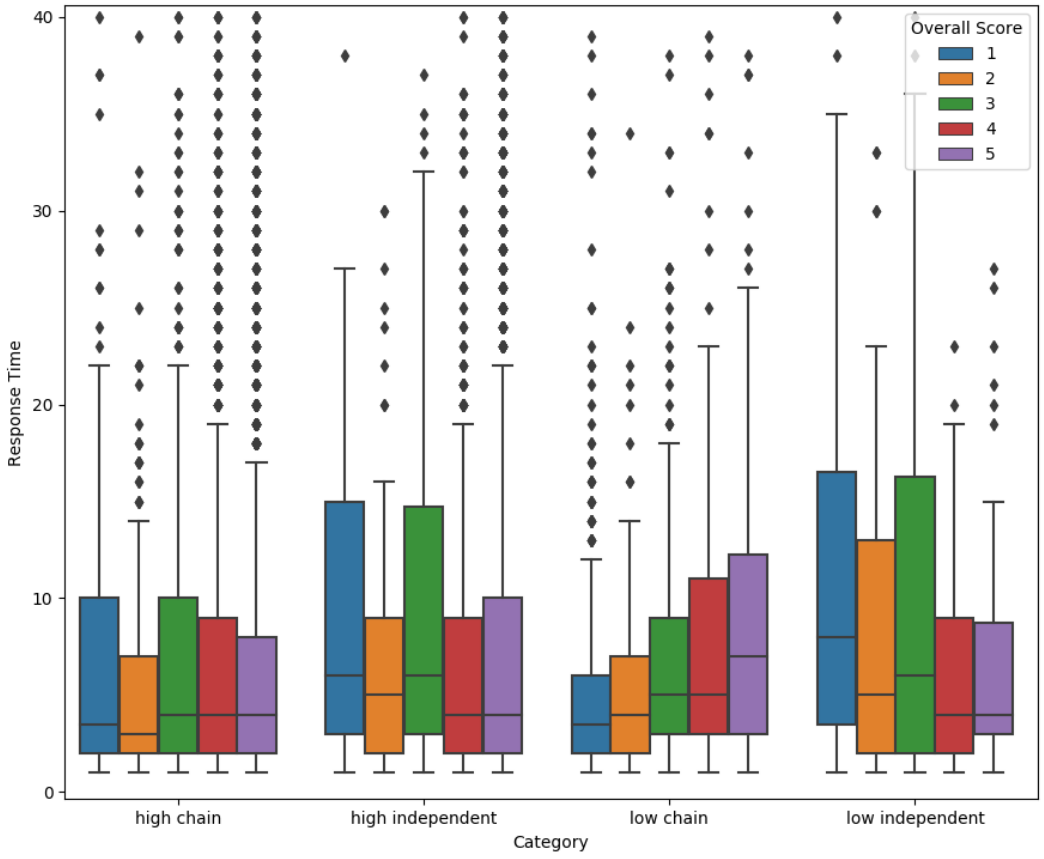

FIGURE 3.26: Boxplot of the response time by overall score and category.

a given time shift, is the sum, for all times, of the products of the overlapping samples of the two signals, one of which is shifted. That is, if two signals have a size of two, we will have three correlation coefficients in which the position that contains the highest value corresponds to the lag. In this case, the two vectors that we have, vary in size depending on the desired time interval but both have the same size.

We tried four ways of correlating signals by changing temporal periods. We did a correlation between the monthly, quarterly and annual average overall score and response rate and monthly overall score over a year. We only present in table 3.16 the last five years because the results of the previous years are the same as those of the presented years.

To determine the lag of the correlation, it is necessary to know the size of each array (or signal) of values of the response rate and the average of the overall score. Since the size of the two signals is equal, if the position of the highest value of the correlation is equal to the size of a array then we have no lag. If the position is smaller than the size, it means that the delay of the second signal is negative. That is, the response rate influences the overall score. Otherwise, the

TABLE 3.16: Correlation between responses ratio and average overall score

			Month	Quarter	Anual	2014	2015	2016	2017	2018
High-end	Chain	Signals size	94	45	9	12	12	12	12	12
		Lag	0	0	0	0	0	0	0	0
		Correlation/Variance	0.06	0.01	0.02	0.06	0.05	0.05	0.05	0.05
	Independent	Signals size	164	62	17	12	12	12	12	12
		Lag	0	0	0	0	0	0	0	0
		Correlation/Variance	0.05	0.01	0.00	0.06	0.06	0.07	0.05	0.05
Low-end	Chain	Signals size	122	45	12	12	9	12	12	8
		Lag	-3	0	0	0	0	0	0	0
		Correlation/Variance	0.14	0.03	0.14	0.13	0.18	0.36	0.46	0.60
	Independent	Signals size	121	48	14	12	12	12	11	8
		Lag	1	1	0	0	0	0	0	0
		Correlation/Variance	0.01	0.04	0.17	0.78	0.31	0.12	0.13	0.16

overall score influences the response rate. The table 3.16 summarizes the values of the correlations between these two vectors by showing the size of the arrays and the position where the highest correlation between these two signals is found in the four categories.

From the table, with exception of three results, one in the low chain category that shows that there is a negative lag, and the other two in the low-independent category that show a positive lag, we can notice that most of the highest correlations positions are equal to the size of the signals. This leads us to conclude that, if there is an influence of the response rate, it is not reflected in a delay of time.

Consequently, to determine whether there is, indeed, an influence of the responses ratio on the overall score, we divided the best value of each correlation by the variance of each correlation, has shown in table 3.16. The values are between 1 and 0 and the higher the value the more correlated the signals are. As table 3.16 shows, the values of the correlation divided by the variance are really low so there is no relation between response's ratio and the overall score.

To summarize the hotel mangers responses analyses, there is no evidence that the interval between the review and the response influences the overall score nor the response ratio influences the overall score. However, there is still a possible influence by analyzing the lag of other temporal periods as weeks. That is, the influence of the response ratio in overall score can be reflected, for example, in the following week or other temporal windows but we let this analysis to future work.

Chapter 4

Conclusions

With the ease of getting information, tourists require more of the hotels and touristic places. Nowadays, tourists rely on reviews in the decision-making process of choosing a hotel. In turn, hotel managers also rely on reviews to improve the service provided by their hotels. The major problem of this state of affairs is that it is really hard to read every review we got. Therefore, the main goal of this thesis was to know which aspects most influence the overall score of the hotels in order to help hospitality and tourism industry improve the services they provide to customers. We split the hotels we collected into four categories, high-end chain hotels, high-end independent hotels, low-end chain hotels and low-end independent hotels to find out if there are differences in the aspects' importance across categories.

We studied several aspects, those that can be directly scored in Tripadvisor, namely, "Service", "Value", "Cleanliness", "Sleep", "Location" and "Rooms", plus the most analyzed aspects in literature besides the previous ones, namely, "Food", "Guests", "Tourism" and "Decoration".

Once the results of trying to predict the overall score with all aspects were not good, we restricted only to the aspects that can be scored directly in Tripadvisor. After that, we applied the Gini index to the prediction algorithms that achieved the best performances and we got the importance of each aspect. The "Service" stood out among the aspects with a Gini importance of around 0.8 (from 0 to 1) in three categories, high-end chain hotels, high-end independent hotels and low-end independent hotels. In low-end chain hotels, the "Service" also had the best Gini importance value, with 0.35, but "Cleanliness" and "Rooms" had similar results, with 0.28 and 0.26 respectively. "Rooms" had the second best Gini importance, with around 0.15 in the same hotel categories in which the "Service" stood out.

As for the other aspects, "Value", "Cleanliness", "Sleep" and "Location", they had Gini index values near 0, except for the "Value" in low-end chain hotels that had a result of 0.15.

In the second phase of the work we focused on three aspects: the one that most stood out, the "Service", and two aspects that had a Gini index near 0, "Sleep" and "Location". We did a time series analysis in which we found out that "Service" is the most relevant aspect followed by "Sleep" and "Location". This analysis showed that "Sleep" is more important in low-end hotels.

We also did an analysis of the dynamics of the overall score and these three aspects with Jaccard index and we obtained coherent values with the previous analysis but "Sleep" had slightly lower importance than "Location". Finally, the secondary goal of the satisfaction dynamics was to know whether the response ratio from the hotels managers and the time it takes to respond to a review are a cause of changes in the overall score. Although we obtained that there is no correlation between the responses and the overall score, these are not sufficient to draw a definitive conclusion of any of these later questions.

4.1 Innovations and contributions

In this work, there are some innovations and contributions that can be used in academic and practical way. First, we describe the contributions that can be used in further research and then we point out some contributions that hotel managers can rely on to a better comprehension of tourists' perception of hotel quality.

4.1.1 Theoretical contributions

We proposed a way of categorizing hotels that we found to be helpful in addressing differences between categories. We noticed some differences mainly between low end and high-end. Particularly, in the independent low-end hotels, we noticed that there were quite a few differences compared to the other categories.

Regarding the influence of the several aspects studied in this work, we did not find literature that used a method similar to ours. However, the results are consistent with the reviewed literature and constitute evidence that ours is an effective method that can be used by everyone. Every analysis that we did are based on open source software that is easy to use.

Regarding the dynamics of the overall score, we were not able to detect touristic seasonality but applying Jaccard index seems to be fairly easy and effective to detect sudden changes and patterns.

4.1.2 Practical contributions

The major contribution of this thesis was finding the influencing factors of the overall hotel scores of the four hotel categories. Generally, hotel managers should focus on the "Service". From the hotel managers perspective, there is always room to improve. One can also apply this analysis to particular cases, for example to a hotel chain or even to a single hotel.

Regarding the practical contributions of the analyses of the dynamics of tourists' satisfaction, these easily detected patterns and sudden changes allows hotel managers to better prepare for them. In order to get a better insight about on tourists' perceptions on more particular cases, our analysis can also be applied to a hotel chain or to an independent unit.

These analyses contributed to verify and differentiate the tourist trends of each hotel category. Hotel managers can rely on these results to increase the satisfaction of tourists, in particular, in their hotels by applying these analyses. With a better knowledge of the tourists' perceptions, besides the increase of their satisfaction, they can differentiate from the category and consequently have a great competitive advantage over their category. Due to this methodology, in the case of "The Savoy", we found that they are following the high-end independent category, however "Service" is not that related with the overall score as in their category. It would be beneficial for them, to cross, in the moments of those changes, internal knowledge with these analyses in order to improve and possibly differentiate from the category.

4.2 Limitations

The main limitation in this work was assigning scores to the sentences of the written comments associated to many reviews. When predicting the overall score, the sentiment analysis from the text of the review's features had the worst performance. As mentioned in the "Sentiment analysis" section, there is a lot of research in SA that can be done. In this work, the lack of research of this field let us with some limitations. Sarcasm, for example, could not be detected in several sentences that would have the opposite score. Also, some sentences with more

than one aspect have the same score even if the sentence has a contrast connector. With a better sentiment analysis tool, we would have more aspects, therefore, this work would be richer.

Another limitation we had was the distribution of the scores in the hotel categories. That is, the high-end hotels have much more five star ratings than one star ratings, which unbalances the data for the learning algorithms and causes learning biases. Well-balanced data would let the algorithms we tested have better performances without overfitting.

4.3 Future work

During the development of this master thesis, there are some ideas that came to our mind and we would like to try. The following paragraphs describe our ideas that we let for future work due the lack of time or by not being of major interest.

Concerning the web-scraping, one limitation that we had and that we left for future work was the distribution of the ratings in the four categories. Our first goal was to automatically get all the reviews of a hotel by simply using the URL. But, somehow, the program kept stalling in an apparently random process. More than half the time of the thesis development was spent with the web-scraping program, however we could not further explore web-scraping and make the program robust and fully automatic.

Concerning the overall prediction, experiments have been left for future work.

More machine learning algorithms can be tested, unsupervised learning seems to have great performances, SVM prediction, Neural networks or Naive Bayes algorithms would add value to this work.

Strategies to mitigate the missing values issue can also be better explored. Nowadays there are really good imputation strategies that go beyond a simple average or mode, as for example, imputation using multivariate imputation by chained equation or using k nearest neighbors. Moreover, the features to predict the overall score can be better explored. First, due the mentioned sentiment analysis limitations, text from the reviews did not help much. It would be great to explore this more deeply. Also, there were some features that we could have included to improve the performance of the predicting algorithms, as the number of words in each review, or other information of the reviewer as the number of contributions. At last, we only tried four kinds of features, "IS", "ES", "IS &

ES" and "Completed". We could have tried more options such as completing the sentiment analysis with the explicit scores.

Regarding the hotel managers responses, we could have explored more options to find out if the responses influence the overall score. We only considered the response ratio and the time it takes to respond to a review while we might have considered the sentiment analysis of the responses. Also, we concluded that the responses ratio does not reflect with a delay in the overall score but the temporal windows we analyzed were monthly, quarterly and annual, we could have analyzed smaller temporal windows as weeks.

Chapter 5

Bibliography

Nuno Antonio, Ana Maria de Almeida, Luís Nunes, Fernando Batista, and Ricardo Ribeiro. Hotel online reviews: creating a multi-source aggregated index. *International Journal of Contemporary Hospitality Management*, 2018.

A George Assaf, Alexander Josiassen, Ljubica Knežević Cvelbar, and Linda Woo. The effects of customer voice on hotel performance. *International Journal of Hospitality Management*, 44:77–83, 2015.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

Néstor Barraza, Sérgio Moro, Marcelo Ferreyra, and Adolfo de la Peña. Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study. *Journal of Information Science*, 45(1):53–67, 2019.

Jonathan D Barsky. Customer satisfaction in the hotel industry: Meaning and measurement. *Hospitality Research Journal*, 16(1):51–73, 1992.

Katerina Berezina, Anil Bilgihan, Cihan Cobanoglu, and Fevzi Okumus. Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, 25(1):1–24, 2016.

Kenneth L Bernhardt, Naveen Donthu, and Pamela A Kennett. A longitudinal analysis of satisfaction and profitability. *Journal of business research*, 47(2): 161–171, 2000.

Cristina Bernini and Silvia Cagnone. Analysing tourist satisfaction at a mature and multi-product destination. *Current Issues in Tourism*, 17(1):1–20, 2014.

Eivind Bjørkelund, Thomas H Burnett, and Kjetil Nørvåg. A study of opinion mining and visualization of hotel reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, pages 229–238. ACM, 2012.

David C Bojanic. Consumer perceptions of price, value and satisfaction in the hotel industry: An exploratory study. *Journal of Hospitality & Leisure Marketing*, 4 (1):5–22, 1996.

Ana Catarina Calheiros, Sérgio Moro, and Paulo Rita. Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, 26(7):675–693, 2017.

Antoni Serra Cantallops and Fabiana Salvi. New consumer behavior: A review of research on ewom and hotels. *International Journal of Hospitality Management*, 36:41–51, 2014.

Yubo Chen and Jinhong Xie. Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management science*, 54(3):477–491, 2008.

Tat Y Choi and Raymond Chu. Determinants of hotel guests' satisfaction and repeat patronage in the hong kong hotel industry. *International Journal of Hospitality Management*, 20(3):277–297, 2001.

Sara Dolnicar and Thomas Otter. Which hotel attributes matter? a review of previous and a framework for future research. In *Proceedings of the 9th Annual Conference of the Asia Pacific Tourism Association*, pages 176–188, 2003.

Harris Drucker. Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115, 1997.

Wenjing Duan, Bin Gu, and Andrew B Whinston. Do online reviews matter?—an empirical investigation of panel data. *Decision support systems*, 45(4):1007–1016, 2008.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

Ismail Erkan and Chris Evans. The influence of ewom in social media on consumers' purchase intentions: An extended approach to information adoption. *Computers in Human Behavior*, 61:47–55, 2016.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Tomohiro Fukuhara, Hiroshi Nakagawa, and Toyoaki Nishida. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *ICWSM*, 2007.

Andrew B Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52. Association for Computational Linguistics, 2006.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Thorsten Hennig-Thurau, Kevin P Gwinner, Gianfranco Walsh, and Dwayne D Gremler. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of interactive marketing*, 18(1):38–52, 2004.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

Paul Ingram. Organizational form as a solution to the problem of credible commitment: The evolution of naming strategies among us hotel chains, 1896–1980. *Strategic management journal*, 17(S1):85–98, 1996.

Karl G Joreskog and Dag Sorbom. Lisrel vii: A guide to the program and applications. *Chicago: SPSS*, 1988.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A sentence model based on convolutional neural networks. In *Procedding of the 52th Annual Meeting of Association for Computational Linguistics*, 2014.

Walter Kasper and Mihaela Vela. Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference*, volume 231527, pages 45–52, 2011.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

Woo Gon Kim and Seo Ah Park. Social media review rating versus traditional customer satisfaction: Which one has more incremental predictive power in explaining hotel performance? *International Journal of Contemporary Hospitality Management*, 29(2):784–802, 2017.

Xiaojiang Lei and Xueming Qian. Rating prediction via exploring service reputation. In *Multimedia Signal Processing (MMSP), 2015 IEEE 17th International Workshop on*, pages 1–6. IEEE, 2015.

Stuart E Levy, Wenjing Duan, and Soyoung Boo. An analysis of one-star online reviews and responses in the washington, dc, lodging market. *Cornell Hospitality Quarterly*, 54(1):49–63, 2013.

Stephen W Litvin, Ronald E Goldsmith, and Bing Pan. Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, 29(3):458–468, 2008.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Jingjing Liu and Stephanie Seneff. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 161–169. Association for Computational Linguistics, 2009.

Sandra Maria Correia Loureiro and Elisabeth Kastenholz. Corporate reputation, satisfaction, delight, and loyalty towards rural lodging units in portugal. *International Journal of Hospitality Management*, 30(3):575–583, 2011.

Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, and Zhigang Gao. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*, 256:56–62, 2017.

LXML. lxml package. `https://lxml.de/`, 2018.

Majdi Mafarja, Ibrahim Aljarah, Hossam Faris, Abdelaziz I Hammouri, Al-Zoubi Ala'M, and Seyedali Mirjalili. Binary grasshopper optimisation algorithm approaches for feature selection problems. *Expert Systems with Applications*, 117: 267–286, 2019.

Sebastián Maldonado and Richard Weber. A wrapper method for feature selection using support vector machines. *Information Sciences*, 179(13):2208–2217, 2009.

Marcello Mariani, Rodolfo Baggio, Matthias Fuchs, and Wolfram Höepken. Business intelligence and big data in hospitality and tourism: A systematic literature review. *International Journal of Contemporary Hospitality Management*, 2018.

Anna S Mattila and John W O'Neill. Relationships between hotel room pricing, occupancy, and guest satisfaction: A longitudinal case of a midscale hotel in the united states. *Journal of Hospitality & Tourism Research*, 27(3):328–341, 2003.

Aurelio G Mauri and Roberta Minazzi. Web reviews influence on expectations and purchasing intentions of hotel potential customers. *International Journal of Hospitality Management*, 34:99–107, 2013.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

Rodrigo Moraes, JoãO Francisco Valiati, and Wilson P GaviãO Neto. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633, 2013.

Kamal Nigam and Matthew Hurst. Towards a robust metric of opinion. In *AAAI spring symposium on exploring attitude and affect in text*, pages 598–603, 2004.

Peter O'Connor. Managing a hotel's image on tripadvisor. *Journal of Hospitality Marketing & Management*, 19(7):754–772, 2010.

David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

Do-Hyung Park, Jumin Lee, and Ingoo Han. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International journal of electronic commerce*, 11(4):125–148, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Hong-gang Peng, Hong-yu Zhang, and Jian-qiang Wang. Cloud decision support model for selecting hotels on tripadvisor. com with probabilistic linguistic information. *International Journal of Hospitality Management*, 68:124–138, 2018.

Gabriele Pizzi, Gian Luca Marzocchi, Chiara Orsingher, and Alessandra Zammit. The temporal construal of customer satisfaction. *Journal of Service Research*, 18(4):484–497, 2015.

Jiangtao Qiu, Chuanhui Liu, Yinghong Li, and Zhangxi Lin. Leveraging sentiment analysis at the aspects level to predict ratings of reviews. *Information Sciences*, 451:295–309, 2018.

Hailin Qu, Peng Xu, and Amy Tan. A simultaneous equations model of the hotel room supply and demand in hong kong. *International Journal of Hospitality Management*, 21(4):455–462, 2002.

Kenneth Reitz. Requests package. `http://docs.python-requests.org/en/master/`, 2018.

Marco Rossetti, Fabio Stella, and Markus Zanker. Analyzing user reviews in tourism with topic models. *Information Technology & Tourism*, 16(1):5–21, 2016.

Shahrzad Saremi, Seyedali Mirjalili, and Andrew Lewis. Grasshopper optimisation algorithm: theory and application. *Advances in Engineering Software*, 105:30–47, 2017.

Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge & Data Engineering*, 28(1):1–1, 2016.

Markus Schuckert, Xianwei Liu, and Rob Law. Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing*, 32(5):608–621, 2015.

Jaspreet Singh, Gurvinder Singh, and Rajinder Singh. Optimization of sentiment analysis using machine learning classifiers. *Human-centric Computing and Information Sciences*, 7(1):32, 2017.

Alex Smola and SVN Vishwanathan. Introduction to machine learning. *Cambridge University, UK*, 32:34, 2008.

Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 300–307, 2007.

Daniela Soldić Frleta and Dora Smolčić Jurdana. Seasonal variation in urban tourist satisfaction. *Tourism Review*, 2018.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267–307, 2011.

Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. User modeling with neural network for review rating prediction. In *IJCAI*, pages 1340–1346, 2015.

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. *proceedings of ACL-08: HLT*, pages 308–316, 2008a.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM, 2008b.

World Travel and Tourism Council. City travel and tourism impact 2018. `https://www.wttc.org/-/media/files/reports/economic-impact-research/cities-2018/city-travel--tourism-impact-2018final.pdf`, 2018a. Accessed: 2018-11-10.

World Travel and Tourism Council. Travel and tourism economic impact. `https://www.wttc.org/-/media/files/reports/economic-impact-research/regions-2018/world2018.pdf`, 2018b. Accessed: 2018-11-10.

Tripadvisor. Top 25 destinations. `https://www.tripadvisor.com/TravelersChoice-Destinations-cTop-g4`, 2018a. Accessed: 2018-12-10.

Tripadvisor. Number of reviews. `https://www.tripadvisor.pt/Hotels-g186338-London_England-Hotels.html`, 2018b. Accessed: 2018-12-10.

Henry Tsai, Bomi Kang, Ronnie J Yeh, and Eunju Suh. Examining the hotel room supply and demand in las vegas: A simultaneous equations model. *International Journal of Hospitality Management*, 25(3):517–524, 2006.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

USNews. The world's 30 best places to visit in 2017-18. `https://travel.usnews.com/gallery/the-worlds-30-best-places-to-visit-in-2017-18`, 2018. Accessed: 2018-12-10.

Norbert Vanhove. *The Economics of Tourism Destinations: Theory and Practice*. Routledge, 2017.

Ivar E Vermeulen and Daphne Seegers. Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism management*, 30(1):123–127, 2009.

Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792. ACM, 2010.

Wei Wei and Jon Atle Gulla. Sentiment learning on product reviews via sentiment ontology tree. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 404–413. Association for Computational Linguistics, 2010.

Cedric Hsi-Jui Wu and Rong-Da Liang. Effect of experiential value on customer satisfaction with service encounters in luxury-hotel restaurants. *International Journal of Hospitality Management*, 28(4):586–593, 2009.

Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. Opinionseer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, 16(6):1109–1118, 2010.

Zheng Xiang and Ulrike Gretzel. Role of social media in online travel information search. *Tourism management*, 31(2):179–188, 2010.

Fu Xianghua, Liu Guo, Guo Yanyan, and Wang Zhiqiang. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-Based Systems*, 37:186–195, 2013.

Xueke Xu, Xueqi Cheng, Songbo Tan, Yue Liu, and Huawei Shen. Aspect-level opinion mining of online customer reviews. *China Communications*, 10(3):25–41, 2013.

Nira Yacouel and Aliza Fleischer. The role of cybermediaries in reputation building and price premiums in the online hotel market. *Journal of Travel Research*, 51 (2):219–226, 2012.

Qiang Ye, Rob Law, and Bin Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182, 2009.

Qiang Ye, Rob Law, Bin Gu, and Wei Chen. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human behavior*, 27(2): 634–639, 2011.

Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.

Amin Zarshenas and Kenji Suzuki. Binary coordinate ascent: An efficient optimization technique for feature subset selection for machine learning. *Knowledge-Based Systems*, 110:191–201, 2016.

Appendices

Appendix A

Materials, methods and results

Table A.1: Hotels description

Name	Category	Rating	Reviews
The Savoy	High-end and Independent	4.5	3 306
The Dorchester	High-end and Independent	4.5	945
Brown's Hotel	High-end and Independent	4.5	1 132
The Langham	High-end and Independent	5	496
The Connaught	High-end and Independent	4.5	652
Claridge's	High-end and Independent	4.5	1 239
The Goring	High-end and Independent	5	1 116
Windsor House Hotel	Low-end and Independent	2.5	206
Euro Queens Hotel	Low-end and Independent	3	609
The Tophams Hotel	Low-end and Independent	2.5	851
Abercorn House	Low-end and Independent	3	170
Britannia Hampstead Hotel	Low-end and Chain	2.5	700
Airport Inn Gatwick	Low-end and Chain	3	398
Russ Hill Hotel	Low-end and Chain	2	630
Europa Gatwick Hotel	Low-end and Chain	2	809
Britannia Lodge Gatwick	Low-end and Chain	2	38
DoubleTree Hyde Park	High-end and Chain	4	1910
DoubleTree Islington	High-end and Chain	4.5	1870
DoubleTree Chelsea	High-end and Chain	4.5	1504
DoubleTree Marble Arch	High-end and Chain	4	3359
DoubleTree Victoria	High-end and Chain	4	1940
Montcalm Shoreditch	High-end and Chain	4.5	1094
Montcalm Royal London House	High-end and Chain	4.5	746
Montcalm Brewery	High-end and Chain	4.5	2583
The Marble Arch Montcalm	High-end and Chain	5	386

TABLE A.2: Number of ratings assigned to each feature by category

		Service	Value	Cleanliness	Sleep Quality	Location	Rooms
High-end	Chain	8 105	5 349	5 275	4 966	5 339	5 240
	Independent	5 067	3 999	3 928	3 179	3 923	3 525
Low-end	Chain	1 467	1 410	1 425	992	1 402	1 395
	Independent	968	809	794	495	786	737

TABLE A.3: Predicting overall score from the total of reviews.

	IS and ES		ES		IS		Completed	
	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2
RF								
0	1.35	1.25	1.31	1.19	1.35	1.25	1.35	1.25
6	1.12	0.84	1.10	0.79	1.35	1.25	1.12	0.82
Avg	1.13	0.86	1.11	0.81	1.35	1.25	1.13	0.84
Mode	1.12	0.83	1.09	0.78	1.35	1.25	1.12	0.82
Categorical	1.31	1.17	1.29	1.11	1.35	1.25	1.24	0.99
DT								
0	0.90	0.46	0.92	0.52	1.09	0.71	0.91	0.47
6	0.89	0.45	0.91	0.49	1.10	0.73	0.88	0.44
Avg	0.90	0.47	0.90	0.49	1.11	0.75	0.88	0.45
Mode	0.92	0.52	0.93	0.54	1.11	0.74	0.92	0.52
Categorical	0.90	0.47	0.92	0.50	1.10	0.73	0.90	0.47
DTB								
0	0.99	0.41	1.00	0.46	1.14	0.58	0.94	0.39
6	0.95	0.39	0.97	0.42	1.14	0.59	0.94	0.39
Avg	1.00	0.41	1.02	0.46	1.19	0.58	0.96	0.40
Mode	1.07	0.44	1.02	0.50	1.17	0.60	1.05	0.50
Categorical	0.95	0.39	0.95	0.44	1.13	0.60	0.94	0.39

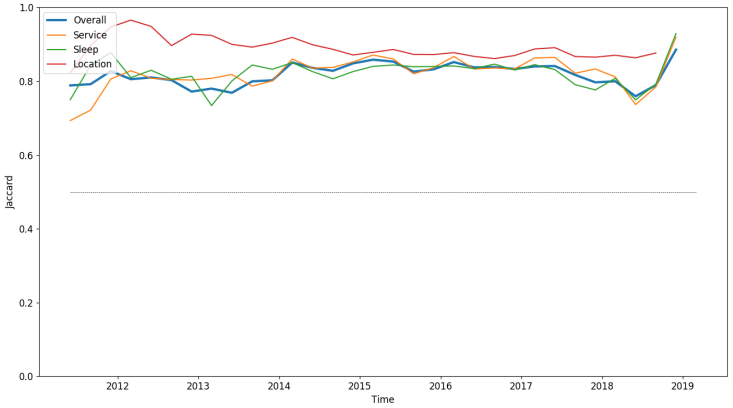

FIGURE A.1: Quarterly Jaccard index in high chain category

74

TABLE A.4: Predicting overall score from high chain reviews.

	IS and ES		ES		IS		Completed	
	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2
RF								
0	1.02	0.55	1.00	0.53	1.07	0.60	1.04	0.58
6	0.91	0.44	0.89	0.43	1.07	0.60	0.92	0.45
Avg	0.91	0.44	0.90	0.43	1.07	0.60	0.94	0.47
Mode	0.90	0.44	0.89	0.43	1.07	0.60	0.92	0.45
Categorical	1.04	0.58	1.02	0.57	1.07	0.60	1.05	0.59
DT								
0	0.73	0.25	0.75	0.27	0.86	0.35	0.75	0.27
6	0.73	0.26	0.74	0.27	0.87	0.37	0.72	0.25
Avg	0.73	0.25	0.74	0.27	0.87	0.37	0.72	0.25
Mode	0.75	0.27	0.76	0.29	0.86	0.36	0.75	0.28
Categorical	0.75	0.27	0.75	0.27	0.87	0.37	0.73	0.25
DTB								
0	0.82	0.24	0.80	0.25	0.93	0.30	0.78	0.22
6	0.81	0.23	0.80	0.26	0.94	0.32	0.85	0.23
Avg	0.83	0.24	0.78	0.25	0.96	0.32	0.78	0.22
Mode	0.93	0.28	0.79	0.27	0.95	0.30	0.93	0.28
Categorical	0.84	0.25	0.78	0.25	0.94	0.33	0.78	0.23

TABLE A.5: Predicting overall score from high independent reviews.

	IS and ES		ES		IS		Completed	
	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2
RF								
0	0.74	0.30	0.74	0.30	0.74	0.30	0.74	0.30
6	0.73	0.29	0.70	0.26	0.74	0.30	0.70	0.26
Avg	0.72	0.28	0.70	0.26	0.74	0.30	0.70	0.26
Mode	0.72	0.28	0.70	0.26	0.74	0.30	0.70	0.26
Categorical	0.74	0.30	0.74	0.30	0.74	0.30	0.74	0.30
DT								
0	0.61	0.19	0.61	0.19	0.69	0.25	0.61	0.19
6	0.61	0.19	0.61	0.18	0.69	0.25	0.61	0.19
Avg	0.61	0.19	0.61	0.19	0.69	0.25	0.61	0.19
Mode	0.61	0.19	0.61	0.19	0.69	0.26	0.61	0.19
Categorical	0.62	0.19	0.62	0.19	0.68	0.25	0.62	0.19
DTB								
0	0.95	0.26	0.67	0.19	0.99	0.29	0.75	0.20
6	0.87	0.23	0.67	0.18	0.95	0.28	0.85	0.22
Avg	0.80	0.22	0.67	0.19	1.02	0.30	0.74	0.20
Mode	0.98	0.27	0.67	0.19	1.03	0.30	0.74	0.20
Categorical	0.83	0.22	0.68	0.19	0.83	0.25	0.67	0.19

TABLE A.6: Predicting overall score from low chain reviews.

	IS and ES		ES		IS		Completed	
	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2
RF								
0	1.34	0.53	1.38	0.55	1.57	0.82	1.26	0.50
6	1.36	0.54	1.37	0.54	1.57	0.84	1.12	0.49
Avg	1.22	0.51	1.38	0.86	1.56	0.81	1.28	0.69
Mode	1.37	0.53	1.36	0.53	1.62	0.81	1.19	0.50
Categorical	1.38	0.56	1.36	0.54	1.54	0.79	1.14	0.46
DT								
0	0.96	0.48	0.97	0.52	1.18	0.75	0.91	0.43
6	0.95	0.48	0.98	0.52	1.19	0.76	0.94	0.49
Avg	0.94	0.49	0.97	0.51	1.18	0.75	0.90	0.44
Mode	0.95	0.49	0.97	0.53	1.19	0.77	0.95	0.51
Categorical	0.99	0.52	0.99	0.55	1.20	0.77	0.97	0.50
DTB								
0	0.93	0.57	0.97	0.58	1.17	0.85	0.93	0.57
6	0.89	0.48	0.96	0.55	1.17	0.85	0.90	0.51
Avg	0.93	0.54	0.96	0.55	1.20	0.88	0.88	0.48
Mode	0.92	0.53	0.97	0.55	1.19	0.87	0.94	0.55
Categorical	0.92	0.51	0.97	0.57	1.17	0.83	0.91	0.51

TABLE A.7: Predicting overall score from low independent reviews.

	IS and ES		ES		IS		Completed	
	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2
RF lu								
0	1.58	0.68	1.57	0.68	1.72	0.92	1.44	0.62
6	1.77	0.88	1.72	0.96	1.85	0.99	1.56	0.70
Avg	1.61	0.69	1.59	0.68	1.78	0.96	1.38	0.59
Mode	1.58	0.68	1.57	0.68	1.73	0.93	1.44	0.62
Categorical	1.66	0.73	1.63	0.70	1.87	0.98	1.51	0.67
DT								
0	1.09	0.61	1.10	0.65	1.25	0.82	1.05	0.55
6	1.07	0.59	1.12	0.68	1.27	0.85	1.04	0.56
Avg	1.08	0.59	1.1	0.65	1.29	0.88	1.03	0.55
Mode	1.12	0.63	1.15	0.7	1.28	0.87	1.05	0.56
Categorical	1.10	0.64	1.12	0.68	1.27	0.84	1.08	0.62
DTB								
0	1.05	0.68	1.09	0.70	1.28	0.99	1.04	0.70
6	1.03	0.63	1.11	0.73	1.29	0.98	1.03	0.64
Avg	1.06	0.67	1.10	0.70	1.29	1.02	1.00	0.61
Mode	1.09	0.74	1.14	0.77	1.28	0.99	1.06	0.73
Categorical	1.08	0.67	1.12	0.74	1.28	1.01	1.05	0.65

TABLE A.8: Predicting overall score balancing the data with SMOTE.

		Total		High Chain		High Unit		Low Chain		Low Unit	
		RMSE	Chi2	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2	RMSE	Chi2
RF	6	1.66	1.85	1.79	2.08	1.84	2.23	1.45	1.30	1.69	1.95
	Avg	1.34	1.04	1.32	0.97	1.52	1.38	1.06	0.48	1.17	0.57
DT	6	1.11	0.65	1.11	0.66	1.29	0.94	0.94	0.44	1.03	0.53
	Avg	1.14	0.72	1.18	0.77	1.28	0.90	0.92	0.43	1.04	0.54
DTB	6	1.09	0.65	1.08	0.66	1.26	0.83	0.91	0.47	1.00	0.52
	Avg	1.10	0.66	1.11	0.68	1.24	0.80	0.91	0.46	1.01	0.54

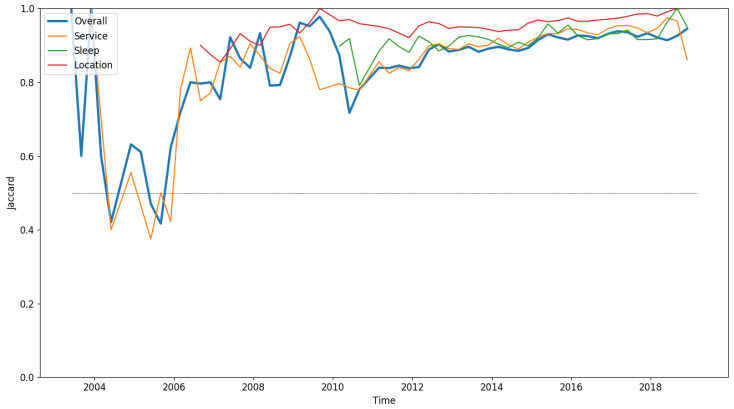

FIGURE A.2: Quarterly Jaccard index in high independent category

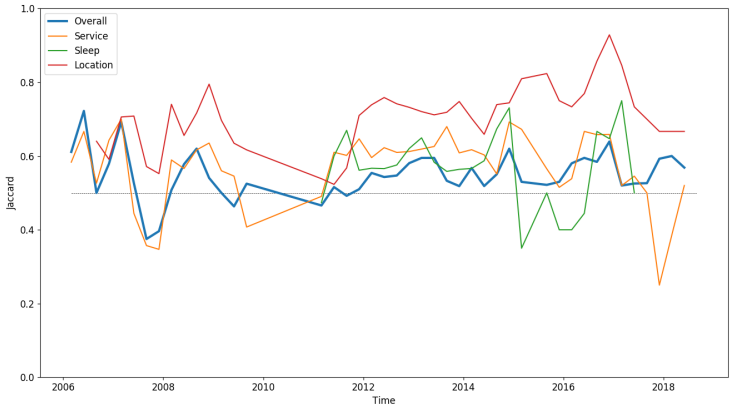

FIGURE A.3: Quarterly Jaccard index in Low chain category

FIGURE A.4: Quarterly Jaccard index in Low independent category

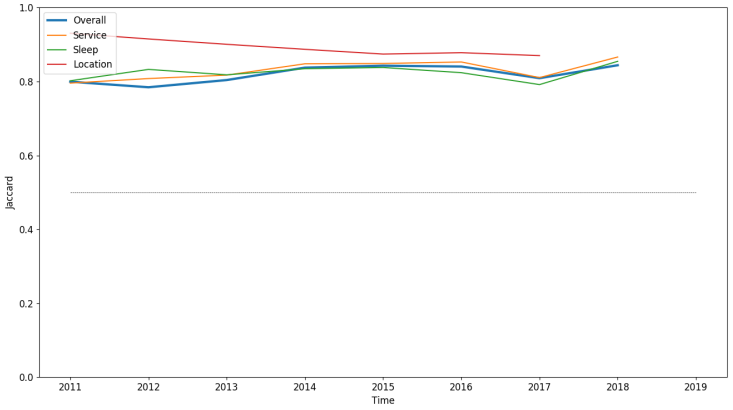

FIGURE A.5: Annual Jaccard index in high chain category

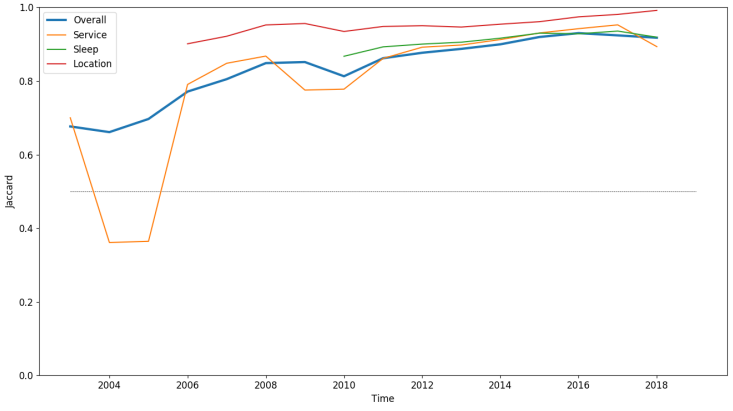

FIGURE A.6: Annual Jaccard index in high independent category

78

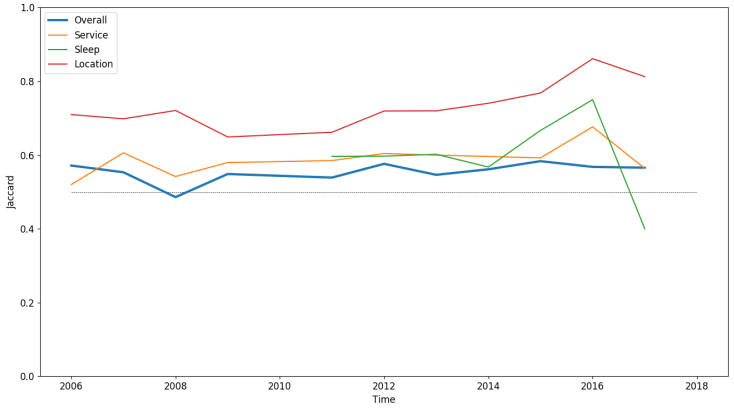

FIGURE A.7: Annual Jaccard index in low chain category

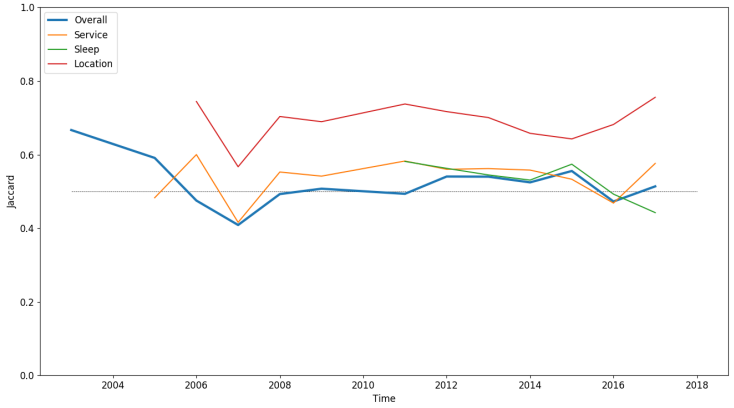

FIGURE A.8: Annual Jaccard index in low independent category