Department of Information Science and Technology

Sentiment analysis in online customer reviews:

The Feels Like Home case

Duarte Rodrigues dos Santos Farinas de Almeida

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

Master in Computer Science and Business Management

Supervisor:

Doctor Raul Manuel Silva Laureano, Assistant Professor,
Department of Quantitative Methods for Management and Economics, ISCTE-IUL

October 2019

(This page was intentionally left blank)

Acknowledgments

Even though this thesis has my name on it, it could not have been finalized without the precious help of several people, either directly or indirectly. They gave me strength, motivation, and courage to reach the finish line of this phase of my life and therefore, I can only be thankful.

To my supervisor, professor Raul Laureano, who was always available to help me, no matter what. Whether I had a quick question or I needed a meeting, he was always there throughout this journey and made sure I always fully understood what, why and how things had to be done. The knowledge he shared with me this year alone will reveal to be very useful to me throughout my career. Thank you.

To my family, the backbone of this work. From my parents up to my grandmother, everyone always showed support when I needed it and that was crucial to keep me going forward towards the finish line. From deep advice to help in day to day tasks, they all made it much easier for me to carry on.

To all my friends who also always told me to keep going and offered help when I needed it the most, without me having to ask for it. In this regard, a special thank you to my friend João, who stood with me every step of the way and made my days significantly better, making it a lot easier to complete certain phases of this work.

(This page was intentionally left blank)

Abstract

Portugal has been, for many years, an attractive destination for tourists from all over the world. This continuous flow of people opens opportunities for companies to explore and for some new other companies to emerge. All the data generated from the interaction of these companies with tourists can be submitted to data mining techniques to extract useful information and, therefore, create knowledge.

This case study uses those data mining techniques to try to explain the polarity of sentiments found in the online reviews of the properties that Feels Like Home, a local accommodation rental platform, manages. Out of the Feels Like Home's portfolio, information regarding negative and positive mentions for each house (monthly) was retrieved from ReviewPro's API, allowing for the final data set to have 1131 entries containing important information to be targeted by data mining.

Through the usage of descriptive analysis and predictive models (CART decision trees), the relationship between the properties and reservations' characteristics and the sentiment polarity found in the reviews is described, as well as the main factors that can help predict those sentiments are revealed. Additionally, the relationship between the monthly occupancy rates and the sentiments' polarity is also described.

This way, this study generates useful knowledge for Feels Like Home and possibly for the rest of the industry to use and adapt to their business needs.

Keywords: Data Mining, CRISP-DM, Sentiment Analysis, Polarity, Rev

Resumo

Portugal tem sido, por muitos anos, um destino atraente para turistas vindos de todo o mundo. Este fluxo contínuo de pessoas abre oportunidades para empresas explorarem e para novas empresas surgirem. Toda a informação gerada a partir das interações entre estas empresas com turistas pode ser submetida a técnicas de data mining para poder extrair informação útil e, assim, gerar conhecimento.

Este estudo de caso usa essas técnicas de data mining para tentar explicar a polaridade de sentimentos encontrada nos comentários online das propriedades que a Feels Like Home, uma plataforma de aluguer de alojamento local, gere. De todo o portfólio da Feels Like Home, informação acerca de menções negativas e positivas para cada casa (mensalmente) foi retirada da API da ReviewPro, permitindo que a amostra final tivesse 1131 entradas contendo informação importante para ser alvo de data mining.

Através do uso de análise descritiva e de modelos preditivos (árvores de decisão CART), a relação entre as características das propriedades e reservas e a polaridade dos sentimentos encontrada nos comentários é descrita, assim como os principais fatores que podem ajudar a prever esses sentimentos são revelados. Adicionalmente, a relação entre as taxas de ocupação mensais e a polaridade dos sentimentos é também descrita.

Desta forma, este estudo gera conhecimento útil para a Feels Like Home e possivelmente para o resto da indústria poderem usar e adaptar às suas necessidades de negócio.

Palavras chave: Mineração de Dados, CRISP-DM, Análise de Sentimentos, Polaridade, Comentários

Index

Index of Tables

Index of Figures

List of Abbreviations

API	Application Program Interface
FLH	FeelsLikeHome
NLP	Natural Language Processing
ML	Machine Learning
KDD	Knowledge Discovery in Databases
POS	Part-of-speech
CRISP-DM	Cross Industry Standard Process for Data Mining
PK	Primary Key
FK	Foreign Key
CART	Classification And Regression Tree

(This page was intentionally left blank)

1. Introduction

The starting point of any research is the identification of a problem or a gap in the literature review, about a specific topic. After this fundamental step, a research question is presented and the objectives that allow for the question to be answered are defined. Finally, the methodological approach of the research is identified.

1.1 Topic's framework

This thesis is a case study related to customer analytics, with a focus on sentiment analysis in online customer reviews of an online local accommodation platform.

As Web 2.0 became a part of our lives, more specifically social media, online interaction between producers and users became easier, generating a big amount of data daily. Comments written by users regarding brands, products, and services in several digital platforms are getting used more often by companies to get better knowledge about consumers (Pang & Lee, 2008). Because of this, the development of techniques that allow for knowledge to be obtained from this data has revealed to be very important, not only for companies but for academic purposes as well (Berry & Kogan, 2010).

This leads to the concept of customer analytics, which can be described as the extensive use of data and different models to make better decisions and actions at an organization level, on a fact-based type of management. Data and models are defined at an individual customer level (Bijmolt et al., 2010). The studied topic for this thesis has, deep down, a connection with this, as the developed work will ultimately focus on improving organizations' decisions and consequential actions based on results and conclusions based on concrete data and models.

Sentiment analysis is a research area that has the intent of analyzing people's opinions and sentiments towards entities such as topics, events, organizations, products, issues, services, individuals, and their respective attributes (Liu, 2012). Heavily related to machine learning, natural language processing, computational linguistics and text mining (Yue, Chen, Li, Zuo, & Yin, 2018), sentiment analysis plays its part by filtering and processing data, providing results that would otherwise be much more difficult to obtain, if not impossible.

The Internet was the main responsible for local accommodation's rapid growth by enabling a cheaper and faster connection between demand and offer (Belk, 2014). As Web 2.0 surfaced, several sharing economy platforms started to emerge, allowing people to share cars (Uber, Cabify), accommodation (Airbnb, Booking), dog-related responsibilities (DogVacay, Rover), food (Feastly) and more (Tanz, 2014).

Regarding the second home phenomenon, it was described by Almeida, who considered second homes as accommodation that belongs to a person who already has a main residence, usually being a city resident (or at least away from the second home), visiting it on holiday and weekends (Almeida, 2009). This phenomenon has been studied not only for the simple act of owning a second house as well as for renting purposes (local accommodation). It is a research topic of great interest to this thesis as the methodology and the final model will both depend on local accommodation related companies, in this case mainly Booking.com, Airbnb and Feels Like Home. Sentiment analysis techniques and local accommodation-based data will both be a big part of this work's development, having data analytics more as a way to relate both these topics.

1.2 Topic's relevance and problem

Feels Like Home (FLH) is a Portuguese accommodation platform founded in 2012 and it manages more than 400 properties for short-term duration rentals, having most of them registered in Airbnb and Booking.com. They manage their costumers' properties in Lisbon, Porto, Algarve, Ericeira and Madeira (Feels Like Home, 2018), compromising to present profit to them.

However, one of the aspects of FLH's processes that is still not well developed is the analysis of the sentiments found in their managed properties online reviews. Up until the start of this work, FLH had a very reactive approach to this subject, instead of having a proactive one, meaning the company would only allocate resources (time and people) to this subject when they realized some properties' reviews were tending to be negative. They realized this had to change, as reviews reflect what their customers think about the service they provide and thus this study comes to contribute to not only science and the local accommodation industry but to FLH as well. Another problem FLH has been facing is irregular occupancy rates throughout the several properties they manage, having high rates in some of them and low rates in others.

This study aims to contribute to FLH's sentiment analysis assessment and evaluation process and to understand what factors affect the polarity of the sentiment. Besides, and using this first contribution as part of the input, this study also aims to contribute to improving FLH's knowledge regarding the relationship between the sentiments and the properties' occupancy rates, by making suggestions to FLH based on the discussion of the results, alongside FLH's manager insights.

Literature focuses a lot on studying platforms like Airbnb and Booking.com, letting companies like FLH, who perform a different kind of service, a bit aside. Furthermore, although sentiment analysis has been used and studied a lot throughout literature in this specific industry (tourism), they focus a lot on major companies with larger datasets at their disposal instead of broadening the research spectrum. Not only that, as a lot of research focuses on reaching conclusions about costumers' habits and sentiment analysis models' accuracy with no model development towards a specific organization's costumers. That's where this topic's relevance relies, on filling a gap in the literature by having the main focus on a specific, smaller company and its customer base, allowing for a solution specifically designed for it.

1.3 Research question and objectives

The main research question is: how do the FLH's managed properties and reservations' characteristics influence the sentiment polarity inherent in their costumers' online reviews and how are these sentiments related to the occupation rates? To answer this investigation question, two main objectives were established.

The first objective is to assess the sentiment associated with each of the FLH's managed properties. Four specific objectives were determined in order to accomplish this first objective:

1. Assess the sentiments in the properties' online reviews, by category (such as location, amenities, room, host, value) and as an aggregated variable (the overall score of a house, regarding its reviews' sentiments), monthly;
2. Characterize the relationship between sentiments and the characteristics of the properties (such as its typology, location, and number of beds);

3. Portray the relationship between sentiments and the characteristics of the reservations (for example, guest's country, most used website to make the reservation and average price per night);
4. Build a sentiment predictive model, using the characteristics of both the properties and the reservations as independent variables.

The second main objective is to evaluate the occupancy rates of FLH's managed properties. Two specific objectives were established to achieve this second objective:

1. Assess the properties' monthly occupancy rates;
2. Characterize the relationship between the monthly occupancy rates and the sentiment polarities.

1.4 Methodological approach

Throughout this thesis, a Cross-industry standard process for data mining (CRISP-DM) methodology will take place, which can be described as an "industry-proven way to guide your data mining efforts" (IBM, 2019). The CRISP-DM project defined a process model that provides a framework to perform data mining projects, independently of both the industry sector and the used technology and it is composed of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Wirth & Hipp 2000).

The developed work will respect these six phases. Therefore, in the first phase, Business Understanding, some business insights will take place regarding Feels Like Home's case in order to fully understand every factor business-related that could influence this work's performance.

Then, in the second phase (Data Understanding), the use of an API's features will provide the needed data to proceed into the next steps. Not only that as Feels Like Home will provide historic data from their managed properties, which, after being properly treated (removing null references and rows with missing data, for example), will allow for its comprehension and, therefore, for this study to go on to the next phases.

Data Preparation will follow the previously described phases, consisting of all actions carried through to properly form a final dataset and feed it to the proposed model. This step's output will serve as input for the next phase.

Modeling will consist of the description of relationships between the relevant variables to this work and the creation of the proposed predictive model for sentiments' polarity (in this case decision trees), in order to fulfill the proposed objectives.

Afterward, Evaluation will analyze how accurate the used model was and how successful was the overall performed work. Besides that, the models will also be evaluated accordingly to a business perspective. Not always the best statistical model is the best model for business and the model that will prevail is the one that is the best for business.

Finally, in the deployment phase, this thesis will be available in the ISCTE-IUL repository and the main results and contributions will be published in a scientific article and, hopefully, disseminate useful knowledge across the scientific community.

1.5 Document structure

This document is structured in five main chapters. These are meant to reflect the developed work, from its conception to its conclusion. The five chapters are the Introduction, State of the Art, Methodology, Results and Discussion and Conclusions.

The Introduction describes the topic and its relevance and the goals this thesis strives to achieve as well as explain the context and the motivation. State of the Art resumes extensive research about what literature on this thesis' framework has already developed to the present day, such as sentiment analysis, text mining, data mining, and local accommodation. Assumptions, conclusions and concrete data results are showcased for the reader to understand everything literature on the subject has to offer.

Methodology describes and explains the work that is developed throughout the thesis, showing how the researcher intends to get to the proposed results and conclusions (elaborating on data retrieval and processing, as well as the chosen analysis methods).

The Results and Discussion chapter discusses the results the previous chapter produced, commenting on expected and unexpected outcomes.

Finally, the conclusions chapter wraps the whole work developed throughout the dissertation, presenting conclusions as well. Limitations of the present study are explained, solutions are proposed to FLH and future research projects are proposed to further develop scientific knowledge about the thesis' topic.

2. State of the Art

In the state of the art, numerous previous studies that present relevance to this thesis' development will be analyzed and reviewed, in order to establish some background on what has already been done in terms of research. The presented topics are the second home phenomenon/local accommodation, sentiment analysis and sentiment analysis applied to customer reviews on the Airbnb and Booking.com platforms.

A literature review protocol was followed to search for scientific information to include in the state of the art. The majority of this research was based on scientific articles, which were found using different combinations of the following factors:

Sources (digital libraries)

IEEE Xplore (www.ieeexplore.ieee.org/Xplore/home.jsp), SpringerLink (www.link.springer.com), b-on (www.b-on.pt/en) and Google Scholar (www.scholar.google.com).

Methods (filters used in the search)

Keywords [("Sentiment analysis" OR "Text mining" OR "Online reviews" OR "Sentiment Polarity" OR "Opinion mining" OR "Occupancy Rates" OR "Hotel performance") AND ("Airbnb" OR "Booking.com" OR "Local accommodation" OR "Hotels")]; [("Factor that influence" OR "Determinants of" OR "What influences" OR "What drives") AND ("Online reviews sentiments" OR "Reviews' sentiments" OR "Occupancy rates") AND ("On booking.com" OR "On Airbnb" OR "Hotels" OR "On local accommodation")].

Time period (2010-2019). There were exceptions throughout the literature review, as some subjects didn't have much information in this time period.

Number of citations (at least 100 citations). Again, there were exceptions, as some subjects lacked articles with the minimum number of citations previously set.

2.1 Sentiment analysis

Sentiment analysis (which is also mentioned as opinion mining) is a research area with the intent of analyzing people's opinions/sentiments towards entities such as topics, organizations, events, products, issues, services, individuals, and their respective attributes (Liu, 2012). Its scientific area is somewhere along the lines of machine learning, data mining, natural language processing (NLP) and computational linguistics and it only started to get proper attention around 2005 (Yue et al., 2018).

2.1.1 Related research areas

Machine Learning

Machine learning (ML) can be described as a computational process that, without being "hardcoded" to produce a specific outcome out of input data, can still perform the desired task (El Naqa & Murphy, 2015). These authors also mention three main ways a computational algorithm can adapt itself in order to learn: supervised (each training example of input data is paired with its known classification label, so the algorithm learns all about the differences and similarities in the input data and becomes able to deal with unseen future data), unsupervised (the algorithm learns in a trial and error manner by adjusting itself after each attempt to get closer to its target, getting to a point where the algorithm reaches its target even if it changes) and semi-supervised (only a part of the input data is labeled, helping the algorithm to learn the unlabeled part). Other authors mention the contribution of ML to society, as ML suitable tasks will replace people at their jobs, making non-suitable tasks more valuable. It will also enable new products, services and processes to society, as well as it will allow for the development of software to become easier due to the possibility of running input data on an already existing ML algorithm instead of creating code from scratch (Brynjolfsson & Mitchell, 2017).

Data mining

Another research area related to sentiment analysis is data mining, which can also be called Knowledge Discovery in Databases (KDD). Its goal is to discover new (and potentially useful) information, having large amounts of data as input. Several fields have benefited from this research field, such as retail sales, bioinformatics and even counter-terrorism (Baker, 2010). Data mining's main economic driver to its tools and

techniques development might have come from the commercial world, as commercial databases keep getting bigger and more numerous and all the data in them, through data mining, might provide a financial return to the owners of these databases. The scientific community also benefits from these techniques, as the mismatch between data and the predictions of a theory can result in advances and innovation (Hand, 2013).

Natural Language Processing

Natural Language Processing (NLP) is a set of techniques to enable an automatic representation and analysis of the human language (Cambria & White, 2014). It can also be described as a field that tries to convert human language into a representation that is easier for computers to use as input (Collobert & Weston, 2008). These "natural occurring texts" (human language) result from natural interactions between humans (when communicating) and can be in any form (oral or written) and any language, mode or genre and should not be created for processing purposes, but naturally, from real interactions (Liddy, 2001).

2.1.2 Psychology background

Sociology and psychology are two main areas that have certain elements present in sentiment analysis (Yue et al., 2018). Upon further research, evidence of this statement can be found in other studies, where even using "sentiment analysis" and "opinion mining" as two terms for the same thing is even questioned, all due to the definition of sentiment itself still being ambiguous. A connection between emotion and sentiment was established, mainly when McDougall said that sentiment usually connects primary emotion with action (Yue et al., 2018). Emotions studies have two main approaches: emotion as finite categories and emotion as dimensions (Devitt & Ahmad, 2013). While the former proposes a finite set of basic emotions experienced universally, the second one delineates multiple dimensions on which every possible emotional state can be represented. Figure 1 shows Watson and Tellegen's two-dimensional emotional model, which expresses opposite emotions in its opposite sides, across four two-dimensional factors (pleasantness/unpleasantness, high/low positive affect, strong/weak engagement, and high/low negative affect).

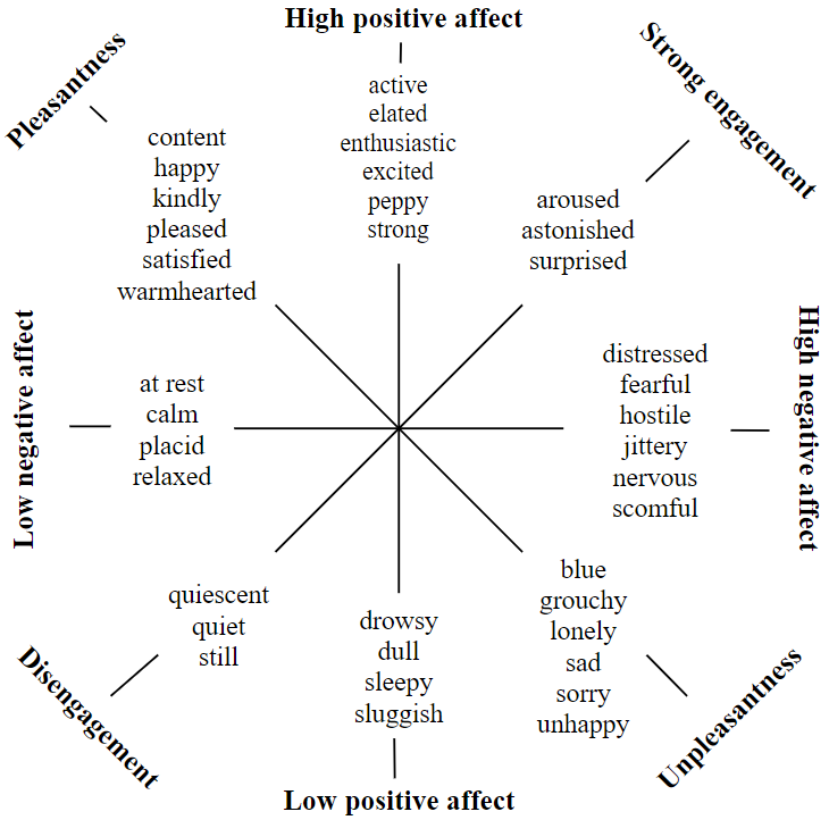

Figure 1: Two-factor emotions model

Source: adapted from Watson and Tellegen (1985, p. 22)

2.1.3 Techniques, approaches and resources

Different approaches take place in sentiment analysis, making use of different techniques and resources such as preprocessing, extraction and the use of classifiers.

Preprocessing

Preprocessing is mandatory before performing any data mining activity (Hemalatha, Varma, & Govardhan, 2012). After gathering all the data from the dataset, its objective is to prepare a text before it undergoes sentiment analysis, preparing it for better results. Preprocessing mainly consists of eliminating arbitrary sequences of whitespaces between words, detecting sentence boundaries, identifying and correcting spelling errors, eliminating arbitrary use of punctuation marks and capitalization and more. Common techniques include sentence splitting (sentence boundary detection), stemming (reducing words to their base form), part-of-speech tagging (labeling tokens – separate words in a sentence represent different tokens - with their corresponding word type) and parsing (revealing the structure of sentences, for example, which words can be grouped into phrases and which words are the subject or the object of a verb) (Petz et al., 2015). Stop words removal is also very common.

Extraction

There are three levels of extraction in sentiment analysis: document level, sentence level, and aspect/feature level. The first one classifies entire documents regarding their polarity towards an entity (so the whole document only produces one opinion, negative or positive) (Lavanya, JC, & Veningston, 2016), with its biggest challenges being cross-domain sentiment analysis and cross-language sentiment analysis (Yue et al., 2018).

The second one is quite self-explanatory as it classifies a sentence depending on whether it expresses a positive, negative or neutral opinion (Lavanya et al., 2016) and a lot of early studies try to focus on identifying subjective sentences. However, complex tasks such as dealing with conditional sentences and sarcasm will be faced. Sentence level sentiment analysis is the way to go in these cases (Yue et al., 2018).

Finally, the aspect level performs a more detailed evaluation/classification of an opinion about an entity (Lavanya et al., 2016), going as far as determining, for a specific entity, opinions expressed on its different aspects or features (Yue et al., 2018). Features/aspects are attributes of an entity, thereby implying that tasks such as entity identification, extracting their respective features and determining the polarity of the features' expressed opinions will be faced (Liu, Hu, & Cheng, 2005).

Classifiers: Machine learning and lexical resources

Furthermore, two categories of techniques are used for sentiment analysis, when it comes to classifying the sentiments present in the extracted and preprocessed text: machine learning and lexicon-based techniques. Machine learning techniques are generally used in the feature and sentence levels (Singh, Singh, & Singh, 2016) and are considered to have a superior performance over the lexicon-based techniques, making them have a higher reputation in the industry (Li, Goh, & Jin, 2018). Several different techniques have been used in sentiment analysis models throughout research, making use of both machine learning and lexicon techniques. These include unigrams, part-of-speech (POS), semantic sentiment analysis, n-gram, social relations for user-level sentiment analysis, topical sentiment analysis, lexicon approaches based on the existence of polarity words, phrase-level sentiment analysis, pattern-based semantic analysis and others (Lavanya et al., 2016).

The most used classifiers, although, are SVM, Naïve Bayes, and decision trees based ones, as seen in many studies (Singh et al., 2016) (Lavanya et al., 2016) (Singla, Randhawa, & Jain, 2017). The latter author even suggested the SVM model has the best accuracy out of the three, being Naïve Bayes the least accurate.

Regarding lexical resources, there are many available to choose from, such as SentiWordNet, WordNet-Affect, General Inquirer, Dictionary of Affect in Language, MPQA, SentiNet, and others (Devitt & Ahmad, 2013) (Musto, Semeraro, & Polignano, 2014). Any of these can be chosen to be a sentiment analysis' classifier, but the same two studies just mentioned above offer some insights on these different options, with Devitt and Ahmad mainly stating that the first four mentioned lexicons, even having different origins, theoretical underpinnings and development criteria, are very consistent regarding what they represent and how they represent it. However, they are still different among them and can represent different impacts on a sentiment analysis system. Musto, Semeraro and Polignano, on the other hand, compared MPQA, SentiWordNet, WordNet-Affect and SentiNet, arguing that the first two had the best results amongst the four lexicons.

2.1.4 Studies using sentiment analysis

There are numerous studies throughout literature using sentiment analysis systems to reach conclusions on many different subjects. Researchers mostly make use of sentiment analysis systems designed by themselves as the main tool to make different datasets undergo their tests and validate their findings, but what data do they test?

Twitter

Social media is a big target for sentiment analysis studies due to its growth and big amount of data (Yue et al., 2018). One of the main social networks used for this kind of research is Twitter, mainly due to it including opinions about products and services (Hao et al., 2011) and its popularity, which has led to the development of applications and research in various domains using it as an information source (Kumar, Morstatter, & Liu, 2014). Twitratr, tweetfeel and Social Mention are some examples of companies that provide Twitter sentiment analysis as one of their services (Kouloumpis, Wilson, & Moore, 2011).

However, one could argue that Twitter's publications (or tweets) have a more casual language when compared to web reviews, and due to its limitations regarding characters per publication (driving to a lot more abbreviations), things can get tougher (Hao et al., 2011). The study led by these previously mentioned authors led to easy to read results, after analyzing tweets using natural language processing techniques to figure out their polarity (see Figure 2). 59,614 tweets were analyzed over five days and all of them were comments on the movie Kung-Fu Panda.

Figure 2: Pixel Sentiment Geo Map
Source: Hao et al. (2011, p. 278)

In 2011, Kouloumpis, Wilson and Moore argued that including microblogging features, such as the presence of intensifiers (e.g. caps lock and character repetition) and positive/negative/neutral emoticons and abbreviations is of great use when using sentiment analysis in the microblogging scenario (in this case, Twitter was used as well). On the flip side, they identified POS (part-of-speech) features to be of less relevance, not contributing to the overall accuracy of their model as much as the other used features (n-gram, lexicon and microblogging features), although this could be justified by the results of the tagger (Kouloumpis et al., 2011). As seen in Figure 3, the addition of the POS features did not contribute much to the average accuracy of the used models (note: HASH data was based on tweets containing a specified set of hashtags, while HASH+EMOT was based on HASH data plus tweets containing the ":)" and ":(" emoticons, excluding tweets that had both of them).

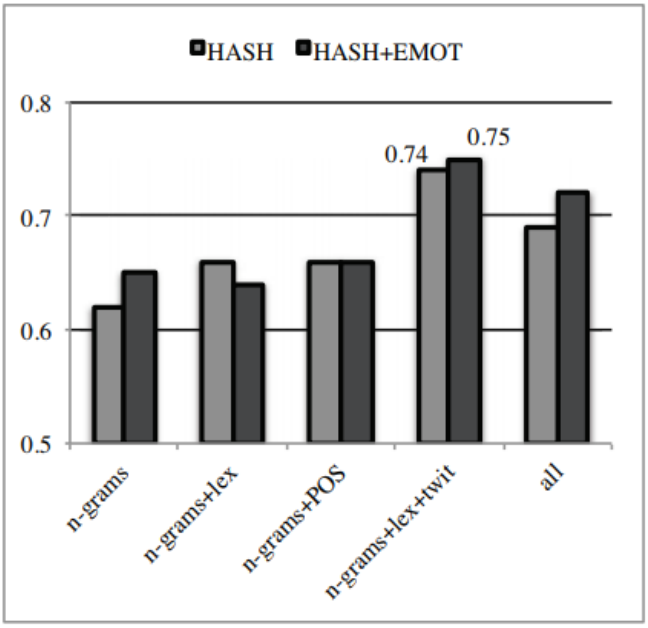

Figure 3: Average accuracy model on the test set over HASH and HASH+EMOT
Source: Kouloumpis et al. (2011, p. 541)

News and blogs

Using a different dataset, Godbole, Srinivasaiah, and Skiena (2007) opted to choose news articles and blogs as their study's dataset. Not only did they associate different opinions to their respective entities, as they aggregated all entities and scored them relatively to each other.

They established a lexicon through path analysis. What they did was they attributed a polarity (positive or negative) to each word and queried WordNet for its synonyms and antonyms. This generated several paths of words, which were then scored, depending on their length and the number of polarity flips they had. They also improved accuracy by limiting the resulting words (synonyms and antonyms) to the top results returned by WordNet and, after converting the scores to z-scores, proceeded to remove what they considered to be ambiguous words. Finally, using the words' appearance frequency, they evaluated polarity (percentage of positive sentiment references among total sentiment references) and subjectivity (reflects the amount of sentiment an entity garners, regardless it being positive or negative). Tables 1 and 2 show the obtained results, clearly showing different opinions between the two used data sets (news articles and blogs) (Godbole, Srinivasaiah, & Skiena, 2007).

Actor	Net Sentiment		Actor	Net Sentiment	
	News	Blogs		Blogs	News
Felicity Huffman	1,337	0,774	Joe Paterno	1,527	0,881
Fernando Alonso	0,977	0,702	Phil Mickelson	1,104	0,652
Dan Rather	0,906	-0,04	Tom Brokaw	1,042	0,359
Warren Buffet	0,882	0,704	Sasha Cohen	1	0,107
Joe Paterno	0,881	1,527	Ted Stevens	0,82	0,118
Ray Charles	0,843	0,138	Rafel Nadal	0,787	0,642
Bill Frist	0,819	0,307	Felicity Huffman	0,774	1,337
Bem Wallace	0,778	0,57	Warren Buffet	0,704	0,882
John Negroponte	0,775	0,059	Fernando Alonso	0,702	0,977
George Clooney	0,724	0,288	Chauncey Billups	0,685	0,58
Alicia Keys	0,724	0,147	Maria Sharapova	0,68	0,133
Roy Moore	0,72	0,349	Earl Woods	0,672	0,41
Jay Leno	0,71	0,107	Kasey Kahne	0,609	0,556
Roger Federer	0,702	0,512	Tom Brady	0,603	0,657
John Roberts	0,698	-0,372	Ben Wallace	0,57	0,778

Table 1: Top positively rated entities, news on the left and blogs on the right
Source: Godbole et al. (2007, p.4)

Actor	Net Sentiment		Actor	Net Sentiment	
	News	Blogs		Blogs	News
Slobodan Milosevic	-1,674	-0,964	John Muhammad	-3,076	-0,979
John Ashcroft	-1,294	-0,266	Sammy Sosa	-1,702	0,074
Zacarias Moussaoui	-1,239	-0,908	George Ryan	-1,511	-0,789
John Muhammad	-0,979	-3,076	Lional Tate	-1,112	-0,962
Lionel Tate	-0,962	-1,112	Esteban Loaiza	-1,108	0,019
Charles Taylor	-0,818	-0,302	Slobodan Milosevic	-0,964	-1,674
George Ryan	-0,789	-1,511	Charles Schumer	-0,949	0,351
Al Sharpton	-0,782	0,043	Scott Peterson	-0,937	-0,34
Peter Jennings	-0,781	-0,372	Zacarias Moussaoui	-0,908	-1,239
Saddam Hussein	-0,652	-0,24	William Jefferson	-0,72	-0,101
Jose Padilla	-0,576	-0,534	King Gyanendra	-0,626	-0,502
Abdul Rahman	-0,57	-0,5	Ricky Williams	-0,603	-0,47
Adolf Hitler	-0,549	-0,159	Ernie Fletcher	-0,58	-0,245
Harriet Miers	-0,511	0,113	Edward Kennedy	-0,575	0,33
King Gyanendra	-0,502	-0,626	John Gotti	-0,554	-0,253

Table 2: Top negatively rated entities, news on the left and blogs on the right
Source: Godbole et al. (2007, p.4)

More recently, Shirsat, Jagdale, and Deshmukh (2017), also using news articles as the chosen dataset, did an article on document-level sentiment analysis. The dataset was from BBC's dataset and contained 2225 documents on many different areas, but mainly Business, Entertainment, Politics, Sport, and Technology. After using different preprocessing techniques, such as data cleaning (they removed URLs, stop words, punctuation, numbers and stripped white spaces from the data, as well as they used r studio tool functions to convert everything into lowercase letters) and steeming

(converting words into their root form), they determined the Term Document Matrix to know how often/frequently terms occurred in the processed documents. Finally, they summarized their results with the use of tables and graphs, concluding about positive and negative opinions on the five main studied areas (Shirsat, Jagdale, & Deshmukh, 2017). Table 3 shows their results.

Sr. No	*Name of Category*	*Total*	*Positive*	*Negative*	*Neutral*
1	Business	510	262	214	34
2	Entertainment	401	136	244	21
3	Politics	417	210	190	17
4	Sport	511	151	327	33
5	Tech	401	136	244	21

Table 3: Documents' polarity on the five studied categories
Source: Shirsat et. al (2017, p. 3)

2.2 The second home phenomenon and local accommodation

Tourism is a very important tool to promote regional economic and social development, being one of the fastest-growing sectors in the whole world (Moswete, Thapa, Toteng, & Mbaiwa, 2008). While there were 25 million international tourists in 1950, this number had a stable growth throughout the years, reaching 1.186 million in 2015, with total revenue of around 1.260 billion US dollars in that same year (UNWTO, 2016). As of 2017, according to the World Tourism Organization, 1.323 million tourist arrivals took place worldwide, representing an increase of 84 million when compared to the previous year (INE, 2018). Being a very broad term and having many subcategories, tourism itself has been thoroughly studied throughout the years.

One of the emerging categories under this umbrella term is second home rentals (or residential tourism), a derivative of the second home phenomenon. Since it was first addressed, second home research work has suffered many changes, especially in recent years. However, many have argued that this recent development consists of individual points of view with little to no connection whatsoever (Müller, 2014). Contributing to the complexity of second home studies, there's the fact that we can find relevant literature for the second home phenomenon not only in tourism studies but also in other related research fields. On top of this, different authors have been using different

terminology to refer to the same phenomenon (second homes) (Müller, 2014). All of this contributes to a lack of consensus around the definition of this phenomenon, especially when combined with the complexity of its origins, frequency of occupancy and the purpose of use of second homes (Roca, Oliveira, Roca, & Costa, 2012). Despite all this, Almeida has referred to second homes as "accommodation belonging to a person who already has a main residence and usually resides in a city or at least away from this villa, visiting it on weekends or holidays" (Almeida, 2009).

One last thing to notice as well regarding these studies is the geographical factor. Nordic and North American regions used to lead the second home research, while nowadays studies from all over the world are available, such as research from China by Huang and Yi in 2011 and Hui and Yu in 2009, Central America by Barrantes-Reynolds in 2011 and South Africa by Visser and Hoogendoorn in several years (Müller, 2014).

2.2.1 Studies on second homes: concrete data

International Studies

Residential tourism is one of the biggest segments of the tourism-based accommodation market, showing regular growth in the case of Spain in recent years, for example. It takes its place in the market as a solution to the lack of rooms that can, from time to time, affect the most desirable destinations for foreigners during high seasons of tourism. Still using Spain as an example, as of 2011, about 16% of all habitations were classified as second homes, with another 14.8% being considered potential second homes (Saló & Garriga, 2011). Using another example, in Denmark, regarding all overnight stays by nonresidential tourists, more than fifty percent are second home rentals (Skak & Bloze, 2017).

However, not everyone rents out their second homes to get some extra income. Bieger, Beritelli and Weinert (2007) identified social differences in the willingness to rent out second homes in Switzerland. Their research pointed out that, depending on which individual consequences of renting their second homes is under scope (such as it being an intrusion to privacy, being financially unattractive or potentially damage personal items), different relationships on a Likert-type scale can be established with, for example, the owners' generation and the owners' age when they bought their second homes (Bieger, Beritelli, & Weinert, 2007). Even with a higher demand for overnights

in the second home market (1.1% growth in the previous 3 years), 97% of the inquired people were unwilling to rent their second homes.

Portuguese studies

In 2017, about 19 million tourist trips occurred in Portuguese territory, including local accommodation stays. Having that said, in that same year, 2,663 establishments were on the local accommodation market (as of July), representing 66,6 thousand available beds. A total of 3,4 million guests were hosted by this kind of establishments and, on average, each stay lasted for 2,35 nights, resulting in a total 8 million overnight stays spent by the guests (INE, 2018).

Private accommodations (we'll dive deeper into those later) in Portugal were not officially registered and taxed for some time (parallel economy), leading the Portuguese government to establish the Decree-Law No. 128/2014 (came into force in November 2014) to regulate this practice (Ramos & Almeida, 2017). According to the same authors, the previously referred legislation was updated in 2015, being published in April 2015 (Decree-Law No. 63/2015).

In 2012, a study (Roca et al., 2012) concluded that the majority of second homes in the West region of Portugal were owned by Lisbon Metropolitan Area residents, followed by Portuguese emigrants (mostly from France and Germany) and foreigners (mainly from the United Kingdom). This study also focused on numerous aspects of second home possession, showcasing the differences between the different groups of people under study (the ones previously mentioned). Those aspects included frequency of use, types of family (for example, couples with children or elderly people), location, propensity to change the second into the first home, and many more. The study also concluded that the impacts of second homes were generally regarded as something positive by the local authorities on most main aspects (cultural, environmental, economic and social) in all types of parishes (rural, semi-rural and urban), mainly because it is seen as a way to compensate for negative demographic trends (such as an aging agricultural population), leading to the diversification of the local economy (Roca et al., 2012).

2.2.2 Booking and Airbnb

Booking and Airbnb are major companies when it comes to local accommodation and renting out second homes. Recent years have been of great growth for both companies. While the first was initially only for hotels and guest houses but has expanded its offer to local accommodation, Airbnb has always had a very broad range of choices, having, in 2016, 2 million places all over the world (Gomes, Pinto & Almeida, 2017).

According to Airbnb itself, "Airbnb's mission is to create a world where people can belong through healthy travel that is local, authentic, diverse, inclusive and sustainable. Airbnb uniquely leverages technology to economically empower millions of people around the world to unlock and monetize their spaces, passions, and talents to become hospitality entrepreneurs" (Airbnb, 2019). As of 2018, there are also the so-called "Experiences", provided by their platform as well. "With Experiences, Airbnb offers unprecedented access to local communities and interests through 15,000+ unique, handcrafted activities run by hosts across 1,000+ markets around the world" can be read on their website (Airbnb, 2019).

Booking, on the other hand, says they "connect travelers with the world's largest selection of incredible places to stay, including everything from apartments, vacation homes, and family-run B&Bs to 5-star luxury resorts, tree houses, and even igloos" (Booking, 2019).

Airbnb and the hotel industry

The disruptive innovation theory - proposed by Clayton Christensen (Bower & Christensen, 1995; Christensen, 1997; Christensen & Raynor, 2003) – can and should be used to examine Airbnb's success (Guttentag, 2015). Disruptive innovation has the power to transform the market. Usually, a disruptive product underperforms the market's most dominant products regarding their key selling points. However, it tries to shine by offering a different set of benefits, such as being cheaper, more convenient or simpler. While in the initial phase this type of innovations appeal to the low end of the market, they tend to improve over time, gradually shifting the market's attention towards the new product/service. This can be a big problem for the companies that used to rule the market before these disruptive innovations came into place, causing them to struggle to compete (Guttentag, 2015).

This is becoming a reality to the hotel industry. Guttentag and Smith studied, in 2017, the travelers' behavior when it comes to choosing between Airbnb's peer-to-peer accommodation and the more traditional approach of the hotel industry. 64.8% of all respondents admitted they would've used a hotel for their stays if Airbnb did not exist at all, leading to the conclusion that Airbnb has taken those customers from hotels, more specifically from mid-range hotels (43.1%) (Guttentag & Smith, 2017). The same study also implies that the type of accommodation Airbnb seems to be replacing also varies depending on factors such as the guest's age, the trip's duration, and the household financial status, to name a few examples. In 2015, Tussyadiah concluded that the three main factors for people to prefer using peer-to-peer accommodation (such as Airbnb) were sustainability, community, and economic benefits, the latter being the one of most relevance.

One thing Airbnb has to overcome though, is the hosts' trust in renters, a problem not seen so much in hotels. As Mittendorf stated, in this kind of market a personal interaction is required to establish a sharing deal between the host and the renter. This brings up high levels of risk and complexity to the sharing economy, because many times, the host is sharing his/her private property with an Airbnb user, who is a stranger to the owner. He concluded that the level of trust the owners have in renters has a significant impact on the owner's intentions, on whether to rent or not. He finalized by recommending both hosts and renters to adopt the trust-building measures provided by Airbnb, as well as recommending Airbnb to improve these (Mittendorf, 2016).

On the flip side, renters' trust in hosts is also something to be accounted for. Research in Stockholm, Sweden, points out that Airbnb customers are influenced by their impressions of the hosts' photos on the listings, turning into one of the many factors when making purchase decisions. Although factors like the listing's characteristics (such as apartment size and location) also affect greatly renters' decisions, the listing itself, through the use of the host's photo, can also influence consumers, as well as the listing's price. This is also true even when the host's reputation varies (Ert, Fleischer, & Magen, 2016).

Speaking of prices, dynamic pricing also proves to play an important role in Airbnb's ecosystem when it comes to the hosts' performance (Oskam, van der Rest, & Telkamp, 2018). Not only frequent adjustments to the prices prove to have better financial results at the end of the day, but properties with a higher mean positive price change also

present better results. Furthermore, even though Airbnb has an integrated dynamic pricing tool, some hosts outperform the platform's built-in tool.

Booking.com and customer behavior

While Airbnb always took a more second home/local accommodation approach to business, Booking started with a more hotel experience kind of approach, changing this later to allow users to register their apartments on the platform.

One of the main strategic purposes for most organizations is customer loyalty and Booking.com is no exception, considering the positive impact it has over their results (Marques & Cruz, 2014). The exponential growth of internet as a means of business and as a platform to create value has been observed in recent years and the same authors point out that customers' personalities have an impact on brand loyalty, as there are customers with a relational profile (these tend to be more loyal towards a certain brand) and customers with a researcher profile (these tend to explore new options and are willing to try new things, therefore representing a lower level of loyalty towards a specific brand). Customers' satisfaction and Booking's trustworthiness are also factors that contribute towards brand loyalty, according to the same authors.

There are, besides brand loyalty, other factors customers consider before picking Booking.com as their preferred booking platform. As showed in a 2017 study, although functional value (as perceived by customers) was somewhat similar when comparing Booking to Airbnb, social and hedonic values were not. Also, hedonic and social values proved to have a larger impact in Airbnb's case (Schaffner, Georgi, & Federspiel, 2017). Díaz and Rodríguez also stated that the fewer variables platforms like Booking.com, TripAdvisor or Holiday Check have to evaluate the listing's performance, the less will customers be able to make useful decisions (Díaz & Rodríguez, 2018).

Finally, and regarding prices, Borges, Pereira, Matos, and Borchardt did a study on Booking's listings prices, depending on customers' satisfaction. No relevant relationships were found, taking into account customers' reviews and other factors (Borges, Pereira, Matos, & Borchardt, 2015).

Airbnb and Booking as P2P (Peer-to-Peer) networks

As previously mentioned, Airbnb is an example of a for-profit P2P (Peer-to-Peer) network platform (Oskam & Boswijk, 2016). Oskam and Boswik also referred the existence of other network platforms, besides the profit purpose, referring the

community's benefit as a result of the platform's use (such as Wikipedia and Linux) and the controlled/open variant, being Wikipedia considered open and Facebook controlled. In the case of Airbnb's type of network though, consumers deal with their own assets and thus gain the power to be co-creators of value (Boswijk, Peelen, & Olthof, 2015). In this case, consumers rent out their houses and apartments in order to create value.

Other examples of value networks include eBay, Uber, and SnappCar, as seen below in Figure 4, that establishes the differences in some P2P networks regarding the organization's control over interactions, purpose (for the community's benefit/for profit) and other factors.

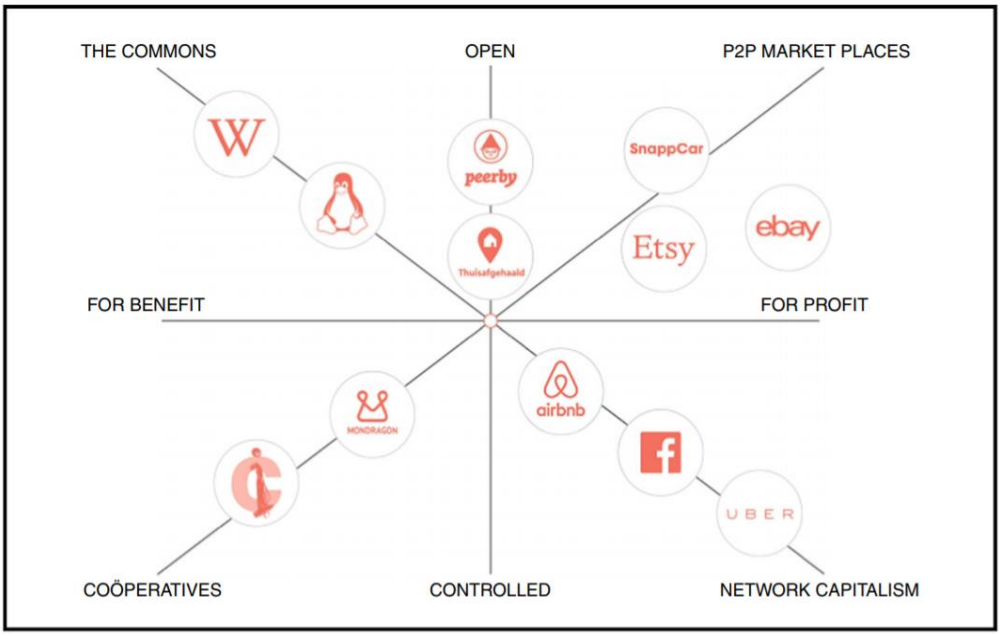

Figure 4: Types of value networks
Source: Oskam & Boswijk (2016, p. 25)

2.3 Online customer reviews and business

2.3.1 Online customer reviews

Before online opinion-sharing platforms appeared, consumers' biggest way of communication was the power of word-to-mouth (Dellarocas, Zhang & Awad, 2007). However, that changed for good. Online communities maintain their records for a very long time for everyone to see and are easily accessible, in contrast with word-to-mouth that keeps no records whatsoever and therefore is much harder to analyze.

Literature has a broad offer of studies on online customer opinions, with many results worth mentioning. Many of them, though, use social media platforms as datasets, as the growth of these platforms and the consequential enormous volume of data generated (in the form of users' opinions) coincides with the growing importance of sentiment analysis. Human behavior plays a big role in this, as our decisions are, more often than not, influenced by other people's opinions. This also applies to organizations, as they need to know what their costumers' (and potential customers) opinions are. As a result, sentiment analysis is being applied in almost every business and social domain (Liu, 2012). It is needed to extract valuable insights from large quantities of online reviews (in the case of product-reviewing online platforms), in order to make a classification of reviews regarding their polarity (positive or negative).

Li and Hitt even argued consumers can write biased reviews and therefore influence other consumers who read those reviews. These researchers suggest that the habit of being an early buyer can correlate with the likelihood of being satisfied with the reviewed products, either negative or positive, so even though a review is completely truthful as to what the review's writer feels about the product, it could be biased because of this factor. Their results (based on online book reviews and sales data collected from Amazon's website) find that, for the majority of the analyzed books, users' reviews posted in the early stages of the product's listing were systematically positively biased. They also say that later buyers tend to not set early reviews apart, so these positively biased reviews get a chance to influence later buyers' decisions (Li & Hitt, 2008).

2.3.2 Studies on online product reviews

Online reviews play a major role in reinforcing consumers' communication on a global scale and e-commerce companies tend to provide online platforms for them to share their experiences and therefore influence potential future buyers' decisions (Singla et al., 2017). All these online reviews, when properly analyzed by the platforms' owners, can be of great value to companies, who can use it to monitor consumers' behavior and their opinions on their products and services to promptly adapt their manufacturing, distribution and marketing strategies accordingly to the results they get (Dellarocas et al., 2007). Furthermore, the field of data analytics also stands as one of the main tools for these companies, as it can discover popular trends in consumers via software and sophisticated computational methods. Product designers can use this data to be provided with insights into product design, for example (Ireland & Liu, 2018).

Many studies have been done on online product reviews, with different intents. Some aim at analyzing consumer behavior, while others focus on the sentiment analysis aspect, reaching conclusions about the proposed sentiment analysis models they develop and propose.

Motion pictures

Dellarocas, Zhang and Awad's article (back in 2007), about motion pictures, showed how online reviews can be used in revenue forecasting, for example. Using data from Yahoo! Movies (movies.yahoo.com), BoxOfficeMojo (www.boxoffice mojo.com) and the Hollywood Reporter (www.holly woodreporter.com), they identified independent variables (such as Genre, MPAA ratings, prerelease marketing and availability, star actor appearance, professional critics, user reviews, release strategy and early box office revenues), external influences (such as the independent variables), internal influences (like word-to-mouth) and other factors and, using diffusion theory-based models, got their conclusions.

This study was able to show that an early volume of online reviews can be used as a proxy of early sales, allowing revenue forecasting to take place before early box office results are published. This implies that organizations can estimate their competitors' sales using online reviews as input. They also concluded about the average valence of user reviews as a predictor of the rate of decay of a movie's coefficient of external publicity and the about the importance of the gender entropy of online reviewers when predicting a movie's initial appeal.

Retailers

Using online reviews for approximately 4500 mobile phones from the Amazon platform (resulting in over 400,000 reviews), Singla, Randhawa and Jain's study proposed to classify them into positive and negative sentiments using, of course, sentiment analysis. They used three classifiers and all of them were machine learning-based (Support Vector Machine, Naïve Bayes and decision trees), concluding about their performances as well using 10-fold cross variation.

After collecting the data and preprocessing it, only three features of the products were used for the study (product name, brand name, and reviews). Afterward, using an inbuilt package of R, called "Syuzhet", they performed sentiment analysis and extracted eight

different emotions and their corresponding polarity, resulting in a total of ten emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive and negative). Finally, they tagged each review with its corresponding polarity (positive or negative), classify it using three above mentioned methods (SVM, Naïve Bayes, and decision trees) and tested their accuracies (10-fold cross variation). Not only did they find a significantly more positive polarity for the reviews (Figure 5) as they argued that the SVM classifier was the most accurate out of the three tested models (Tables 4 and 5).

Model Name	**Accuracy**
Naïve Bayes	66.95
SVM	81.77
Decision Tree	74.75

Table 4: Models' predictive accuracy
Source: Singla et. al (2017, p. 4)

Runs	**Naïve Bayes**	**SVM**	**Decision Tree**
1	67.11	80.12	74.31
2	64.57	79.44	77.13
3	67.57	78.96	70.80
4	66.77	82.25	78.00
5	67.57	80.52	72.63
6	64.57	78.92	76.22
7	67.77	77.70	71.84
8	66.71	79.72	77.62
9	68.31	78.00	67.68
10	68.57	81.75	81.25

Table 5: 10-fold cross variation
Source: Singla et al. (2017, p. 4)

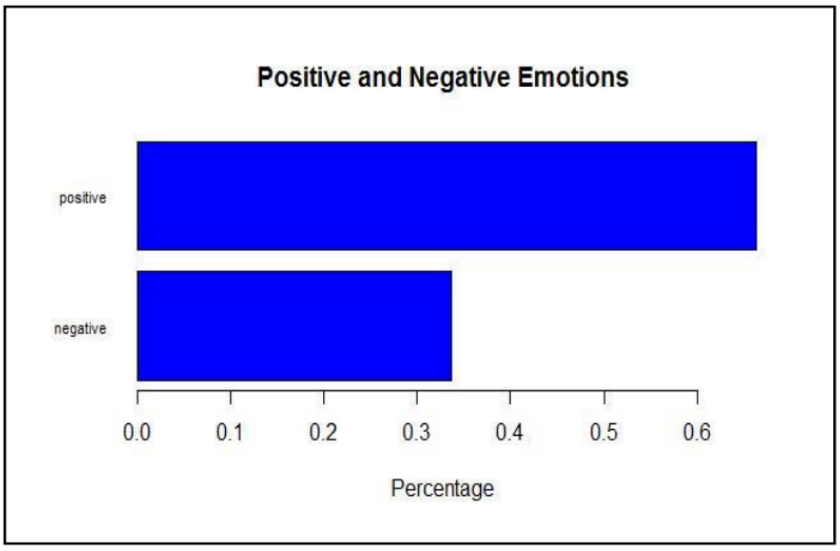

Figure 5: Percentage of positive and negative emotions
Source: Singla et. al (2017, p. 3)

2.3.3 Local accommodation online reviews

Like previously mentioned, accommodation companies are no exception to the online review phenomenon. As of 2015, the prior decade witnessed a massive growth in the use of global Internet-based reservation systems (Gössling & Lane, 2015). There are studies regarding Airbnb, Booking.com and other accommodation websites (hotels' online platforms, for example) that have contributed to the literature by focusing on the customers' online reviews, using, many times, sentiment analysis.

Online reviews on Booking.com and other agencies

Booking's website has the feature of posting reviews, just like many hotels' online platforms. Users have an experience when paying these companies for services and can latter post their opinions for other users to see them and take them into account when making their own decisions. Booking encourages and almost forces travelers to judge their experiences, either positively or negatively. They endorse critical perspectives explicitly and studies are proving that an increase in the customer review rates leads to an increase in sales when it comes to hotel rooms (Gössling & Lane, 2015).

These types of studies can reach conclusions in many aspects, like Chanwisitkul, Shahgholian and Mehandjiev's study did in 2018. They used Booking.com's online reviews of ten hotels located in the vicinity of "Khao San Road", in the center of Bangkok, Thailand and focused on an approach of topics, in this case. They evaluated the human resources aspect, complimentary service, room and bathroom interiors,

location, cleanliness and sleep quality, drawing conclusions on what customers are interested in when staying at a hotel (Chanwisitkul, Shahgholian, & Mehandjiev, 2018).

Sometimes, though, not even the full review is used to perform sentiment analysis. Another 2018 study focused on Booking's online reviews using only the titles of the testimonies. They concluded about the costumers' satisfaction using the titles' polarities, which were obtained by extracting data using WebHarvy software, preprocessing it and analyzing it using PLSA (Probabilistic Latent Semantic Analysis), which was used to calculate the probable polarity of words and documents from customer testimonials that the SentiWordnNet's library could not detect. They stated their model's accuracy was 76% and that their study could provide insights to business managers using sentiment analysis (Khotimah & Sarno 2018).

Finally, Ye, Law, Gu, & Chen also studied the influence user reviews can have on hotel room online sales, suggesting that influence is quite significant, confirming the importance of online word-of-mouth for the tourism industry. Thus, these researchers state hotel managers should be aware of the reviews that are posted on online travel agencies and consider them when making business decisions. All data was extracted from Ctrip.com, a major online travel agency in China, and the study says a ten percent increase in the ratings of user reviews could boost the index of online hotel bookings by five percent.

Online reviews on Airbnb

Millions of reviews are posted on Airbnb's online platform, enabling customers to make use of Airbnb's trusted community marketplace (Zervas, Proserpio, & Byers, 2015). User reviews are key references for potential travelers when it comes to compare and evaluate accommodation options, as they are an important tool to reduce risk when traveling to foreign places (Papathanassis & Knolle, 2011).

Guests do not use only reviews to guide them in their decision making, they also look for photos of the property they are looking to rent and search for a description of it, written by the landlord. When it comes to reviews though, while users look for reviews of the listing from other previous guests, hosts look not only for the reviews their potential renter has written in the past for other properties, as they also look for reviews that other hosts that have previously interacted with that potential renter have written (Chen & Chang, 2018).

But what factors do Airbnb users care about to the point that they mention them in their reviews? Generally speaking, the same factors usually associated with hotel stays, even though the order of importance might be different (Cheng & Jin, 2019). The same authors refer that the three main aspects found in Airbnb's online reviews are location, amenities, and host, as a result of using sentiment analysis on Sydney's properties registered in Airbnb's portal (this study's dataset). Figure 6 is a visual representation of the results, with the previously mentioned three main attributes (and their associated terms inside each zone) plus "recommend", which was the fourth most mentioned term.

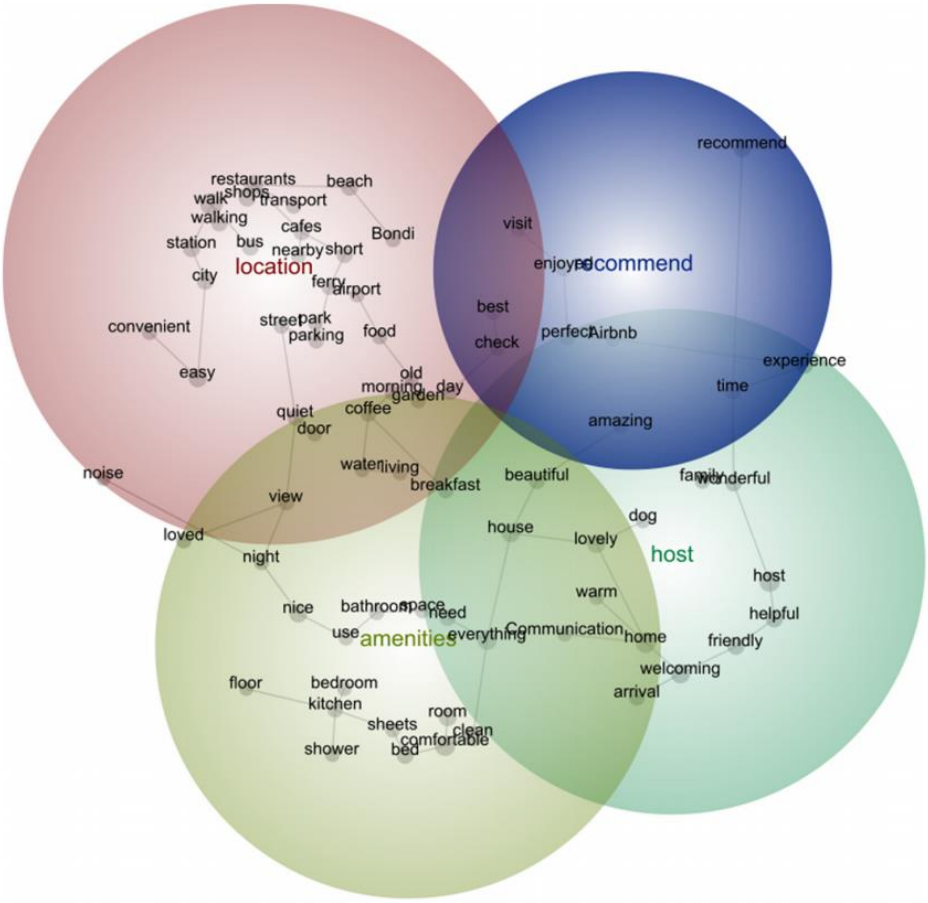

Figure 6: Visual representation of the most mentioned topics and their terms
Source: Cheng & Jin (2019, p. 63)

Other researchers have conducted studies about Airbnb online reviews, but with a different focus. As of 2015, 95% of all properties registered in Airbnb were rated, on average, 4.5 or 5 stars (the maximum rating) and almost no properties were rated lower than 3.5 stars. The same author even compared these ratings with TripAdvisor reviews and the average in the latter was 3.8 stars. Furthermore, comparing the properties that were mutually registered in Airbnb and TripAdvisor, they found that proportionally more properties had the highest ratings (4.5 and above) in Airbnb, even though the

average ratings were similar (Zervas et al., 2015). Bridges and Vásquez also studied this phenomenon, stating that negative aspects of the experience may be minimized by the user when writing the review, due to factors such as built trust between the guest and the host, reciprocity between them when it comes to reviewing and rating each other, lack of anonymity, politeness and possible review removal by Airbnb, if they violate their guidelines. This study used Airbnb listings in Portland, Albuquerque, Philadelphia, and Atlanta. While not explicitly using sentiment analysis, this study's methodology has some traces of it, like data collecting (it was performed manually) and the resort for software to analyze big data. In this case, it was used to measure the degree of word variation in a set of texts (AntConc was used to measure type-token ratios or TTRs) (Bridges & Vásquez, 2018).

2.4 Sentiments prediction models

A study in 2017 used decision trees to create a predictive model of sentiment in Booking.com and Tripadvisor.com retrieved data. After preprocessing the data (tokenization, stemming, stopwords removal and other methods), the investigator attributed a score to the data based on the created word vectors, using TF-IDF scores for each word in the word vector. The investigator used RapidMiner for the classification of the reviews, claiming it was a classifying algorithm very similar to CART and Quinlan's C4.5. Using three different training sets, one for each trained classifier, the investigator got very good results, considering the obtained overall accuracy levels were between 80.79% and 86.67% (Table 6) and between 80.62% and 85.90% for unseen data (Table 7) (Yordanova & Kabakchieva, 2017).

Results (%)	Unbalanced 4-600	Balanced 4-600	Balanced 20-200
Accuracy	86.67	85.53	80.79
Positive Class Recall	86.58	81.32	71.05
Negative Class Recall	86.84	89.74	90.53
F-measure	82.37	86.11	82.53
Positive Class Precision	92.15	88.79	88.24
Negative Class Precision	78.38	82.77	75.77
AUC	0.901	0.892	0.814

Table 6: Results from the Evaluation of the Trained Classifiers
Source: Yordanova & Kabakchieva (2017)

Results (%)	Unbalanced 4-600	Balanced 4-600	Balanced 20-200
Accuracy	85.46	80.62	85.90
Positive Class Recall	88.26	81.69	86.85
Negative Class Recall	42.86	64.29	71.43
Positive Class Precision	95.92	97.21	97.88
Negative Class Precision	19.35	18.75	26.32

Table 7: Results from the evaluation of the trained classifiers on unseen data
Source: Yordanova & Kabakchieva (2017)

2.5 Sentiments' impact on occupancy rates

In 2017, a study aimed to analyze the impact of sentiment polarity found in the hosts' descriptions of the online listings in occupancy rates. They calculated the polarity scores based on the AFINN-111 dictionary, which is a list of 2477 English words that are rated by sentiment, ranging from -5 to +5. Then, they multiplied the number of times each of those words appeared on a review and multiplied it by the dictionary's scores, subtracting the positive and negative results to obtain a final score. After these steps, correlation methods were used to find out the relationship between occupancy rates and

sentiments' polarities and, as revealed by Figure 7, there was not a significant correlation between the two variables (Martinez, Carrington, Kuo, Tarhuni, Abdel-Motaal, 2017).

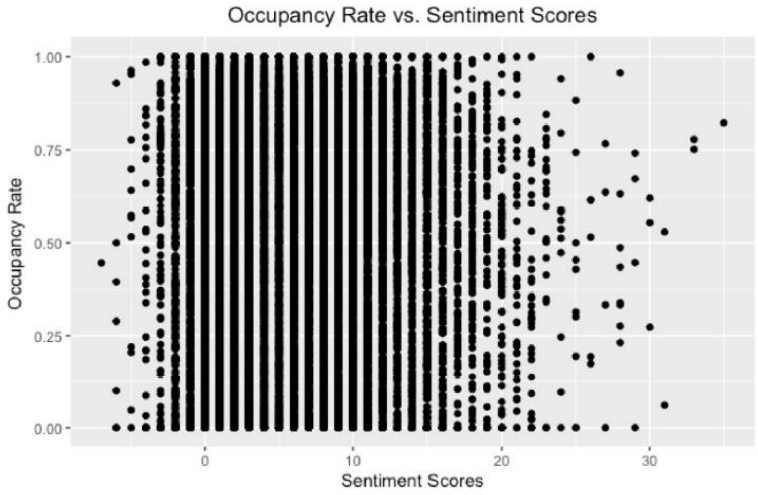

Figure 7: Correlation between occupancy rates and sentiments in listing descriptions
Source: Martinez et. al (2017)

Another study in 2017 analyzed several factors that could potentially influence occupancy rates and revenue per available room (RevPAR), getting information regarding those factors from 442 hotels' online reviews, in Switzerland. The variables belonged to three logical related areas to hotels, such as physical characteristics of the hotels (grounds, building, ambiance, rooms, and Internet), food and drink quality and service quality. TrustYou's software (a German company that offers online reputation management tools to the hospitality industry) was used to reach conclusions about the polarity (negative/positive) of the sentiments found in the reviews regarding the factors in scope, to further relate them with the dependent variables (occupancy rates and RevPAR) using partial least squares path modeling (PLS-PM) to get the R squared values of each one of the hotel related areas. A path coefficient of 0.907 was found for the Hotel category when it comes to positive influence over the dependent variables and 0.770 for negative influence, having the subcategory "Rooms" as its predominant factor (path coefficient of 0.649 and 0.715, respectively). Afterward, running Sobel tests, they realized Rooms, Internet and Building were all strong enough that, having "Hotel" as a mediator variable, they strongly influenced Demand (measured in occupancy rates), which was strongly related to Revenue (measured by RevPAR), as the PLS-PM analysis revealed ($\beta = 0.543$, $t = 14.286$, $p < .0015$) (Phillips, Barnes, Zigan, & Schegg, 2017).

(This page was intentionally left blank)

3. Methodology: CRISP-DM

As previously mentioned, CRISP-DM (Cross Industry Standard Process for Data Mining) will be the standard process used for the data mining that will occur throughout this work, due to it being more business-focused than other models used for data mining. Also, CRISP-DM can be applied to several business areas, such as financial, human resources, provided services, and others.

Four different software programs were used in the different CRISP-DM phases. Excel, Eclipse and ReviewPro were used for the data understanding and preparation phases, while IBM SPSS Statistics was used in the later phases, to model and process the previously treated data.

CRISP-DM englobes six main stages of work: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment (Chapman et al., 2000). Figure 8 illustrates these phases and how they all relate to each other.

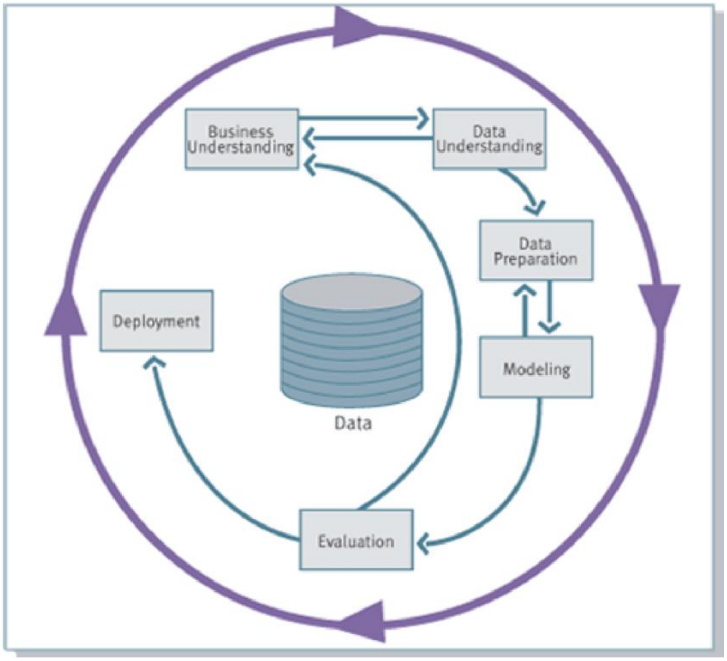

Figure 8: The six CRISP-DM phases
Source: Chapman et al. (2000, p. 13)

All these phases require different software and techniques, which will be described in the following sub-topics of this document. An FLH specialist (Francisco Cruz) and the supervisor of this work were consulted throughout all the phases as well, to fully achieve all the proposed objectives.

3.1 Business Understanding

In this phase, the main problem this work will try to solve has to be translated into a data mining problem, so that a solution can be proposed at the end of this document. In order to do this, the business goals and requirements have to be analyzed, leading to the definition of both criteria and a plan to successfully achieve all of the objectives (Chapman et al., 2000).

FLH is a Portuguese company that manages several properties in Portuguese territory, including Lisbon, Porto, Algarve, Ericeira, and Madeira (Feels Like Home, 2018). The problem that is being faced, though, is that they have no optimized way of analyzing the sentiments found in the reviews of the properties they manage and have a lot of uncertainties about what factors have the biggest impact on them.

The chosen approach in this investigation will focus on the impact of properties and reservations' characteristics on the sentiment polarity associated with the online reviews. This approach, in a final stage, will translate into a predictive classification model of the sentiment's polarity, aiming for an accuracy of at least 50%. This low value was established knowing previously the sample is too small to allow for the identification of many patterns in the data. Moreover, many of the properties have a low total number of reviews, limiting the available dataset regarding the sentiment polarity variables used to calculate the total polarity of each property. To complement the study, the relationship between sentiments' polarity and occupancy rates was described.

3.2 Data Understanding and Data Preparation

The Data Understanding phase consists of collecting the available data and analyzing it in order to get familiar with it and identify potential quality problems it might present. The possible hidden information is sought after through the creation of new hypotheses, which emerge by detecting interesting subsets in the data (Chapman et al., 2000).

The Data Preparation phase, on the other hand, consists of the set of activities that build the final dataset that will be used in the modeling phase, from the initial dataset. These activities are likely to be performed several times and not in any specific order. Some

of these activities include table, record and attribute selection as well as the transformation and cleaning of data for modeling tools (Chapman et al., 2000). In this phase, some variables had to be treated before they could be used to calculate new variables or to be sent to the final dataset.

There were two main sources of information throughout this work: Feels Like Home data tables and the available data from ReviewPro's API, regarding sentiments associated with the reviews of the properties. ReviewPro is a Guest Intelligence company that focuses on the hotel industry, creating cloud-based solutions that allow their clients to achieve a deep comprehension of the reputation's performance, as well as it allows them to identify both weak and strong topics about the client's operational processes and services. They offer insight to increase the client's satisfaction and score in evaluation pages, online tourism agencies and, finally, revenue (ReviewPro, 2019). FLH had access to their API for a limited amount of time and in that time frame, all the data used in this investigation was retrieved.

The chosen time interval the extracted data refers to was from 1st January 2017 to 31st May 2019. The reason behind the start date was the fact that FLH's manager, Francisco Cruz, specifically reported data anomalies and incoherencies in data prior to January of 2017. Because of this, no data before that date could ever be trusted as veridic and not harmful to the conclusions obtained in this work. On the other side of the spectrum, the end date was the date the files started being extracted from the database.

3.2.1 Feels Like Home dataset

Several excel tables were extracted from Feels Like Home's database. The manager created credentials for the researcher to be able to access a Platerit based platform, which made a lot of data available for this work. Tables regarding information about topics such as the properties and their characteristics, reservations (including average rates per night), occupancy rates of each house and guests' nationalities were extracted and analyzed, in order to extract valuable information with the intent of making a new Excel file that contains all the information in one table, to allow for an easy and fast usage of SPSS Statistics in later phases of this work. The following tables illustrate the extracted data from the available database.

Properties Table

This table contains information about each property (house), ranging from its location details to the number of extra beds, for example. Each extracted Excel file contained information about all the houses regarding a specific month in a specific year, which lead to the later junction of all these files (twenty-nine files, precisely, as there was one for each analyzed month of this research). Table 8 refers to the original table extracted from the database.

Original Column Name	Description	Nature	CRISP-DM usage	Justification for usage
ClientID	The ID of the house	Nominal	Data preparation	Used to relate data from Excel files to the correct house
Neighborhood	The neighborhood of the house	Nominal	Modeling	Used in objective 1
City	City of the house	Nominal	Excluded	All retrieved data was from Lisbon
Typology	Typology of the house	Nominal	Modeling	Used in objective 1
FloorNumber	Floor number of the house	Nominal	Modeling	Used in objective 1
DoubleBeds	Number of double beds	Scale	Modeling	Used in objective 1
SingleBeds	Number of single beds	Scale	Modeling	Used in objective 1
SofaBeds	Number of sofa beds	Scale	Modeling	Used in objective 1
ExtraBeds	Whether there are extra beds or not	Nominal	Modeling	Used in objective 1
Elevator	Whether there is an elevator or not	Nominal	Modeling	Used in objective 1
Parking	A detailed description of where is the private parking spot for that house	Nominal	Modeling	Used in objective 1
MaxOccupancy	Maximum number of occupants	Scale	Modeling	Used in objective 1
AptAlias	Online name of the house, when searched in accommodation online platforms	Nominal	Data preparation	Used to find the PID (ReviewPro id for each house), to extract sentiments
PostCode1	Post Code 1 of the house	Nominal	Excluded	More relevant data about the location was already obtained
PostCode2	Post Code 2 of the house	Nominal	Excluded	More relevant data about the location was already obtained
GeoLocation	Geo-coordinates of the house	Nominal	Excluded	More relevant data about the location was already obtained
Inactive	Whether the house is inactive or not	Nominal	Excluded	Being historical data or present data is irrelevant for the objectives
StartDate	The date the house was first available to rent	Nominal	Data preparation	Used to find out how many months of data would be analyzed for each house
InactiveDate	The date the house stopped being available to rent	Nominal	Data preparation	Used to find out how many months of data would be analyzed for each house

Table 8: Properties table

Reservations table

This table refers to information regarding the reservations' details, individually (each row is a reservation). Variables such as distribution channel used for the reservation and check-in dates and check out dates are included. Although there is a variable that refers to the total price of sale of the reservation and it could be used to get the average price per night, there was another file from the available data that contained that information already, eliminating the need to do further calculus to obtain this value. Although there is a variable that refers to the client's nationality for each reservation, there was another table with more detailed information regarding that topic. This time, only one file had to be extracted, as it already contained information about all the houses and all the relevant months to this work. Table 9 resumes the original information contained in the file.

Original Column Name	Description	Nature	CRISP-DM usage	Justification for usage
ClientID	The ID of the house	Nominal	Data preparation	Used to relate data from Excel files to the correct house
AccessID	The ID of the reservation	Nominal	Excluded	Irrelevant to the objectives
CheckInDate	Check-in date of the reservation	Nominal	Data preparation	Used in calculations of other variables
CheckOutDate	Check out date of the reservation	Nominal	Excluded	Irrelevant to the objectives
GuestCountry	Client's country	Nominal	Excluded	More detailed data can be found in another table
NumOfGuests	Number of guests in the reservation	Scale	Excluded	More detailed data can be found in another table
Canceled	Whether or not the reservation was canceled	Nominal	Data preparation	Used in calculations of other variables, as a filter
ReservationDateCreated	The date the reservation was created	Nominal	Excluded	Check-In Date was more relevant
ReservationSoldValue	The total price of the reservation	Scale	Excluded	More detailed data can be found in another table
City	City of the booked house	Nominal	Excluded	All retrieved data was from Lisbon

Table 9: Reservations table

Aggregated reservations table

Another table regarding reservations was used, as it contains information displayed in an aggregated way when compared to the previous reservations table. While the previous had a line for each reservation made, having the house in question as one of the columns, each line of this table contains one house and the columns have the total and average values for each of the variables. Once again, just like in the properties' table, one file had to be retrieved from FLH's database for each of the months in the chosen time period. After that task, all the files had to be joined in the same Excel table, to facilitate the transition of the values to the final Excel table. The following table illustrates an example of the original tables, retrieved from the database.

Original Column Name	Description	Nature	CRISP-DM usage	Justification for usage
ClientID	The ID of the house	Nominal	Data preparation	Used to relate data from Excel files to the correct house
AptAlias	Online name of the house, when searched in accommodation online platforms	Nominal	Data Preparation	Used to find the PID (ReviewPro id for each house), to extract sentiments
ClientName	Name of the owner of the house	Nominal	Excluded	Irrelevant to the objectives
Typology	Typology of the house	Nominal	Modeling	Used in objective 1
MaxGuest	Max Occupancy of the house	Scale	Excluded	Already had this variable in the properties table
NumReservations	Number of reservations made that month	Scale	Excluded	Irrelevant to the objectives
TotalValue	Sum of all the sell values, of all reservations, that month	Scale	Excluded	Irrelevant to the objectives
TotalDays	Sum of days of all reservations	Scale	Excluded	Irrelevant to the objectives
TotalGuests	Sum of the number of guests of all reservations	Scale	Excluded	Irrelevant to the objectives
AvgDuration	The average duration of the stays, in days	Scale	Modeling	Used in objective 1
AvgRate	Average rate per night for that month	Scale	Modeling	Used in objective 1
AvgGuests	The average number of guests for that month	Scale	Modeling	Used in objective 1

Table 10: Aggregated reservations table

Occupancy rates table

This table, as the name indicates, contains the details about the occupancy rates of every house, for the selected month when retrieving the file from the database. This table is especially important for objective 2, as it gives us the required values to use regarding occupancy rates. Several tables were merged into one, once again.

The Relative Occupancy variable had to be created (dividing the number of days occupied by a customer by the number of days the house was available in that month, subtracting the number of days the owner occupied the house), to avoid rates that were too low because the owner of the house stayed there for a long period of time. For example, assuming 30 days in a month, if a house was occupied by a customer for 2 days and the owner occupied it for the other 28, this variable returns 100% ((2/2) *100, instead of 6,67% ((2/30) *100). The following table represents an example of one of those Excel tables downloaded from the database.

Original Column Name	Description	Nature	CRISP-DM usage	Justification for usage
ClientID	The ID of the house	Nominal	Data preparation	Used to relate data from Excel files to the correct house
AptAlias	Online name of the house, when searched in accommodation online platforms	Nominal	Data Preparation	Used to find the PID (ReviewPro id for each house), to extract sentiments
ClientName	Name of the owner of the house	Nominal	Excluded	Irrelevant to the objectives
Typology	Typology of the house	Nominal	Modeling	Used in objective 1
NumReservations	Number of reservations made that month	Scale	Excluded	Irrelevant to the objectives
TotalDays	Number of days the house was occupied by a costumer	Scale	Data Preparation	Used to calculate relative occupation rates
TotalExtraDays	Number of days the house was occupied by the owner	Scale	Data Preparation	Used to calculate relative occupation rates
CalendarDays	Number of days the house was available in a month	Scale	Data Preparation	Used to calculate relative occupation rates
Occupation	Number of days occupied by customer/Total days in the month	Scale	Excluded	Irrelevant to the objectives
RelativeOccupation	(Number of days occupied by customer/Number of days the house was available to costumers) *100	Scale	Modeling	Used in objective 2
TotalOccupation	Number of days the house was occupied/Number of days in the month	Scale	Excluded	Irrelevant to the objectives

Table 11: Occupancy Rates table

Countries table

This table contains the data regarding the country that represents the majority of the reservations' revenue for each month. Once again, several files had to be extracted from the database, as each one of them had information only for a specific month. The reason a file containing all the months at once could not be downloaded is that then there would be no way of knowing for what month was each set of rows referring to. The following table illustrates the data from one of these tables.

Original Column Name	Description	Nature	CRISP-DM usage	Justification for usage
CountryCode	Country Code (e.g.: PT, FR, BR, EUA)	Nominal	Data preparation	Used to identify the country with the highest revenue percentage
Country	Name of the country	Nominal	Data Preparation	Used to identify the country with the highest revenue percentage
NumReservations	Number of reservations made that month	Scale	Modeling	Used in objective 1
TotalValue	Revenue generated by customers from that country	Scale	Excluded	More detailed data can be found in another variable
TotalDays	Total of days costumers from that country occupied the properties	Scale	Excluded	More detailed data can be found in another variable
AvgRate	The average price of the reservations by costumer from that country	Scale	Excluded	Irrelevant to the objectives
PercentOfTotal	Total revenue generated by customers of that country/Total revenue that month, by all properties	Scale	Data Preparation	Used to obtain the country with the highest percentage of occupation

Table 12: Countries table

Relationship between the retrieved tables

The following diagram showcases the extracted tables and the relationships among them, in order to fully understand how the data was prepared and how the different variables were connected.

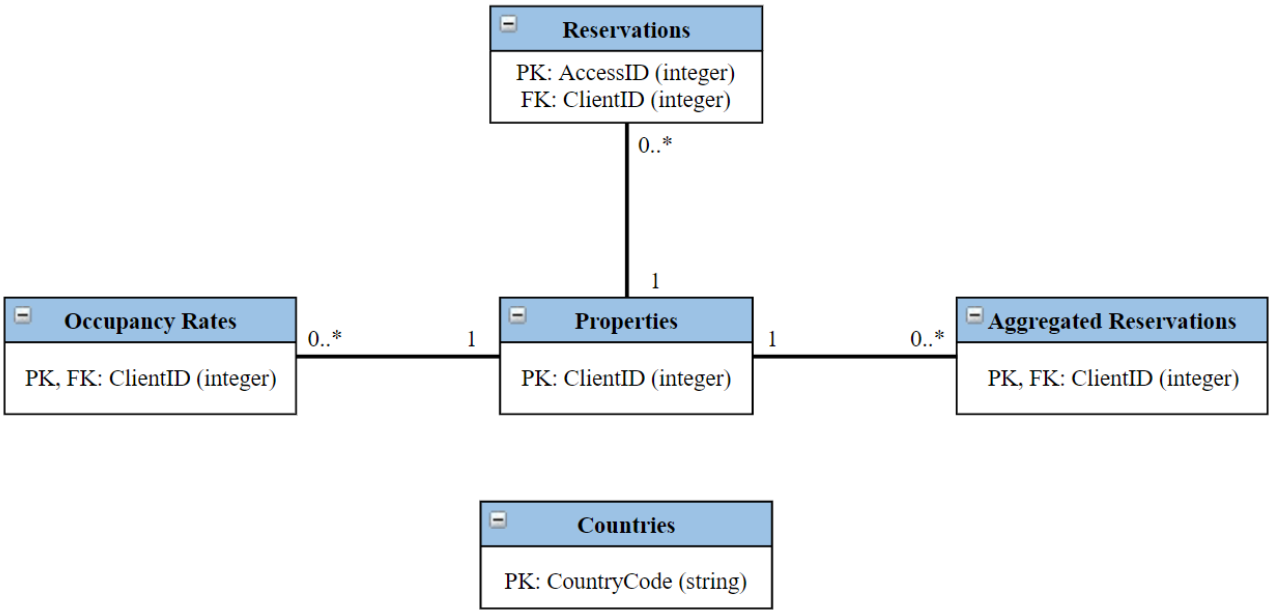

Figure 9: FLH's tables diagram

3.2.2 ReviewPro dataset

The other main source of information to include in the available dataset was ReviewPro and its sentiment extraction features. Their API has a lot of different functions that return different outputs, depending on what the user intends to do with the extracted data. However, the common thing about all the different functions and their outputs was the format in which the results were presented to the user: JSON. JSON is a lightweight data-interchange format and is easy for humans to read and write, as well as for machines to parse and generate. Based on a subset of the JavaScript Programming Language Standard ECMA-262 3rd Edition - December 1999, JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python, and many others (JSON, 2019).

The chosen function was the "Semantic Mentions", as its output was the most suited to this work. When used, a textbox with JSON notation written can be seen. This JSON formatted output returns, for the chosen PID (property's ID for ReviewPro), the number of negative and positive mentions for each of the terms present in the ReviewPro's dictionary. This process was very difficult to automatize, as the only way to choose a specific house and time interval was to insert the from and to dates in a textbox, as well as the PID.

This limitation had the consequence of only being able to retrieve information for sixty properties, as each one could take up to 30 minutes to retrieve all the JSON files needed, in the established time interval of analysis, for reasons described below.

Eclipse (Java)

A program had to be developed to analyze all the retrieved data from ReviewPro. A folder was created to store a folder per property. Then, each one of the properties' folders contained all the JSON files that ReviewPro returned for that house that were not empty (no mentions to that house, in any of the categories, that month).

The process to create and store the JSON files consisted of inserting the parameters in ReviewPro's API (PID, from and to dates), copy the returned text, create a new file in the property's folder (in Eclipse), name it the month and year it referred to and paste the text inside the file. Each house could potentially make this process be repeated twenty-nine times (the number of analyzed months, throughout this work), making it a very long process. Eventually, the number 60 was fixed as the number of retrieved houses, due to time constraints.

After retrieving all the data, the developed program would analyze the files, check each one of the concepts (words found in the ReviewPro's dictionary), check what sub-category the concept belonged to (using a file extracted from the API that contains all the sub-categories and their concepts) and iterate the corresponding category's counter (either positive counter or negative counter). These categories are "Location" (only the "Location" sub-category was chosen), "Amenities" (the "Internet", "Facilities". "Maintenance", "Cleanliness", "Ambience", "Bathroom" and "Decoration" sub-categories were chosen), "Room" ("Room" and "Bed" sub-categories") Host (only the "Service" sub-category was chosen) and "Value" (only the "Value" sub-category was chosen).

The final output was returned by the console (example of it in Figure 10), mentioning for each of the categories the number of positive and negative mentions. Then, all this data was manually put in the final Excel file (the file that was used as input for SPSS

Statistics), making for another very long process, as there were more than 1100 entries to be filled[1].

```
Date: 2019-04                    Date: 2019-05
From: 2019-04-01                 From: 2019-05-01
To: 2019-04-30                   To: 2019-05-31

Location NEGATIVE: 0             Location NEGATIVE: 0
Location POSITIVE: 0             Location POSITIVE: 4

Amenities NEGATIVE: 12           Amenities NEGATIVE: 4
Amenities POSITIVE: 2            Amenities POSITIVE: 2

Room NEGATIVE: 14                Room NEGATIVE: 1
Room POSITIVE: 2                 Room POSITIVE: 3

Host NEGATIVE: 0                 Host NEGATIVE: 0
Host POSITIVE: 0                 Host POSITIVE: 0

Value NEGATIVE: 2                Value NEGATIVE: 0
Value POSITIVE: 0                Value POSITIVE: 0
```

Figure 10: Example of output of the Java program

3.2.3 Final table and new variables

This table was where the majority of the Data Preparation phase occurred. Due to its size, it can be viewed in Appendix A. New variables were created to use in the statistical models and tests, such as the sentiment polarity variables used throughout the models' constructions and the univariate and bivariate analysis. Four variables were created to assess sentiment: "Total Polarity" (negative, neutral or positive), "Last Six Months Total Polarity" (negative, neutral or positive), "Positive Sentiment" (yes or no) and "Last Six Months Positive Sentiment" (yes or no). These were all obtained from the scores' variables previously calculated, which were given by subtracting positive mentions and negative mentions, both for only the month in question as for the previous six months (the accumulated variables). Besides these variables, some new others were added, such as "Season", "High Season" and "Holiday". These can all be found in appendix A, alongside their descriptions.

The final table is the Excel table that was used as input for SPSS Statistics, to complete the methodological approach of this work. It can only be called the final table now, as

[1] Code can be obtained at https://github.com/drsfa1/Sentiment-analysis-in-online-customer-reviews-The-Feels-Like-Home-case.git

this document is being written, because throughout this work the table suffered multiple changes, as CRISP-DM allows for this kind of agility of going back to data understanding and preparation phases after starting the modeling phase. This final version contains 1131 entries of data.

Only the variables that served as input to SPSS were included in this table, as some of them were merely used as auxiliary variables in data preparation, such as "Date", "Concat" (concatenation of date with property ID) and the variables regarding sentiments that were used to calculate polarity scores.

3.3 Modeling

3.3.1 Descriptive Analysis, Univariate Analysis, and Bivariate Analysis

Descriptive statistics consist of methods for organizing, displaying and describing data through the use of tables, graphs and summary measures (Laureano & Botelho, 2017). The objective of this type of statistic is to provide simple summaries, either quantitative (summary statistics) or visual (for example easy-to-interpret graphs). While usually it is used as an early stage of a bigger investigation, it can sometimes be enough for an entire investigation.

Univariate analysis is the simplest form of descriptive analysis, as it only deals with one variable and provides summarized data about it. Its major purpose is to describe variables, as opposed to explaining relationships between two or more variables.

The univariate analysis for each of the variables' themes (properties' characteristics, reservations' characteristics, and sentiment variables) is described in the Results section, while the tables found in appendix B represent the univariate analysis of all the different types of variables present in the data set.

The main difference between bivariate and univariate analysis is that while univariate only describes one variable in an easy to interpret way, bivariate attempts to describe relationships between two variables. It analyses how two variables simultaneously change together, analyzing if there is a relationship, and, if there is, its direction and intensity.

Throughout this investigation, several variables are formulated into hypothesis, so that afterward, significance tests can be performed to assess the validity of the formulated hypothesis**.** The type of test used depends on the type of data, the form of the population distribution, the method of sampling and sample size**.** The independence tests were separated into objectives (characteristics of the properties crossed with sentiments and reservations' characteristics crossed with sentiments).

Usually, these tests compare the hypothesis proposed by the researcher to the null hypothesis, which represents precisely the opposite of what the researcher thinks is true. After testing the null hypothesis, the test returns a value that determines the likelihood of the null hypothesis being correct. If this value is superior to the established significance level (usually 0.05), the null hypothesis is confirmed and the original hypothesis, proposed by the researcher, is dropped (Laureano, 2013).

Bivariate analysis with sentiment polarity

The bivariate analysis will describe the relationship between two variables. The existence of a significant relationship is given by the independence tests, while the strength of that relationship is given either by Crámer's V or Spearman's correlation test, depending on what type of variables we are trying to test. The main variable we want to test here is, on the Y-axis, the sentiment's polarity. The chosen variable to do this analysis was "Total Polarity", which assumes the values "Negative", "Neutral" and "Positive".

Chi-squared test of independence

In this investigation, the used hypothesis test was the Chi-Squared test (considering a significance level of 0.05), due to it being the most suitable to work with categorical variables. In this study the variable that expresses sentiment for each house is either ordinal or nominal, depending on the phase of the work and the obtained results. Either way, it will always be categorical. Due to this, if the other variable is categorical the test is ready to be made, while if it is quantitative it only needs to be treated as an ordinal variable for this test to be suitable.

Association metrics

To assess the association/correlation between two variables, two different indicators had to be used, depending on the case (which depends on what type of variables are being used).

Cramér's V and Spearman's rank correlation coefficient

When analyzing two nominal variables, their association was measured using Cramér's V. This measure ranges from 0 (absence of relationship) to 1 (perfect relationship) and indicates the intensity of the relation between the two variables. When the two variables are ordinal or one is ordinal and the other is quantitative, the Spearman's correlation coefficient can be used as a measure of the correlation between the two variables. This measure ranges from -1 (negative perfect correlation) to 1 (positive perfect correlation), while 0 value means an absence of correlation.

Finally, to compare the mean values of the occupation rate among the three categories of the sentiments a one-way ANOVA was performed (considering a significance level of 0.05) and, since no significant differences were found, the post-hoc Scheffé test was also performed.

3.3.2 Predictive Model

One of the main objectives of this work is to define a predictive model of the sentiments associated with each house, monthly. This phase will, in theory, give us the factors about the properties and reservations that can explain the most the fluctuance observed in the sentiments for each house, in order to reach results that will be discussed with FLH's manager, to understand the results from a business perspective.

Decision Trees

The chosen predictive model was a decision tree. Decision trees are one of the most popular techniques, both for classification (when the target variable is qualitative) and regression (when the target variable is quantitative) models, due to its easy to understand output and being easy and quick to perform. It divides a training set in examples of a class, in a recursive manner. There's a point of division in each node of the tree, consisting of a test on one or several attributes to determine how the data should be subdivided, to create two more nodes. Each node is used to increase its generalization and its precision of prediction in the given training dataset. The division of the data,

made recursively, ends when the division is considered pure or small (Turban, 2011). Several factors can be analyzed to determine the tree's complexity, such as the total number of nodes, number of leaves, tree depth and the number of analyzed attributes (Maimon & Rokach, 2010).

CART decision tree

A CART decision tree was the chosen predictive model for this work, as it is one of the most popular classification predictive models used in investigations. It can be found in almost any area, such as financial research and medical topics, is easy to implement and is provided by lots of code developing platforms such as Matlab, which is easy to be further implemented to compare with other methods. Finally, the results presented by this type of tree are easier for humans to interpret than other models and present advantages over some of them, like for example the capability of modeling complex relationships between independent and dependent features in the task without strong model assumptions (Li, Sun, & Wu, 2010).

CART's parametrizations

Parametrizations (the process of defining parameters) play a big role in decision trees, as they allow different trees to be tested as well as achieving different results. The parametrizations used in this work for the CART decision tree include the tree's depth, number of cases for parent and child nodes and the prior probabilities of each of the sentiment polarity's values. The usage of prune to avoid overfitting (when the model presents a good performance in the training dataset and poor performance in the testing dataset) was used as well, but it was immediately obvious it was consistently presenting worse results.

Many different parametrizations were tested, generating several different models. The "Positive Sentiment" variable was the influenced variable, as it was more relevant to the business. Only the best model was chosen due to its reliability for FLH to make decisions based on it. Table 13 contains the five parametrizations that generated the best CART decision trees.

	Model A	Model B	Model C	Model D	Model E
Tree depth	7	4	6	7	6
Minimum records in parent node	4	2	10	10	2
Minimum records in child node	2	1	5	5	1
Prior probabilities	Training sample	Training sample	Equal	Training sample	Training sample

Table 13: CART parametrizations

3.4 Evaluation

In the Evaluation phase, the models and their construction are evaluated, to ensure no mistakes were made. Only after this phase is done can the investigator make decisions regarding the results obtained in the research (Chapman et al., 2000).

3.4.1 Evaluation metrics

Confusion matrix

The tested predictive models have proper metrics to be evaluated. To calculate them, we'll use confusion matrices. A confusion matrix (or classification matrix) is a tool to evaluate classification models' performances. As we can see in the example confusion matrix in Table 14 there are two main entries: the predicted class (in this case, sentiment) by the model and the observed class (Delen, Kuzey, & Uyar, 2013). While true positives (TP) are positive tuples correctly classified by the model, the false positives (FP) are incorrectly classified negative tuples (which means they were negative but the model predicted they would be positive). The same logic applies to the other two quartiles, as false positives (FP) are negative tuples that were classified as positive and true negatives are negative tuples that the model predicted correctly.

In this study, positive sentiment was considered a successful case, while non-positive sentiment was considered an unsuccessful case.

		Predicted sentiment	
		Successful (positive sentiment)	**Unsuccessful (non positive sentiment)**
Observed sentiment	**Successful (positive sentiment)**	True Positives (TP)	False Negatives (FN)
	Unsuccessful (non positive sentiment)	False Positives (FP)	True Negatives (TN)

Table 14: Confusion Matrix

For this study, the used evaluation metrics were overall accuracy, specificity, and sensitivity. These were the used metrics to evaluate all the built models.

Accuracy

Overall accuracy represents the total percentage of cases that are correctly classified by the model. It is a ratio between correctly predicted cases and the total number of cases.

$$\boldsymbol{Accuracy} = \frac{True\ positives + true\ negatives}{Total\ cases}$$

Specificity

Specificity represents the proportion of negative tuples that are correctively identified when compared to the total of negative tuples.

$$\boldsymbol{Specificity} = \frac{True\ negatives}{True\ Negatives\ + False\ Positives}$$

Sensitivity

Sensitivity is the proportion of true positives when compared to the total of positive tuples.

$$\boldsymbol{Sensitivity} = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

3.4.2 Validation method

In order to validate the created models, two techniques can be used to assess the model's predictive capability. The holdout method separates the dataset into two samples. One is used to train the model (around two-thirds of the data) and the other one is used to test the model (the remaining of the data). When the sample is considered to be small, the cross-validation method is used (Li et al., 2010).

K-fold cross-validation is a variation of the holdout method, where the data is separated in k different subsets with the same size. One of the subsets is used for testing while the rest of them is used for training. All the subsets end up being used for testing, as this is a circular process (Kohavi, 1995). This is the best method when the dataset size is small, as is the case of this work (1131 entries of data). K-fold cross-validation is the technique used in this work, using k-values of 10 and 20.

After all the different models are tested using this validation method, the best one (the one presenting the highest Cross-Validation accuracy) will be showcased to FLH in order to get more detailed conclusions about what measures they should take to make the most out of this work.

3.5 Deployment

This phase will be done by the deployment of this thesis. It would be important that Feels Like Home develops an implementation of the proposed models in this thesis so that the results and conclusions can be of good use to them, as well as for the rest of the local accommodation industry

(This page was intentionally left blank)

4. Results and Discussion

In this chapter, all the obtained results from the investigation will be displayed and analyzed. This discussion includes not only the investigator's interpretations of the results as it includes FLH's manager insights and interpretations of the potential results' usage as a contribution to FLH.

4.1 Sample characterization

4.1.1 Properties' characteristics

Table 15 allows for a quick analysis of the dataset, as it is simple to interpret. Most of the dataset refers to historical neighborhoods (95.6%), while the touristic distribution is much more even (only 46.8% are considered touristic). Also, most entries have a T2 typology and the houses with no double beds are a minority (4.5%), being 1.2 the mean value for the number of double beds in each of the dataset entries. Another thing to notice is that 64.4% of the analyzed properties do not have an elevator and 87.5% do not have a dedicated parking spot, numbers that could be explained by how old a large portion of Lisbon's downtown buildings are (and some of FLH's properties being villas instead of flats) and the fact that Lisbon has a high population density (and therefore less space to have dedicated parking spots), respectively.

		Count	%	Mean	Minimum	Median	Maximum	Standard Deviation
Historical Neighborhood	No	50	4.40%					
	Yes	1081	95.60%					
Turistic Neighborhood	No	602	53.20%					
	Yes	529	46.80%					
Couple	No	1001	88.50%					
	Yes	130	11.50%					
Typology	T0	108	9.50%					
	T1	225	19.90%					
	T2	582	51.50%					
	T3	170	15.00%					
	T4	46	4.10%					
Floor Number	-1	27	2.40%					
	0	135	11.90%					
	1	211	18.70%					
	2	305	27.00%					
	3	214	18.90%					
	4	94	8.30%					
	5	92	8.10%					
	6	20	1.80%					
	8	13	1.10%					
	12	20	1.80%					
Double Bed	No	51	4.50%					
	Yes	1080	95.50%					
Single Bed	No	500	44.20%					
	Yes	631	55.80%					
Sofa Bed	No	680	60.10%					
	Yes	451	39.90%					
Extra Bed	No	814	72.00%					
	Yes	317	28.00%					
Elevator	No	728	64.40%					
	Yes	403	35.60%					
Parking	No	990	87.50%					
	Yes	141	12.50%					
Max Occupancy				4.4	4	7	2	1.4
N° of Double Beds				1.2	1	3	0	0.5
N° of Single Beds				1.3	2	5	0	1.3

Table 15: Descriptive statistics of the properties' characteristics

Note: 1131 observations.

Looking at table 16, regarding the neighborhoods, the immediate conclusion is that the most represented neighborhood is Bairro Alto, representing 18.5% of the analyzed dataset, while the least represented one is Ajuda, appearing only in 0.9% of the final data used for this investigation. Santa Catarina (13.7%), Baixa (11.3%) and Martim Moniz (9.1%) are also highly represented in the dataset.

		Count	%
Neighborhood	Ajuda	10	0.90%
	Alfama	78	6.90%
	Av. da Liberdade	24	2.10%
	Av. Novas	65	5.70%
	Bairro Alto	209	18.50%
	Baixa	128	11.30%
	Campo de Ourique	40	3.50%
	Chiado	14	1.20%
	Estrela	14	1.20%
	Expo	20	1.80%
	Graça	16	1.40%
	Intendente	18	1.60%
	Lapa	23	2.00%
	Marquês de Pombal	31	2.70%
	Martim Moniz	103	9.10%
	Mercês	13	1.10%
	Principe Real	60	5.30%
	Restelo	20	1.80%
	S. Bento	25	2.20%
	S. José	20	1.80%
	Santa Catarina	155	13.70%
	Sé	45	4.00%

Table 16: Descriptive statistics of the properties' neighborhoods

Note: 1131 observations.

4.1.2 Reservations' characteristics

Table 17 describes the univariate statistics of the reservations' characteristics available in the dataset.

		Count	%	Mean	Minimum	Median	Maximum	Standard Deviation
Main Nationality	Brazil	86	7.60%					
	France	550	48.60%					
	Germany	45	4.00%					
	Spain	409	36.20%					
	United Kingdom	41	3.60%					
Main Distribution Channel	Airbnb	65	5.70%					
	Booking	993	87.80%					
	Feelslikehome	3	0.30%					
	None	66	5.80%					
	Other	4	0.40%					
Average Advance Booking				2.8	0	2.5	9	1.3
Average Number of Guests				3.6	1	3.6	7	1.1
Average Stay Duration				3.2	1	3	17	1
Price per Night				102.6	30.6	95.6	423.1	41.7

Table 17: Descriptive statistics of the reservations' characteristics

Note: 1131 observations.

Once again, quick conclusions can be drawn from the table. From the dataset, the main nationalities (when it comes to generated revenue) of costumers are France (48.6%) and Spain (36.2%) and the main distribution channel (the platform the client used to make the reservation) was Booking.com. Besides that, the mean values are 2.8 months for average booking advance (how many months between the reservation date and the

check-in date), 3.6 guests for the average number of guests and 3.2 days for the average number of days of the reservations.

4.2 Sentiments assessment

Table 18 contains information about the sentiments and their ratios between positive, neutral and negative sentiments, as well as for only positive and non-positive sentiments (last row). Total polarity is obtained subtracting the number of negative mentions of the property to the positive mentions of the property.

		Count	%
Total Polarity	Negative	400	35.40
	Neutral	77	6.80
	Positive	654	57.80
Positive Sentiment	No	477	42.20
	Yes	654	57.80
Last 6 Months Polarity	Negative	370	32.70
	Neutral	25	2.20
	Positive	736	65.10
Last 6 Months Positive Sentiment	No	395	34.90
	Yes	736	65.10

Table 18: Descriptive statistics of the sentiments
Note: 1131 observations.

Looking at the table, we can quickly conclude that for the most part, the properties in the dataset have positive sentiments attached to them. 57.8% of the entries in the dataset have positive polarities, against 42.2% of non-positive (negative or neutral) polarities. When looking at the aggregated score of the previous six months the numbers are even more promising, getting 65.1% positive polarities and 34.9% non-positive polarities.

To get a better understanding of how the shown values for positive polarity are originated, a classification CART decision tree was created to identify what categories of sentiments are the most important to the overall polarity. Table 19 displays the importance of each of the independent variables to the calculation of the "Positive Sentiment" variable, for example[2].

[2] The tree was obtained using the default parametrization of SPSS Statistics (except for crossvalidation) and got an accuracy level of 92.5% in training (resubstitution) and 90.1% in test (crossvalidation 20-fold).

Independent Variable	Importance	Normalized Importance
Room NEG	0.128	100.00%
Room POS	0.122	96.00%
Amenities NEG	0.121	94.80%
Amenities POS	0.093	73.20%
Location POS	0.08	62.90%
Amenities Mentions	0.075	59.00%
Location Mentions	0.068	53.30%
Room Mentions	0.066	51.40%
Location NEG	0.029	22.80%
Value NEG	0.024	18.50%
Host POS	0.019	15.20%
Value POS	0.013	10.10%
Host Mentions	0.012	9.60%
Value Mentions	0.012	9.40%
Host NEG	0.011	9.00%

Table 19: Independent Variable Importance

The variables related to rooms (both negative and positive mentions), amenities (both negative and positive mentions) and location (positive mentions) seem to be the most impactful in the calculation of the overall score for each entry of the dataset, looking at their normalized importance. The total number of mentions related to amenities and location also has relevant importance (59.0% and 53.3%). This table did not surprise the FLH's manager, but on the other hand, it confirmed what FLH already thought about the importance of each category's positive or negative mentions to the calculation of an overall score. However, the very different impacts of positive and negative mentions about location on the calculation of the score did bring some insight. While the negative mentions have a normalized importance of (22.8%), the positive ones have a 62.9% normalized importance. Even though people sometimes leave bad mentions about location, its impact on the overall sentiment score of the house is not nearly as big as the impact of the positive ones.

Taking a more detailed look at these categories and their variables, Table 20 gives us a better perspective of how each category performs in FLH's online reviews. This table surprised FLH's manager, as he did not expect so many negative reviews regarding the "Room" category. He also did not quite understand the number of negative scores related to amenities, but upon talking to a person in FLH that was more dedicated to sentiment assessment than him, we quickly understood that many times what is preventing perfect or very good scores are precisely the amenities, which can explain the high percentage of negative scores (38.64%) in this category.

		Count	%	Sum
Room	No mentions	141	12.47	
	Negative	413	36.52	
	Neutral	101	8.93	
	Positive	476	42.09	
Room Mentions				6950
Amenities	No mentions	172	15.21	
	Negative	437	38.64	
	Neutral	97	8.58	
	Positive	425	37.58	
Amenities Mentions				6505
Location	No mentions	234	20.69	
	Negative	69	6.1	
	Neutral	52	4.6	
	Positive	776	68.61	
Location Mentions				3923
Host	No mentions	1045	92.4	
	Negative	38	3.36	
	Neutral	5	0.44	
	Positive	43	3.8	
Host Mentions				192
Value	No mentions	863	76.3	
	Negative	126	11.14	
	Neutral	19	1.68	
	Positive	123	10.88	
Value Mentions				640

Table 20: Categories' polarity and mentions

Note: 1131 observations.

4.3 Relationship between sentiments and the characteristics of the properties

After performing the bivariate analysis, only two variables related to the physical characteristics of the houses and their locations revealed to have a significant relationship with the sentiment's polarity found in the properties' mentions in the online reviews, as we can see in Table 21.

	Total Polarity							
	Negative		Neutral		Positive		Total	
	Count	%	Count	%	Count	%	Count	%
Elevator (χ2(2)= 10.724; p= 0.005; Cramer V= 0.058)								
No	282	38.7	50	6.9	396	54.4	728	100.00
Yes	118	29.3	27	6.7	258	64	403	100.00
Neighborhood (χ2(42)= 68.023; p= 0.007; Cramer V= 0.173)								
Ajuda	3	30.00	0	0.00	7	70.00	10	100.00
Alfama	26	33.30	0	0.00	52	66.70	78	100.00
Av. da Liberdade	2	8.30	1	4.20	21	87.50	24	100.00
Av. Novas	25	38.50	3	4.60	37	56.90	65	100.00
Bairro Alto	89	42.60	13	6.20	107	51.20	209	100.00
Baixa	36	28.10	15	11.70	77	60.20	128	100.00
Campo de Ourique	13	32.50	5	12.50	22	55.00	40	100.00
Chiado	2	14.30	0	0.00	12	85.70	14	100.00
Estrela	6	42.90	0	0.00	8	57.10	14	100.00
Expo	8	40.00	2	10.00	10	50.00	20	100.00
Graça	7	43.80	1	6.30	8	50.00	16	100.00
Intendente	9	50.00	2	11.10	7	38.90	18	100.00
Lapa	12	52.20	1	4.30	10	43.50	23	100.00
Marquês de Pombal	10	32.30	2	6.50	19	61.30	31	100.00
Martim Moniz	38	36.90	6	5.80	59	57.30	103	100.00
Mercês	7	53.80	0	0.00	6	46.20	13	100.00
Principe Real	23	38.30	6	10.00	31	51.70	60	100.00
Restelo	4	20.00	2	10.00	14	70.00	20	100.00
S. Bento	2	8.00	0	0.00	23	92.00	25	100.00
S. José	4	20.00	1	5.00	15	75.00	20	100.00
Santa Catarina	62	40.00	12	7.70	81	52.30	155	100.00
Sé	12	26.70	5	11.10	28	62.20	45	100.00

Table 21: Bivariate analysis of properties' variables and sentiments

Note: 1131 observations.

The p values in both the "Elevator" (p=0.005) and "Neighborhood" (p=0.007) variables are lower than the significance level considered in this study (α=0.05). This means they have a significant relationship with the variable in question here, "Total Polarity". However, the strength of the associations is different. While "Elevator" has a Cramér's V of 0.058 (very weak), "Neighborhood" presents a 0.173 value (weak).

These values indicate that even though both variables have significant relationships with sentiment polarity, the neighborhood variable has a much stronger association with sentiment polarity than the elevator one.

4.4 Relationship between sentiments and the characteristics of the reservations

Next, the reservations' characteristics (including the average price per night) were tested for significant relationships. Out of all the tested variables, only price per night revealed a significant relationship, revealed by the p-value of 0.015, as we can see in

Table 22. Although the original table displayed the correlation between all these variables, a summarized version is presented here to make it easier to read the values we want to analyze.

	Total Polarity			
	Negative	**Neutral**	**Positive**	**Total**
Average Advance Booking (p= 0.342; Spearmn's correlaion coefficient= 0.028)				
Valid N	400	77	654	1131
Mean	2.77	2.77	2.80	2.79
Minimum	0.00	0.63	0.00	0.00
Median	2.50	2.55	2.57	2.55
Maximum	8.23	6.50	9.00	9.00
Standard Deviation	1.40	1.19	1.32	1.34
Average Number of Guests (p= 0.106; Spearmn's correlaion coefficient= 0.-0.048)				
Valid N	400	77	654	1131
Mean	3.67	3.69	3.57	3.62
Minimum	0.00	0.00	0.00	0.00
Median	3.57	3.60	3.50	3.50
Maximum	6.63	7.00	7.00	7.00
Standard Deviation	1.15	1.19	1.17	1.17
Average Stay Duration (p= 0.395; Spearmn's correlaion coefficient= 0.025)				
Valid N	400	77	654	1131
Mean	3.19	3.25	3.17	3.18
Minimum	0.00	0.00	0.00	0.00
Median	3.00	3.17	3.00	3.00
Maximum	17.00	6.00	10.00	17.00
Standard Deviation	1.28	0.84	0.85	1.02
Average Price per Night (p= 0.015; Spearmn's correlation coefficient= 0.072)				
Valid N	400	77	654	1131
Mean	98.6	99.77	104.33	101.99
Minimum	0.00	0.00	0.00	0.00
Median	91.85	92.4	98.14	95.41
Maximum	352.50	258.63	423.10	423.10
Standard Deviation	39.93	42.44	43.7	42.37

Table 22: Spearman's coefficient of Reservations and Sentiments' polarity

Note: 1131 observations.

Even though we could find this significant relationship, the correlation coefficient (Spearman) is considered very low, standing at 0.072. This correlation is positive, meaning when the price goes up, the sentiment goes up as well, but not very much. Even though this correlation is weak, it can be explained by the more the client pays, the better are the conditions of the rented property, ultimately leading to a better level of satisfaction from the client. When clients pay less for their rented properties, the level of satisfaction is lower. These conclusions were supported by FLH's manager.

4.5 Sentiment predictive model

After many models of the CART decision trees were created and evaluated, only the ones with the best results were chosen to showcase in this work. Table 23 resumes the results of the best models, alongside their evaluation metrics (sensibility, specificity, and accuracy).

	Model A	Model B	Model C	Model D	Model E
Parametrizations					
Tree depth	7	4	6	7	6
Minimum records in parent node	2	2	10	10	2
Minimum records in child node	1	1	5	5	1
Prior probabilities	Training sample	Training sample	Equal	Training sample	Training sample
Results					
Accuracy Resubstitution	73.8%	64.7%	64.9%	70.8%	70.0%
Sensibility Resubstitution	83.4%	76.9%	60.6%	85.2%	82.1%
Specificity Resubstitution	60.2%	48.0%	70.9%	51.2%	53.5%
Accuracy Crossvalidation (k=20)	52.1%	52.0%	53.2%	51.8%	51.9%
Accuracy Crossvalidation (k=10)	53.1%	53.9%	53.2%	53.3%	54.0%

Table 23: Predictive models results for Positive Sentiment

The chosen predictive model was Model E, as it was the model with the highest cross-validation accuracy (54.0% for 10 fold and 51.9% for 20 fold) out of every parametrization tested. Besides that, model E shows a high resubstitution sensibility (82.1%) and this was appraised by FLH' manager, who gives priority to correct positive sentiment prediction rather than the correct negative sentiment prediction, from a business point of view. Even though this model's quality is quite similar to model D, the FLH manager preferred the set of rules of model E instead of the set of rules of Model D.

Looking at the relative importance of each of the independent variables towards the prediction of the independent variable ("Positive Sentiment"), in Table 24, we can see the neighborhood is the most important variable, which goes along the results of the bivariate analysis of the properties' characteristics and sentiment. Right after neighborhood, the "Average Stay Duration" (92.40% of normalized importance) and "Average Price per Night" (83.80% of normalized importance) also present big importance in the sentiment's prediction. Finally, the month variable and the average number of guests per reservation also have moderate normalized importance in the variable prediction.

Independent Variable	Importance	Normalized Importance
Neighborhood	0.032	100.00%
Average Stay Duration	0.029	92.40%
Average Price per Night	0.027	83.80%
Month Name	0.025	79.40%
Average Number Guests	0.015	46.30%
Floor Number	0.015	45.70%
Main Nationality	0.012	37.90%
Number of Reservations	0.011	34.20%
Typology	0.009	28.90%
Max Occupancy	0.009	28.40%
Turistic Neighborhood	0.005	17.30%
Season	0.005	17.20%
Nº of Single Beds	0.005	16.10%
Elevator	0.005	15.10%
Single Bed	0.004	13.10%
Sofa Bed	0.004	12.50%
Nº of Double Beds	0.004	12.20%
Extra Bed	0.003	10.20%
Parking	0.003	9.20%
Double Bed	0.001	4.00%
High Season	0.001	1.90%
Holiday	0.001	1.90%
Historical Neighborhood	0	0.70%

Table 24: Independent variable importance

Concluding, while four of the reservations' characteristics play a big role in sentiment explanation, regarding the properties' characteristics, only the neighborhood and floor number seem to be important, even though the location characteristic (neighborhood) is much more important than the property's characteristic (the floor number).

Decision tree rules

Some rules (sets of conditions) can be formulated from the decision tree paths for each node. The proposed model is very complex (39 terminal nodes, which means 39 possible rules to define with the tree), so only a few rules are presented here to classify entries with positive sentiment and with negative sentiment, as examples. A high confidence level of 80% was established and a support number of cases to the rule of more than 10 was chosen.

Rules that predict positive sentiment:

- Node 13: IF the neighborhood is Restelo, Chiado, S. José Alfama, S. Bento, Ajuda, Av. Liberdade AND the average duration of stay is less or equal to 5.1 days AND there is no sofa bed AND the month is March, April, May, June, July, August, September, November or December THEN the sentiment is positive (with a confidence of 100% and a support of 43 cases, which represent 3.8% of the sample);
- Node 47: IF the neighborhood is Restelo, Chiado, S. José Alfama, S. Bento, Ajuda, Av. Liberdade AND the average duration of stay is less or equal to 5.1 days AND there is a sofa bed AND the main nationality of costumers is Spain, France, United Kingdom or Brazil AND the average price per night is less or equal to 69.73€ AND the month is February, March, July, November or December, THEN the sentiment is positive (with a confidence of 80% and a support of 20 cases, which represent 1.8% of the sample);
- Node 50: IF the neighborhood is Restelo, Chiado, S. José Alfama, S. Bento, Ajuda, Av. Liberdade AND the average duration of stay is less or equal to 5.1 days AND there is a sofa bed AND the main nationality of costumers is Spain, France, United Kingdom or Brazil AND the average price per night is more than 69.73€ AND the month is January, February, April, May, June, August, September, November or December, THEN the sentiment is positive (with a confidence of 85.2% and a support of 54 cases, which represent 4.8% of the sample).

Rules that predict negative sentiment:

- Node 62: IF the neighborhood is Lapa, Martim Moniz, Baixa, Marquês, Graça or Príncipe Real AND there is no elevator AND the number of monthly reservations is more than 6.5 AND the month is February, August, September or December AND the average number of guests is more than 2.9 THEN the sentiment is negative (with a confidence of 85.0% and a support of 20 cases, which represent 1.8% of the sample).

4.6 Monthly occupancy rate assessment

Table 25 showcases the descriptive statistics of the occupancy rates available in the dataset.

	Mean	Minimum	Median	Maximum	Standard Deviation
Relative Occup Rate	71.3	0	74.07	100	17.67

Table 25: Descriptive statistics of occupancy rates

Note: 1131 observations.

As we can see, from the dataset, on average, the monthly occupation rate is 71.3% and there are entries with a 0% rate, contrasting with some others with 100%.

4.7 Relationship between monthly occupancy rates and sentiments

The analysis of the relationship between sentiment's polarity and occupancy rates, showcased in Figure 11, reveals that not every sentiment polarity type (negative, neutral or positive) leads to similar occupancy rates, as the One-way ANOVA test shows ($F(2;1128)= 5.846$; $p= 0.003$). In fact, the Scheffé test allowed to conclude that the mean of occupation rates associated with neutral polarity (64.8%) is significantly lower than the mean of the other two types of sentiment polarity (both around 72%), which do not even differentiate one from another (p-value of 0.822).

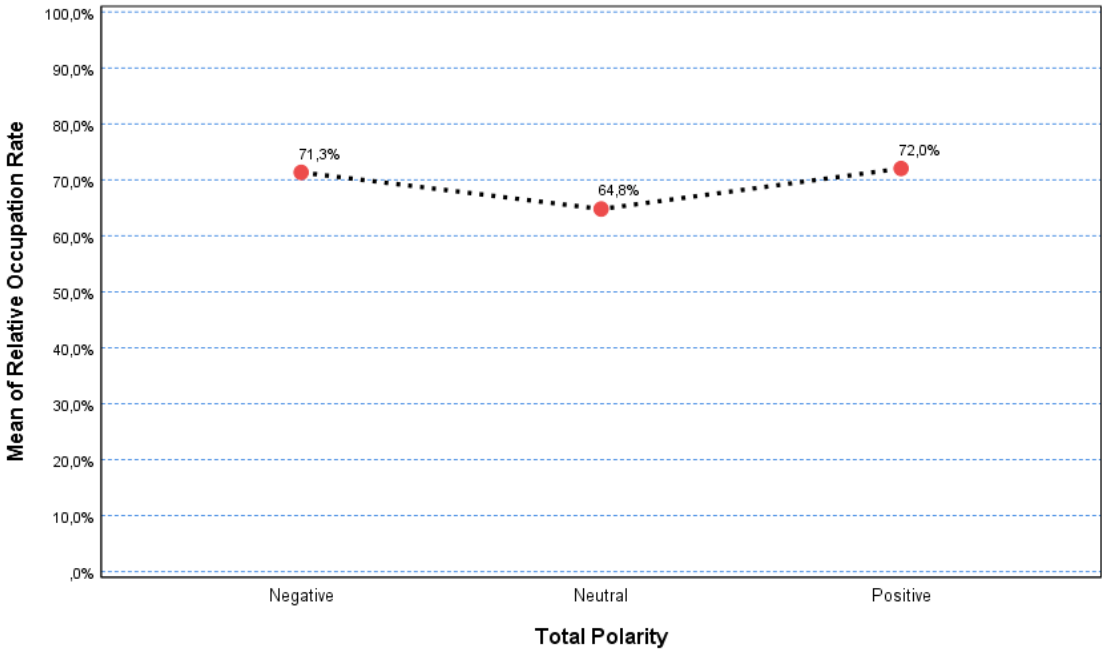

Figure 11: Occupation rate by sentiment polarity

Additionally, the relationships between the occupancy rates and the sentiment polarity from the previous month and the aggregated sentiment polarity from the previous six months were analyzed. However, while there was a similar conclusion regarding the previous month's polarity (comparing to the analysis described above, regarding the polarity from the same month), no significant differences among the three occupation means were found when analyzing the previous six months polarity (results not reported).

(This page was intentionally left blank)

5. Conclusion

5.1 Summary

In this investigation, we proposed to develop a solution for FLH's not optimized processes of sentiment assessment of their managed properties and tried to find a significant relationship between those sentiments and occupancy rates. The main research question was: how do the FLH's managed properties and reservations' characteristics influence the sentiment polarity inherent in their costumers' online reviews and how are these sentiments related to the occupation rates?

Adopting a CRISP-DM methodology, both objectives were accomplished through the usage of data mining and statistical techniques, such as chi-squared tests of independence, correlation coefficient tests, association tests and predictive models (in this case, CART decision trees). Both the objectives were translated into data mining problems and solutions were proposed.

Having this said, the research question was clearly answered. The neighborhood and the presence of an elevator in the property revealed to have a relationship with sentiments, although both considered being weak. Also, the price per night revealed to be the only characteristic of the reservations that influenced the sentiments' polarity (once again, considered weak). Although this bivariate analysis created useful knowledge, the predictive analysis must consider all available variables and their interactions, as their importance can increase when trying to predict the sentiment. For instance, the same month can generate either a positive or a negative sentiment polarity, depending on its interactions with other variables (see rules 62 and 47, for example, in section 4.5). Concluding, the interactions between the different properties' (namely, the neighborhood) and reservations' (namely the duration of stays, the price per night, the number of guests and the month of the occupation) characteristics allow to better predict the sentiment's polarity, as shown by the set of rules of the predictive model. Therefore, one should analyze these variables as a whole instead of individually.

5.2 Contributions

Answering the research question allows for contributes of the investigation to be pointed out, both for scientific knowledge and professionals in the same industry as, in the case of this study, FLH.

5.2.1 Scientific contributions

This investigation contributes to the expansion of knowledge on sentiment analysis on local accommodation. This type of case study is not very common. Usually, researchers tend to make investigations having the input of big datasets, as is the case of all studies using reviews from several different hotels in the same city. Also, most studies focus on analyzing sentiments and their influence on revenue or occupancy rates. This investigation, being a case study, took a more specific approach by studying data from one company only. Also, even though there are many studies on what factors influence sentiment polarity, many of those times they are limited when it comes to what factors they can consider to the research hypothesis, as they do not have access to more detailed information like this study had to FLH's managed properties details. Finally, these studies are not very common in Portugal yet, so this comes as a contribute to expand knowledge on this particular field in this country.

5.2.2. Industry and FLH

Professionals of the local accommodation industry gained a new study about the Portuguese reality on the industry and, in particular, about this industry in the capital of a city that relies so much on tourism, and, consequentially, on the local accommodation industry. In this regard, this study can help professionals understand what factors they might have to start worrying about when making choices about their businesses. Even the simple conclusion of location (a factor you cannot control about a house) having an important role in sentiment polarity can already make a business owner rethink about where he wants to start his/her business, or even consider a change of location for his/her business.

This study's contributions to FLH were a first-time analysis and problem-solving planning of the sentiments present in the online reviews of their properties' portfolio. The way the company managed this topic was not very planned and the processes were not very well defined. For the first time, a big scale analysis was performed on their

dataset, even though there were limitations to this study regarding the used sample. While some of the results were expected by FLH, such as "Location" being so well rated when it comes to polarity scores and "Amenities" having an unfavorable ratio of negative reviews, the number of negative scores on the "Room" category took them by surprise. FLH can now take the approach of this study and further develop it to get an even more specially tailored and detailed solution to implement on their business.

5.3 Limitations

Regarding the limitations of this study, the factor with the biggest impact was, without a doubt, the sample size (1131 entries) and the quality of the data retrieved from ReviewPro's API.

While originally (after retrieving the tables from FLHS's platform and treating the data) the dataset contained more than 6000 entries, it later had to be reduced to 1131 entries. The reason behind this decision was the very long (but needed) process of extraction of data from the API, due to its difficulty of being automatized (Booking.com was contacted in order to get information regarding their API but the platform was not allowing access to it during the time of this work's development). This led to several hours of extracting JSON files and eventually a limit number of properties to extract information from had to be decided due to time constraints.

Besides, many properties did not have mentions in online reviews (related to the categories studied in this work chosen due to FLH's context) in every month from the date they became available to the final date this work established (31st May 2019) and in many cases, there were few mentions. This may be related to FLH's market share and the number of clients they have per month, which may reflect in the number of comments the overall sample has.

5.4 Future research

Future research based on this study includes having a bigger and better sample than the one used (maybe using data from a bigger local accommodation rental platform), as well as including data from other cities. For patterns to be discovered in the data, creating good predictive models, large datasets must be used.

Other predictive models can be used to predict sentiments' polarity, such as artificial neural networks, other algorithms of decision trees (namely CHAID, C4.5, and C.5 trees) and multiple regressions. Besides that, other variables can be used in models, such as the six previous months aggregated polarity scores and the sentiment's polarities found in the houses' listings (for example, the owner's description of the house).

References

Airbnb (2019). About us. Retrieved on January 6, 2019, from https://press.airbnb.com/about-us/

Almeida, C. (2009). Aeroportos e Turismo Residencial. Do conhecimento às estratégias. URI: http://ria.ua.pt/handle/10773/1851

Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, 7(3), 112-118.

Belk, R. (2014). Sharing versus pseudo-sharing in Web 2.0. *The Anthropologist*, *18*(1), 7-23. https://doi.org/10.1080/09720073.2014.11891518

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. John Wiley & Sons.

Bieger, T., Beritelli, P., & Weinert, R. (2007). Understanding second home owners who do not rent—Insights on the proprietors of self-catered accommodation. *International Journal of Hospitality Management*, 26(2), 263-276. https://doi.org/10.1016/j.ijhm.2006.10.011

Bijmolt, T. H., Leeflang, P. S., Block, F., Eisenbeiss, M., Hardie, B. G., Lemmens, A., & Saffert, P. (2010). Analytics for customer engagement. *Journal of service research*, *13*(3), 341-356. https://doi.org/10.1177/1094670510375603

Booking.com (2019). About Booking.com. Retrieved in January 6, 2019, from https://www.booking.com/content/about

Borges, I. R., Pereira, G. M., Matos, C. A. D., & Borchardt, M. (2015). Análise da relação entre a satisfação dos consumidores e os preços ofertados no sítio booking. com. *Tourism & Management Studies*, 11(2), 64-70. http://dx.doi.org/10.18089/tms.2015.11208

Boswijk, A., Peelen, E. and Olthof, S. (2015), Economie van Experiences, Pearson, Amsterdam.

Bower, J. L., & Christensen, C. M. (1995). Disruptive technologies: catching the wave. DOI:10.1016/0024-6301(95)91075-1

Bridges, J., & Vásquez, C. (2018). If nearly all Airbnb reviews are positive, does that

make them meaningless?. *Current Issues in Tourism*, 21(18), 2057-2075. https://doi.org/10.1080/13683500.2016.1267113

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534. DOI: 10.1126/science.aap8062

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, *9*(2), 48-57. DOI: 10.1109/MCI.2014.2307227

Chanwisitkul, P., Shahgholian, A., & Mehandjiev, N. (2018). The Reason behind the Rating: Text Mining of Online Hotel Reviews. In 2018 IEEE 20th Conference on Business Informatics (CBI) (Vol. 1, pp. 149-157). IEEE. DOI: 10.1109/CBI.2018.00025

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide*. USA: SPSS Inc.

Chen, C. C., & Chang, Y. C. (2018). What drives purchase intention on Airbnb? Perspectives of consumer reviews, information quality, and media richness. *Telematics and Informatics*, 35(5), 1512-1523. https://doi.org/10.1016/j.tele.2018.03.019

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

Christensen, C. M. (1997). The innovator's dilemma: When new technologies cause great firms to fail. Boston, MA: Harvard Business School Press.

Christensen, C. M., & Raynor, M. E. (2003). The innovator's solution: Creating and sustaining suc- cessful growth. Boston, MA: Harvard Business School Press.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). ACM. DOI: 10.1145/1390156.1390177

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications*, 40(10), 3970-

3983. https://doi.org/10.1016/j.eswa.2013.01.012

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, *21*(4), 23-45. https://doi.org/10.1002/dir.20087

Devitt, A., & Ahmad, K. (2013). Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language resources and evaluation*, *47*(2), 475-511. https://doi.org/10.1007/s10579-013-9223-6

Díaz, M. R., & Rodríguez, T. F. E. (2018). Determining the reliability and validity of online reputation databases for lodging: Booking. com, TripAdvisor, and HolidayCheck. *Journal of Vacation Marketing*, *24*(3), 261-274. https://doi.org/10.1177/1356766717706103

El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *Machine Learning in Radiation Oncology* (pp. 3-11). Springer, Cham. https://doi.org/10.1007/978-3-319-18305-3_1

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*, *55*, 62-73. https://doi.org/10.1016/j.tourman.2016.01.013

Feels Like Home (2018). Onde estamos. Retrieved on December 31, 2018, from https://www.feelslikehome.pt/onde-estamos/

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. *Icwsm*, *7*(21), 219-222.

Gomes, R. D. S. D. E., Pinto, H. E. D. R. S., & Almeida, C. M. B. R. D. (2017). Second home tourism in the Algarve: The perception of public sector managers. *Revista Brasileira de Pesquisa em Turismo*, *11*(2), 197-217. http://dx.doi.org/10.7784/rbtur.v11i2.1246

Gössling, S., & Lane, B. (2015). Rural tourism and the development of Internet-based accommodation booking platforms: a study in the advantages, dangers and implications of innovation. *Journal of Sustainable Tourism*, *23*(8-9), 1386-1403. https://doi.org/10.1080/09669582.2014.909448

Guttentag, D. (2015). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. *Current issues in Tourism*, *18*(12), 1192-1217. https://doi.org/10.1080/13683500.2013.827159

Guttentag, D. A., & Smith, S. L. (2017). Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. *International Journal of Hospitality Management*, *64*, 1-10. https://doi.org/10.1016/j.ijhm.2017.02.003

Hand, D. J. (2013). Data MiningBased in part on the article "Data mining" by David Hand, which appeared in the Encyclopedia of Environmetrics . . *Encyclopedia of Environmetrics*. https://doi.org/10.1002/9780470057339.vad002.pub2

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L. E., & Hsu, M. C. (2011). Visual sentiment analysis on twitter data streams. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 277-278). IEEE. DOI: 10.1109/VAST.2011.6102472

Hemalatha, I., Varma, G. S., & Govardhan, A. (2012). Preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *1*(2), 58-61.

IBM (2019) IBM SPSS Modeler 15.0 welcome page. CRISP-DM Help. Retrieved on January 6, 2019, from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm

INE (2018) Estatísticas do Turismo 2017. Portugal.

Ireland, R., & Liu, A. (2018). Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP Journal of Manufacturing Science and Technology*, *23*, 128-144. https://doi.org/10.1016/j.cirpj.2018.06.003

JSON (2019). Retrieved on 13 September, 2019, from https://www.json.org/json-en.html

Khotimah, D. A. K., & Sarno, R. (2018). Sentiment Detection of Comment Titles in Booking. com Using Probabilistic Latent Semantic Analysis. In *2018 6th International Conference on Information and Communication Technology*

(ICoICT) (pp. 514-519). IEEE. DOI: 10.1109/ICoICT.2018.8528784

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Proceedings of the 14th international joint conference on Artificial intelligence. 2. pp. 1137-1143.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. In *Fifth International AAAI conference on weblogs and social media*.

Kumar, S., Morstatter, F., & Liu, H. (2014). *Twitter data analytics* (pp. 1041-4347). New York: Springer. https://doi.org/10.1007/978-1-4614-9372-3

Laureano, R. (2013) Testes de Hipóteses com o SPSS: O Meu Manual de Consulta Rápida, 2ª edição, Lisboa: Edições Sìlabo

Laureano, R. M., & do Carmo Botelho, M. (2017). IBM SPSS Statistics–O meu manual de consulta rápida. *Lisboa, 3ª Edição*.

Lavanya, T., JC, M. J. P., & Veningston, K. (2016). Online review analytics using word alignment model on Twitter data. In *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 1-6). IEEE. DOI: 10.1109/ICACCS.2016.7586388

Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications*, *37*(8), 5895-5904. https://doi.org/10.1016/j.eswa.2010.02.016

Li, L., Goh, T. T., & Jin, D. (2018). How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. *Neural Computing and Applications*, 1-29. https://doi.org/10.1007/s00521-018-3865-7

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, *19*(4), 456-474. https://doi.org/10.1287/isre.1070.0154

Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web* (pp. 342-351). ACM. DOI: 10.1145/1060745.1060797

Maimon, O., & Rokach, L. (Eds.). (2010). *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US. *Retrieved from http://link.springer.com/10.1007/978-0-387-09823-4*

Marques, A., & Cruz, R. S. (2014). Determinants of loyalty toward Booking. com brand. *The International Journal of Management Science and Information Technology (IJMSIT)*, (11), 96-123. URI: http://hdl.handle.net/10400.8/3729

Martinez, R. D., Carrington, A., Kuo, T., Tarhuni, L., & Abdel-Motaal, N. A. Z. (2017). The Impact of an AirBnb Host's Listing Description'Sentiment'and Length On Occupancy Rates. *arXiv preprint arXiv:1711.09196*.

Mittendorf, C. (2016). What Trust means in the Sharing Economy: A provider perspective on Airbnb.com. *Americas Conference on Information Systems*, 1–10.

Moswete, N., Thapa, B., Toteng, E. N., & Mbaiwa, J. E. (2008). Resident involvement and participation in urban tourism development: A comparative study in Maun and Gaborone, Botswana. In *Urban Forum* (Vol. 19, No. 4, pp. 381-394). Springer Netherlands. https://doi.org/10.1007/s12132-008-9041-x

Müller, D. K. (2014). Progress in second-home tourism research. *The Wiley Blackwell companion to tourism*, 389-400. https://doi.org/10.1002/9781118474648.ch31

Musto, C., Semeraro, G., & Polignano, M. (2014). A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts. In *DART@ AI* IA* (pp. 59-68).

Oskam, J., & Boswijk, A. (2016). Airbnb: the future of networked hospitality businesses. *Journal of Tourism Futures*, *2*(1), 22-42. https://doi.org/10.1108/JTF-11-2015-0048

Oskam, J., van der Rest, J. P., & Telkamp, B. (2018). What's mine is yours—but at

what price? Dynamic pricing behavior as an indicator of Airbnb host professionalization. *Journal of Revenue and Pricing Management*, *17*(5), 311-328. https://doi.org/10.1057/s41272-018-00157-3

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135. http://dx.doi.org/10.1561/1500000011

Papathanassis, A., & Knolle, F. (2011). Exploring the adoption and processing of online holiday reviews: A grounded theory approach. *Tourism Management*, *32*(2), 215-224. https://doi.org/10.1016/j.tourman.2009.12.005

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2015). Reprint of: Computational approaches for mining user's opinions on the Web 2.0. *Information Processing & Management*, *51*(4), 510-519. https://doi.org/10.1016/j.ipm.2014.07.011

Phillips, P., Barnes, S., Zigan, K., & Schegg, R. (2017). Understanding the impact of online reviews on hotel performance: an empirical analysis. *Journal of Travel Research*, *56*(2), 235-249. https://doi.org/10.1177/0047287516636481

Ramos, D., & Almeida, L. (2017). Tourism Porto and North of Portugal–Case Study Concerning Private Accommodation.

ReviewPro. (2019). Sobre a ReviewPro. ReviewPro. Retrieved in June 20, 2019, from https://www.reviewpro.com/pt-pt/companhia/sobre-a-reviewpro/

Roca, M., Oliveira, J., Roca, Z., & Costa, L. (2012). Second Home Tourism in the Oeste Region, Portugal: Features and Impacts. *European Journal of Tourism, Hospitality and Recreation*, *3*(2), 35–55.

Saló, A., & Garriga, A. (2011). The second-home rental market: A hedonic analysis of the effect of different characteristics and a high-market-share intermediary on price. *Tourism Economics*, *17*(5), 1017-1033. https://doi.org/10.5367/te.2011.0074

Schaffner, D., Georgi, D., & Federspiel, E. (2017). Comparing customer experiences and usage intentions regarding peer-to-peer sharing platforms with conventional online booking websites: The role of social, hedonic, and functional values.

In *Marketing at the Confluence between Entertainment and Analytics* (pp. 1049-1056). Springer, Cham. https://doi.org/10.1007/978-3-319-47331-4_208

Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2017). Document Level Sentiment Analysis from News Articles. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)* (pp. 1-4). IEEE. DOI: 10.1109/ICCUBEA.2017.8463638

Singh, J., Singh, G., & Singh, R. (2016). A review of sentiment analysis techniques for opinionated web text. *CSI transactions on ICT*, *4*(2-4), 241-247. https://doi.org/10.1007/s40012-016-0107-y

Singla, Z., Randhawa, S., & Jain, S. (2017). Sentiment analysis of customer product reviews using machine learning. In *2017 International Conference on Intelligent Computing and Control (I2C2)* (pp. 1-5). IEEE. DOI: 10.1109/I2C2.2017.8321910

Skak, M., & Bloze, G. (2017). Owning and letting of second homes: what are the drivers? insights from Denmark. *Journal of Housing and the Built Environment*, *32*(4), 693-712. https://doi.org/10.1007/s10901-016-9531-4

Tanz, J. (2014): How Airbnb and Lyft Finally Got Americans to Trust Each Other. Available at: https://www.wired.com/2014/04/trust-in-the-share-economy/ (accessed in 13 /07/2019

Turban, E. (2011). *Business intelligence: a managerial approach*. Boston: Prentice Hall.

UNWTO (2016). World Tourism Organization. Tourism Highlights 2016. UNWTO edition.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. Psychological Bulletin, 98(2), 219–235.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). Citeseer.

Yordanova, S., & Kabakchieva, D. (2017). Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning. *International Journal of Computer*

Applications, *158*(7).

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2018). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 1-47. https://doi.org/10.1007/s10115-018-1236-4

Zervas, G., Proserpio, D., & Byers, J. (2015). A first look at online reputation on Airbnb, where every stay is above average. *Where Every Stay is Above Average (January 28, 2015)*. http://dx.doi.org/10.2139/ssrn.2554500

(This page was intentionally left blank)

Appendix

A- Final table

Name	Description	Source	Nature
ID	Row identifier. Used to prevent discrepancies between the Excel data and SPSS	Automatically generated, one ID per row	Nominal
Property	Number of the property	FLH database	Nominal
N Month	Month's number	Applied a formula to obtain it from "Date"	Scale
Name Month	Name of the month	Applied a formula to obtain it from "Date"	Nominal
Season	Season of the year	Applied a formula to obtain it from "Date"	Nominal
High Season	Whether it is High Season or not	Obtained from "Season". June, July, August and December return "Yes".	Nominal
Holiday	Whether it is Holiday or not	Obtained from "Date". February, April and December return "Yes".	Nominal
Total Mentions	Total of mentions in the reviews	Obtained from the sum of positive and negative mentions	Scale
Total Score Classification	Classification of the total score	Obtained from the subtraction of positive mentions and negative mentions. Negative values until -11= Very Negative; -11 to 1= Negative; 0= Neutral; 1 to 10= Positive; from 11 = Very Positive	Ordinal
Total Score Classification 6 Months	Classification of the aggregated score from the previous six months	Obtained from the subtraction of positive mentions and negative mentions in the six months prior to the referred month. Negative values until -11= Very Negative; -11 to 1= Negative; 0= Neutral; 1 to 10= Positive; from 11 = Very Positive	Ordinal
Average Advance Booking	The average number of months between reservation date and check-in date	Obtained calculating the average distance, in months, from check-in dates and reservation dates	Scale
Relative Occupancy Rate	Relative occupancy rate	Explained in the occupancy table subtopic	Scale
Previous Relative Occupancy Rate	The relative occupancy rate of the previous month	Explained in the occupancy table subtopic, but from the previous month	Scale
Main Nationality	Main nationality of the costumers	Obtained from the Countries table. Filters were used to obtain the highest percentage of the users, to relate it with their nationality	Nominal
Main Distribution	Main distribution channel	Obtained from the reservations table. Pivot tables were used.	Nominal
Average Num Guests	The average number of guests	Obtained from the aggregated reservations table	Scale
Average Duration	The average duration of stay, in days	Obtained from the aggregated reservations table	Scale
Average Price per Night	Average price per night of the reservations made for that month	Obtained from the aggregated reservations table	Scale

Table 26: Final table for the dataset (i)

Name	Description	Source	Nature
Neighborhood	Neighborhood	Obtained from the properties table	Nominal
Historical Neighborhood	Whether the neighborhood is considered historical or not	Obtained from "Neighborhood". Google Maps was used, after retrieving information from Lisbon's city hall website.	Nominal
Turistic Neighborhood	Whether the neighborhood is considered touristic or not	Obtained from "Neighborhood". Google Maps was used, after retrieving information from Lisbon's city hall website.	Nominal
Max Occupancy	Maximum number of people allowed per reservation	Obtained from the properties table	Scale
Typology	Typology of the property	Obtained from the properties table	Nominal
Floor number	Floor number of the property	Obtained from the properties table	Nominal
N Double Bed	Number of double beds in the property	Obtained from the properties table	Scale
Double Bed	Whether the house has double beds	Obtained from the properties table. Values other than zero return "Yes"	Nominal
N Single Bed	Number of single beds in the property	Obtained from the properties table	Scale
Single Bed	Whether the house has single beds or not	Obtained from the properties table. Values other than zero return "Yes"	Nominal
Sofa Bed	Whether the house has a sofa bed or not	Obtained from the properties table	Nominal
Extra Bed	Whether the house has an extra bed or not	Obtained from the properties table	Nominal
Elevator	Whether the house has an elevator or not	Obtained from the properties table	Nominal
Parking	Whether there is a dedicated parking spot or not	Obtained from the properties table	Nominal

Table 27: Final table for the dataset (ii)

B- Descriptive analysis

		Count	Mean	Minimum	Median	Maximum	Mode	Standard Deviation
Month	April	123						
	August	89						
	December	70						
	February	79						
	January	86						
	July	76						
	June	81						
	March	118						
	May	137						
	November	83						
	October	93						
	September	96						
Season	Autumn	246						
	Spring	341						
	Summer	261						
	Winter	283						
High Season	No	815						
	Yes	316						
Holiday	No	859						
	Yes	272						
Average Advance Booking			2,79	,00	2,55	9,00	2,00	1,34
Relative Occup Rate			71,30	,00	74,07	100,00	80,65	17,67
Previous Rel Month Occup			70,32	,00	74,19	100,00	80,65	19,05
Main Nationality	Brazil	86						
	France	550						
	Germany	45						
	Spain	409						
	United Kingdom	41						
Main Distribution Channel	Airbnb	65						
	Booking	993						
	Feelslikehome	3						
	None	66						
	Other	4						
Average Number of Guests			3,64	1,00	3,55	7,00	2,00	1,13
Average Stay Duration			3,20	1,00	3,00	17,00	3,00	,99
Price per Night			102,63	30,65	95,64	423,10	50,40[a]	41,73
Historical Neighborhood	No	50						
	Yes	1081						
Turistic Neighborhood	No	602						
	Yes	529						
Couple	No	1001						
	Yes	130						
Max Occupancy			4	2	4	7	4	1
Typology	T0	108						
	T1	225						
	T2	582						
	T3	170						
	T4	46						
Floor Number			2	-1	2	12	2	2
Nº of Double Beds			1	0	1	3	1	1
Double Bed	No	51						
	Yes	1080						
Nº of Single Beds			1	0	2	5	0	1
Single Bed	No	500						
	Yes	631						
Sofa Bed	No	680						
	Yes	451						
Extra Bed	No	814						
	Yes	317						
Elevator	No	728						
	Yes	403						
Parking	No	990						
	Yes	141						
Total Polarity	Negative	400						
	Neutral	77						
	Positive	654						
Previous Six Months Polarity	Negative	370						
	Neutral	25						
	Positive	736						

a. Multiple modes exist. The smallest value is shown

Table 28: Descriptive statistics of all the variables (i)

		Count	%
Neighborhood	Ajuda	10	0.90%
	Alfama	78	6.90%
	Av. da Liberdade	24	2.10%
	Av. Novas	65	5.70%
	Bairro Alto	209	18.50%
	Baixa	128	11.30%
	Campo de Ourique	40	3.50%
	Chiado	14	1.20%
	Estrela	14	1.20%
	Expo	20	1.80%
	Graça	16	1.40%
	Intendente	18	1.60%
	Lapa	23	2.00%
	Marquês de Pombal	31	2.70%
	Martim Moniz	103	9.10%
	Mercês	13	1.10%
	Principe Real	60	5.30%
	Restelo	20	1.80%
	S. Bento	25	2.20%
	S. José	20	1.80%
	Santa Catarina	155	13.70%
	Sé	45	4.00%

Table 29: Descriptive statistics of all the variables (ii)