UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER THESIS

OPPORTUNITIES AND BARRIERS OF USING TEXT MINING IN THE HOSPITALITY INDUSTRY – THE CASE OF SOTELIA HOTEL

Ljubljana, May 2022 SIMIĆ MAJA

AUTHORSHIP STATEMENT

The undersigned Maja Simić a student at the University of Ljubljana, School of Economics and Business, (hereafter: SEB LU), author of this written final work of studies with the title Opportunities and barriers of using text mining in the hospitality industry – the case of Sotelia hotel, prepared under supervision of prof. dr. Jurij Jaklič

D E C L A R E

1. this written final work of studies to be based on the results of my own research;

2. the printed form of this written final work of studies to be identical to its electronic form;

3. the text of this written final work of studies to be language-edited and technically in adherence with the SEB LU's Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the SEB LU's Technical Guidelines for Written Works;

4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;

5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the SEB LU in accordance with the relevant SEB LU Rules;

6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;

7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;

8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the SEB LU Study Information System;

9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;

10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, _____ Author's signature: _____

TABLE OF CONTENTS

INTRODUCTION **1**

1 USE OF INTERNET AND INFORMATION TECHNOLOGIES IN HOSPITALITY INDUSTRY **3**

1.1 Current trends in the Information Technology **4**

1.2 Current trends in the Hospitality industry **7**

1.3 Importance of customer satisfaction and the eWOM **9**

2 TEXT MINING IN THE HOSPITALITY INDUSTY **11**

2.1 Text mining tools and techniques **12**

2.2 Sentiment Analysis for analysing online hotel reviews **17**

3 METHODOLOGY **21**

3.1 Research methodology **21**

3.2 Description of Hotel Sotelia – Podčetrtek, Slovenia **24**

4 ANALYSIS AND INTERPRETATION OF RESULTS **25**

4.1 Explanation of the processes done in RapidMiner **25**

4.2 Analyses performed in Microsoft Excel **31**

5 DISCUSSION **44**

5.1 Interview and interpretation of results **45**

5.2 Opportunities of using text mining in hospitality industry **50**

5.3 Limitations of using text mining **51**

5.4 Detecting hidden patterns with text mining **54**

CONCLUSION **55**

REFERENCE LIST **58**

APPENDICES **65**

LIST OF FIGURES

Figure 1: Processing of Text Mining 15
Figure 2: Sentiment analysis of hotel review 18
Figure 3: Tripadvisor website, hotel Sotelia 19

Figure 4: Fine-granted Sentiment Analysis 19
Figure 5: Octoparse final extraction 21
Figure 6: Research flow 23
Figure 7: Hotel Sotelia 24
Figure 8: First process in RapidMiner (data preparation and tokenization) 27
Figure 9: Process of getting aspects in RapidMiner 28
Figure 10: Word Frequency analysis done with RapidMiner 29
Figure 11: Word Cloud 30
Figure 12: Total number of textual reviews per site and their percentages 32
Figure 13: number of aspects by polarity, Tripadvisor.com data 35
Figure 14: Polarities of most occurred aspects, compared to the whole dataset (in %) 37
Figure 15: Percentage of comment by rate from one to five, depending on the site 38
Figure 16: Two-dimensional contingency table 42

LIST OF TABLES

Table 1: Overview of top ten hospitality trends according to EHL 8
Table 2: Business intelligence vs data science 13
Table 3: Text Mining vs Data Mining 13
Table 4: Word Frequency Analysis 29
Table 5: Explanation for rate standardization 31
Table 6: Number of Google Reviews with only a few words and rate provided 32
Table 7: Examples of Google Reviews with a few words and rate provided 33
Table 8: Number of Booking.com reviews with a few words and rate provided 33
Table 9: Examples of Booking.com reviews with a few words and rate provided 33
Table 10: Tripadvisor.com numbers and percentages of aspects and their polarities 34
Table 11: Sentiments towards food/drinks indicated in numbers and percentages 35
Table 12: Sentiments towards facilities indicated in numbers and percentages 36
Table 13: Sentiments towards staff indicated in numbers and percentages 36
Table 14: Sentiments towards room amenities indicated in numbers and percentages 36
Table 15: Overview of given rates per each site expressed as a percentage 38
Table 16: Negative reviews and their aspects recognized by RapidMiner 39
Table 17: Calculating aspects association 43
Table 18: Cramer's coefficient 44

LIST OF APPENDICES

Appendix 1: Povzetek (Summary in Slovene language) 1
Appendix 2: Interview questions 4

LIST OF EQUATIONS

Equation 1: Calculation for Cramer's coefficient (V) 42
Equation 2: Calculation for Cramer's coefficient (V), predicted values 42

LIST OF ABBREVIATIONS

ABSA - Aspect-based Sentiment Analysis
AI - Artificial Intelligence
AJAX - Asynchronous JavaScript
API - Application Programming Interface
BI - Business intelligence
CAGR - Compound Annual Growth Rate
CLV - Customer lifetime value
EHL - The École hôtelière de Lausanne
eWOM - Electronic Word of Mouth
HTTP - Hypertext Transfer Protocol
ICT - Information and communication technologies
IR - Information Retrieval
IT - Information Technology
NLP - Natural Language Processing
RQ - Research Question
SA - Sentiment analysis
UGC - User-generated content
WWW - World Wide Web
XML - Extensible Markup Language

INTRODUCTION

Over the last two decades, the hospitality industry has been growing rapidly. Tatulli (2019) claimed in his research work that there are more than 700,000 hotels worldwide, which were providing over $3.41 trillion to the global economy every year. Today, this number did not change a lot, because of the Covid-19 pandemic that stopped the growth of the hospitality industry for 2 years, resulting in global lockdowns and dramatically changes in consumers' behaviour. However, estimation is that the Covid-19 pandemic will eventually pass and that, the hospitality industry will recover and will look forward to a prosperous future (Deloitte, 2020). The leader in terms of growth and development among industries is the Information Technology (IT) industry, since it became crucial for our jobs, personal life, and all economy branches, which are becoming more digital, day-by-day. A boom in information and communication technologies (ICT) has led to fast development of Web 2.0, which is characterized mostly by electronic word of mouth (eWOM) or user-generated content (UGC) (Zelia, Rui, Gorete, & Martins, 2020). From the time when the information technology was invented, the early 19th century, the ability to generate massive amounts of data had never been so powerful (Xindong, Xingquan, Gong-Qing, & Wei, 2013). Various platforms and constant availability of the Internet allow exchanging enormous amounts of data in real-time, which can, if analysed properly, bring precious knowledge about the customers and about actions, which need to be undertaken in order to excel business (Taşkın, 2016). Social media and travel-related websites, where electronic word of mouth stands for number one priority, play the vital role in today's world, especially in the hospitality sector. Slightly more than ninety percent of travellers say that online hotel reviews are especially important for the decision making while making plans of their vacation (Zhao, Wang, Guo, & Law, 2015). The participation on the Internet is more than welcome, and the websites had never been more open for customer's opinion, reviews, recommendation, and ratings. In the online hotel reviews, both, numerical and textual evaluations can be found, and both are important as they represent trustworthy and objective feedback from the other guests (Gavilan, Avello, & Martinez-Navarro, 2018). Depending on the results of these numerical and textual evaluations, the future customers will decide upon their stay.

The global IT industry is projected to increase from $8,384.32 billion in 2021 to $9,325.69 billion in 2022 at a compound annual growth rate (CAGR) of 11.2%, whereas the expectation for 2026 is that the IT market will reach $13,818.98 billion at CAGR of 10.3%. (The Business Research Company, 2022). Reason for this growth is mainly due to the rearrangement of the operations in the companies and their recovering from the Covid-19 impact. Both, hospitality and IT industries, are very important for the global economy. Thus, to ensure a good position on the market and to get insights from the guests, hotel managers are advised to find the best way of managing the hotel by combining information technology with other managerial sciences, and business domains (Calheiros, 2015).

The thesis will tend to inspect if the reviews on the online travel sites can help management better understand their quests, if these reviews can have impact on hotel's reputation and customer's decision making while booking accommodation online. Primarily, if the usage of mining techniques such as aspect-based sentiment analysis could help, not only in gathering and in analysing data, but also in inspecting if these techniques can improve the business and give the real picture of the customer's opinion shared on the travel sites. In other words, the aim of the thesis is to reveal if the aspect-based sentiment analysis and similar text mining approaches can lead to general improvements and competitive advantage if conducted and analysed properly. Nevertheless, another important thing covered in the thesis is if text mining approach have some gaps and limitations, which are essential and in that way are blocking the truth and validity of text.

Hence, three research questions (RQ) are defined:

RQ1: On which analysis can hotel managers rely on while using text mining, techniques for analysing reviews in the hospitality industry? In other words, what are the opportunities of using text mining in hospitality industry?

RQ2: Are there any pitfalls in analysing reviews while using text mining tools and techniques? If so, what actions should be undertaken to resolve them?

RQ3: Is text mining technique the right one for detecting hidden patterns in customer's sentiments and can that action lead to business improvement and competitive advantage?

In terms of methodological approach, the thesis is conducted considering qualitative and quantitative method. The rational for choosing mixed research methodology is that text posted online is, most of the time trustworthy, and accurate measurement of customer satisfaction. Based on that, along with the theoretical part, it was important to conduct several analyses, and to describe and present all results in the right way. Regarding quantitative method, data for this thesis are gathered from three websites (Tripadvisor.com, Booking.com and Google.com) dating from January 2009 to January 2020, and then analysed with different text mining techniques. The text mining approach is put on to inspect key insides from the reviews. Moreover, these insights are analysed in terms of their aspects with help of software/platform called RapidMiner. Once known, the analysed results are compared with the original review posted on travel site. This helped in revealing whether the analysis was reliable, and if it this type of analysis is worth conducting in hospitality industry. When it comes to qualitative method, it was used to provide literature overview of the topic, to describe the current trends in hospitality and IT industry, to interpret the analysis and finally to propose the potential improvements regarding business aspects, which are, according to the analysed data, the most important ones for the customers. After all, it was meaningful to conduct the interview in order to check the validity of results obtained from the analyses. Thus, semi-structured interview with the sales representative of hotel Sotelia, who is familiar with hotel's mission, vision, aims, and strategies, is conducted to better

understand the current business of hotel. Furthermore, interview helped to inspect if the analysis in the theses was accurate, helpful and if these kinds of analysis can lead managers to new findings and consequently, to new decisions, which will lead to better and more pleasant environment for future guests.

In the first chapter, main trends in hospitality and information technology industry are defined, but also the importance of the electronic word of mouth is provided. In the second chapter, general information regarding text mining is written, along with the description of the aspect-based sentiment analysis and potential limitations of using text mining. In the third part, the methodology was described, while in the fourth part the analyses and interpretation of results are presented. In this part, it is explained how data are collected, how the processes in RapidMiner were run and interpretation of the results is provided. In the fourth part, the summary of the interview and remarks regarding the previous chapters is presented, along with the answers on the defined research questions. Last, but not the least, main points, potential improvements, and limitations of the study, are provided in the conclusion. Nevertheless, the focus and the overall purpose remained on examining opportunities and barriers of using text mining in the hospitality industry.

1 USE OF INTERNET AND INFORMATION TECHNOLOGIES IN HOSPITALITY INDUSTRY

Internet age started at 1960 and in later years it was believed that use of internet in conducting surveys and collection of electronic data may revolutionize many fields of study having larger samples and bigger reach, by easier data collection, and therefore more representative data. However, the others were sceptical of internet's usability, but also with its practical value (Benfield & William, 2006). Introduction of Hypertext Transfer Protocol (HTTP) and World Wide Web (WWW) enabled less paperwork, cost reductions, including labour costs, mailing costs etc. Up to now, internet has transformed education, business, government, healthcare, the way how people communicate with each other. It became one of the most important factors from most of the society changes. Today, it can be said that traditional ways of communication are fading, and almost everything is based on the internet, online communication, and technology. Social media networks, web sites and forums represent a never-ending source of opinions and discussions regarding wide variety of topic. The world switched to digital environment and thus, the digital documents, comments, posts, photos, and other ways of communication became extremely important and for many users, number one source of truth. To gather these data, different techniques can be applied, depending on data classification. Data can be classified as:

– Structured data, also known as tabular data, are usually presented with columns and rows in the rational database. Examples of structured data are dates, names, credit card numbers, addresses and other similar data.

- Semi-structured are documents and information that do not contain structured data, nevertheless they are not without any structure. Examples include email, XML and JSON documents, CSV files and other similar documents.
- Unstructured data are presented as the information that is not organized in the defined manner or for which there is no pre-determined data model. In this category may represent data such as audio, video, dates, different facts, and binary data files which do not have specific structure (Vidhya & Aghila, 2010).

Sometimes, when data are unstructured or when the structure is not easily defined, classic database management systems are not able to store and analyse all data (Laudon & Laudon, 2004). Helping business leaders and hotel owners to identify and reveal hidden patterns and trends in the hospitality industry is inevitable and that is why, so called "Big data approach", is an especially important approach to understand. Big data relates to a combination of complex, large-volume data sets with multiple sources, which can be structured, unstructured and semi-structure (Rouse, 2019; Xindong, Xingquan, Gong-Qing, & Wei, 2013). One way of dealing with the big data is by using data-mining tools for collecting, organizing, cleansing, and visualizing data. These tools are used to better understand the patterns of customer behaviour, reveal hidden information, and predict new trends (Lee & Siau, 2001; Hoontrakul & Sahadev, 2008). That is why the text mining techniques are used in analysing the eWOM including reviews from travel sites. Text mining, also known as text analytics, tries to solve problem of too much information by combining different techniques from natural language processing (NLP), machine learning, knowledge management, data mining, information retrieval (IR) and in that way extract the useful information from textual document (Feldman & Sanger, 2007). The whole process requires and involves information retrieval, storage of the information retrieved, big data, machine leaning, special techniques for analysing given data, statistics, visualization of the results and more (Chang et al., 2018).

Even though text mining techniques certainly represent a much better way of analysing the data and conducting business connected to the hotel industry, there are still some gaps and limitations such as language difficulties, dealing with fake reviews, writing a text which is contradictory to the ratings, etc. These limitations are investigated throughout the practical part of the thesis in chapter 4, whereas in chapter 5.2 the answer on the research question related to limitation can be found.

1.1 Current trends in the Information Technology

IT industry is among a few, which despite the Covid-19 pandemic, has a growing trend. The positive thing for IT industry, during these tough times, is that economy, education, employment, work in general, and personal lives became more digital, more connected in terms of engagement on the internet, and more automated than ever. In the previous years, information sharing was controlled according to firms' needs, nowadays the customers determine which information they want to see, post, and follow (Limberger, Anjos, Meira,

& Anjos, 2014). It is easy to conclude that both, Hospitality, and IT industry are necessary for the global economy growth. As with implementation of IT - storing, retrieving, transmitting, and data manipulation are much faster, companies that implement IT modern solutions are supposed to improve their business, make changes faster and have a huge profit out of it (Stifanich Pilepić, & Šimunić, 2019). Because of the stated benefits, other industries are also more than encouraged to implement any IT solutions, which will bring them better position on the market. In the past twenty years IT sector went through incredible changes and with its innovative technologies, automation, and mechanization it also transformed the global hospitality industry. The IT sector plays the most important role in the hospitality world, as it allows data processing and offers innovative technological solutions (Car, Stifanich Pilepić, & Šimunić, 2019). Because of those possibilities, the hospitality industry revolutionized the way of dealing with customer, it became more agile and more dynamic. The progress and involvement of Artificial Intelligence (AI) had a huge impact on hospitality industry. Having in mind the hospitality sector, this situation helped in building build better connection with customers, getting closer to them and reducing extra costs. Increasing the customer's outreach and introducing the public use of the internet without boundaries led to this situation, which transformed the game and provided many benefits, but also introduced some unfavourable scenarios to travel agencies. Nevertheless, this does not mean that travel agencies will disappear because there are still many advantages of using them. Digitalization had been recognized as one of the key trends transforming society and business in the near and long-term future (Kääriäinen, et al., 2016). Hotel owners are usually familiar with the fact that digital transformation can help them achieve many goals, which are having the highest impact on their business (HotelTechReport, 2021). Hence, digital transformation can help to:

- Increase website traffic and digital revenue
- Reduce operational costs
- Improve service quality
- Improve customer outcomes

Due to the fact that the IT industry is on its highest level and that many of the services are already digitalized, both business and customers can enjoy online reservations, quicker property management at front desk, lower labour costs and faster service, better and improved accuracy, and modern websites where customers can easily exchange opinions, make payments, and reserve the right room, restaurant, spa centre or any other activity which will fulfil their needs. All of these is available with just a few clicks on laptop, cell phone, or tablet.

The generation that is most familiar with these kinds of actions is the millennial generation and IT has a crucial role in engaging these types of travellers to make reservations. With the appropriate online promotion, including banners, social media, mail marketing and site

promotions, the reach to these groups could be much more extensive than using traditional ways of promotion.

Below are stated some of the most popular trends which would not exist without IT industry:

- One of the main trends relates to reservations. It is now possible to book accommodation for anyone via device connected to the internet by going on the travel sites. The possibility of comparing the prices, reading the comments, and choosing the room for any occasion had never been easier. Nowadays, many hotels also have a 24x7 Artificial Intelligence powered chatbot used to increase direct communication with customers and bookings on the website. Sotelia does not have this possibility, and it would be a good proposal for the hotel to introduce Chatbot (Roberts, 2021).

- Another trend is connected with mobile communications. Hotels can send updates, notices, promotions, and offer deals via online communication channels. The customers' reach can be enhanced through text, GPS tagging or messaging emails (Roberts, 2021).

- In-room technology presents a one of the vital functionalities in modern times. As one of the must-have things while on vacation are electronic devices, the internet connection in the room has become one of the top priorities for travellers. Some hotels have their web application, allowing customers to access it via Wi-Fi and enjoy room service options, interactive service, restaurant reservations, and nearby attractions (Roberts, 2021).

The IT industry has had the highest impact on organising, coordinating, and making a self-service booking in the last few years. Because of the rapid hardware and software improvements, information is stored more quickly and accurately than before. These improvements allow companies in all industries to be more efficient and prevent capacity loss. Travel agencies can use enterprise-level software which offers special programs used to track and manage the business, process data in the best way and organize them accordingly. In addition, hotels and travel agencies use this solution to communicate with business partners, sponsors, and outsourcing agencies.

Video conferences, high-speed internet and constant and instant communication worldwide have become less expensive than traditional methods. It is much easier to coordinate the processes as data are sent in real-time, allowing changes to be seen instantly after they are made, reducing time waits, increasing productivity, and providing better organization. With the help of notification, it is easier to track guests, see if any new booking is made and, in that way, not forget about the customers. Moreover, customers do not have to go to a travel agency anymore since many online platforms such as Booking.com, Tripadvisor.com, and other travel sites allow instant booking, car hire, restaurants reservations, and many more in one place.

1.2 Current trends in the Hospitality industry

Before describing the general trends in the hospitality industry, the current Covid-19 situation cannot be overseen, as it, unfortunately, stopped the growth of the mentioned industry in early 2020. From the situation of so-called “over-tourism” (Dodds & Butler, 2019), the global tourism industry moved to “non-tourism”. Besides, in 2020 the whole Global economy was affected due to the Covid-19 pandemic.

Almost overnight, the whole economy was shut down, impacting many businesses, and leaving many people without a job. One of the most impacted industries was the hospitality industry, where this pandemic caused a severe impact on personnel, operations, supply chain, revenue, and overall assessment (Dogan & Chi, 2020). Governments worldwide have raised the restrictions, introduced lockdowns, and claimed social distancing and work from home. All these measures were conducted to flatten the Covid-19 curve (Dogan & Chi, 2020). Consequently, the hospitality businesses worldwide experienced a significant decrease in demand. Depending on the country, some objects such as cafes, restaurants and hotels were totally shut down.

Closing the borders and the other orders issued by the governments led to a tremendous decline in hotel revenue and occupancy of the hotel rooms in the period from 2020 to early 2022. Unlike the other sectors, tourism revenue during the Covid-19 epidemic was irreversibly lost because unreserved rooms could not be booked in the later years (Gössling, Scott, & Hall, 2021). The hospitality sector is forced to ensure a healthy and safe environment for employees and customers and enhance travellers' willingness to invest in their business. While in 2019, managers and scholars were thinking about how to manage their business better to gain more, today they should answer the vital questions, such as: What are the customers' thoughts about visiting a restaurant or a hotel in the Covid-19 pandemic? Are they prepared to come? If not, what will make customers return to the tradition of visiting hotels and restaurants? (Gössling, Scott, & Hall, 2021).

Moreover, research also suggests that 40% of customers are ready to pay more for improved health and safety measures. From this information, it can be concluded that guests' primary focus is switching from luxury, which was previously in the first place, to cleanliness and safety. Many of the customers believe that various technologies used for service will be essential in the pandemic period and after that. Customers believe that service robots, digital menus with QR codes, contactless payments via card or mobile phone, keyless entry, and other adjusted conditions must be impeded by hotels to minimise exposure to other people.

Artificial Intelligence (AI) and the whole IT industry will continue to play a critical role. Hence, it is essential for people in the hospitality sector to examine how AI devices' use in service delivery will impact customers, operations, and employees. Likewise, it is vital to identify the psychological and physical factors which can influence acceptance of using AI devices in service delivery (Gössling, Scott, & Hall, 2021).

Regardless of the Covid-19 pandemic, some trends had and will continue to reshape the hospitality industry.

In Table 1, the top ten hospitality trends can be seen. These trends are defined according to the professors from Hospitality Management School, also known as "The École hôtelière de Lausanne" (EHL), located in Switzerland.

Defined trends are considered the main factors which hotels should follow to entice guests and be competitive in the market.

Table 1: Overview of top ten hospitality trends according to EHL

Trend	Description
Local Experience	These days, it is modern to offer a unique experience to the guest. Experiencing the local way of living in the county guests are visiting is sometimes one of their main goals. To respond to these kinds of requests, the hotel can offer local products and day trips, including visiting farmhouses and hiking tours, enabling their customers to participate in local activities.
Contactless technology and digitalized guest experience	In recent years, almost all businesses experienced the power and importance of digital and contactless services. Mobile check-in, voice control, contactless payments, and biometrics are essential for the hospitality industry. Investments in technology are inevitable.
Personalization	Hotels should treat their guests as individuals and create personalized emails, messages, and welcoming. Mangers are encouraged to use tools that can make massive email campaigns targeting a specific audience.
Experience economy and essentialism	In search of a unique experience and more adventurous trips, customers switch their focus from luxury to more essential and personalized trips. Future guests can choose whatever service and type of vacation they want with digitalised platforms. Thus, he or she can directly reserve a room without involving travel agents.
Generations X and Y	Younger generations, or the so-called millennial generation, has different views and needs than the older generation. Therefore, hotels need to adjust their offers for such guests, providing them with modern solutions and different facilities and programs.
Virtual and augmented reality	Being one of the most exciting trends in the hospitality industry, virtual reality enables guests to see or take a tour within the hotel/restaurant and, in that way, assess whether the hotel fulfils their requirements. In addition, by having 360° view videos on-site, hotels ensure that potential guests are fully aware of what facilities the hotel can offer.

Table continues

Table 1: Overview of top ten hospitality trends according to EHL (continued)

Trend	Description
Solo Travelers	Many people decide to travel alone, unencumbered, experience something new, step away from their comfort zone, interact more with strangers and make new friends. Because staff and guest interaction are essential for these kinds of travellers, management should aim to lower the barriers between the guest and staff to make it a more friendly, informal, but still respective atmosphere. In this way, solo travellers feel more comfortable and welcome.
New Hospitality skills and asset management	As the hospitality industry became more complex in recent years, many new job profiles, such as asset managers, were introduced.
Sustainability	An eco-friendly environment and sustainability have become the most critical trends in recent years. Customers are aware of environmental issues, and they are ready to invest more to support ethical behaviour. Examples of supporting sustainability are restaurants with vegan and vegetarian choices, hotels which already implemented innovative heating, and smart light bulbs. Sustainable materials for accommodation purposes, such as bedsheets and towels, slippers, and bathrobes, are also one way of acting more sustainable. Moreover, reducing and eliminating plastic, paper, and food waste will play a vital role in the following years. To keep pace with competitors and make our planet "healthier", businesses should continuously work on acting more sustainable.
Automation and technology	Technological development and automation are essential to reduce manual work and wait times and create general answers for recurring questions. To optimize and monitor revenues, reservations, trach channels and reputation, hotels use management systems such as SAP, big data solutions, and hiring programmers. These systems, predictive analytics, profiling of the customers, and integrated messages are all part of an incredible shift towards digital and automated.

Source: Masset & Weisskopf (2021).

1.3 Importance of customer satisfaction and the eWOM

A primary source for finding unbiased information in the tourism industry is user-generated content. When customers decide to book the hotel, they refer to the experience that previous guests shared in their reviews across social media and travel-related sites. The primary focus

of many literature studies has been to analyse these text reviews and, in that way, help to understand the customers' experience in the hotels (Kim, Bai, Kim, & Kaye, 2018). As the number of internet users and their content is constantly rising, companies must invest a lot of energy and time to satisfy their customers and expand customer lifetime value (CLV). Evaluating the CLV is among the crucial factors that will determine if the business will be a success or failure (Chuang & Shen, 2008). Hotel managers are therefore looking for the right way to understand which factors influence customer loyalty the most. Spotting these main factors can help them implement the right marketing strategies, which will then ensure constant loyalty and maximisation of their lifetime value. One of the main challenges for managers in hospitality right now is what to offer to their guests to maintain guests' loyalty and satisfaction (Rather & Sharma, 2017). Customer satisfaction and customer loyalty are not the same things. Customer satisfaction measures a customer's attitude toward a specific product, brand, or service. On the other hand, customer loyalty represents a set of attitudes and behaviours that a customer shows which demonstrate loyalty to a specific product, brand, or service.

In the hospitality industry, loyalty can be seen as repeat booking, choosing one hotel over a competitor. Moreover, many studies have proven that customer loyalty is essential for gaining a competitive advantage. Because of the stated, studies also claimed it is much cheaper for the company to maintain old customers than to attract new ones (Chuang & Shen, 2008; Awan, Siddiquei, Jabbar, Abrar, & Baig, 2015). Thus, companies or, in this case, hotels, should understand customers' needs, develop services and offers that meet those needs and should listen to and interact with their customers (Tatulli, 2019; Dawar, 2013; Tepeci, 1999). Once the customer is linked with the company or hotel, he or she will be unwilling to switch to another.

The two most potent websites for sharing traveller's experiences are Booking.com and Tripadvisor.com (eBizMBA Inc., 2021), in which the typical review consists of rating or numerical evaluation and additional textual evaluation (Godnov & Redek, 2019). On stated sites, users can add photos, list amenities and details about the hotel, track bookings and customers' preferences, and most importantly, people can write and respond to the review (Tripadvisor, 2020). Many potential customers will decide based on someone's review because those reviews are considered trustworthy, objective, and unbiased (Godnov & Redek, 2019). Constant availability and ability to use phone, tablet or any other device has led to a massive number of online reviews, which are highly likely to be influential. Thus, the responsible manager must recognize the influence and the predictive effect of the review (Tsang & Prendergast, 2009). In 2019, the number of reviews on Tripadvisor.com was 859 million (Statista, 2019). It can be expected that until 2025, online travel accommodation possibilities will be two or more times bigger than they are now (Bowtell, 2015), which means that by analysing reviews, the potential knowledge and valuable information about the customers could be at least doubled (Taşkın, 2016). With gained knowledge from reviews, hotel managers would be able to track customer satisfaction, improve efficiency

and effectiveness, and improve business. For example, the proper analysis may help find the delighted people willing to pay more, hence offering them a luxury package with higher prices. By analysing reviews, the hotels can also reduce the customers' churn, detect changes in guests' choices, and react fast on those changes (Chittiprolu, Samala, & Bellamkonda, 2021). As reviews are likely to influence future demand, it is possible to predict the future demand of a specific hotel.

Nevertheless, the problem with the current trend is that posted information does not always contain useful data since not all users have the same pattern of writing a review. In many cases, the information from reviews is thrown in various random directions, making it impossible to extract some meaningful knowledge from that information.

2 TEXT MINING IN THE HOSPITALITY INDUSTY

Chapter two provides main terms and definitions related to data science, business intelligence, data mining, and text mining while focusing on common and valuable text mining analysis.

Timely and accurate customer and competitor intelligence improve hotel effectiveness and guest satisfaction. Business intelligence (BI) plays a critical role when directors make operational and strategic decisions. It combines data mining, business analytics, data visualization, infrastructure and data tools, and the best practices for helping organizations make data-driven decisions more often. Companies know they have correctly implemented innovative BI solutions once they have a comprehensive view of their organization’s data, when they can use data to drive changes, once the company can instantly and quickly adapt to any market change, and once they can faster spot and eliminate inefficiencies (Mariani, Baggio, Fuchs, & Höepken, 2018).

In 1960, BI was defined as a system of sharing information among organizations, while nowadays, business intelligence solutions focus on flexible self-service analysis, empowered business users, governed data on trusted platforms, and speed to get any insight (Imhoff & White, 2011). BI solutions should enable easy and timely access to all critical operational data, provide predictive use of business operations, make informed business decisions regarding distribution, rate strategy, marketing campaigns, track daily, monthly, quarterly, or yearly profit and many more. In general, BI solutions should be used to smoothen the processes and, in that way, create more efficient organization (Ranjan, 2009).

During the last years, business intelligence has expanded and now it includes more activities and processes than before (Analiytiks, 2019). Some of the processes, which should be followed for better performance, are (Ukhalkar, Phursule, Gadekar, & Sable, 2020):

– Data preparation, which implies putting multiple data sources together, detecting the dimensions and measurements, organizing, and preparing them for data analysis.

- Data mining, which is used for uncovering trends in large datasets by using machine learning, databases, and statistics.
- Reporting, which is enabling stakeholders to make conclusions and decisions based on data analysis.
- Benchmarking and Performance metrics, which is comparing recent performance data to old data with aim of tracking performance against goals, usually by using customized dashboards.
- Descriptive analytics, where preliminary data analysis is used to find the root cause of the problem.
- Querying, in which business intelligence is trying to find the right answers from the datasets, while examining the specific questions.
- Statistical analysis, which serves for understanding why something happened, and how by taking the results from descriptive analytics to make further investigation with statistics.
- Data visualization, which is used for representing data in the more understandable way, while turning data analysis into visual depictions such as graphs, charts, and histograms.
- Visual analysis, in which visual storytelling is used to explore data and communicate newest insights.

All the stated actions are very important also for hotels, as they need to respond to the changing dynamics of the market in real-time and be the first to capitalise on emerging opportunities. Therefore, hotels need to track their guests' preferences, occupancy, and availability of hotel rooms. They need to spot the main competitors and track satisfaction based on online and on-site reviews. Nevertheless, in the current information explosion era, it is not rare to hear from hospitality experts that they are overloaded by data, but they do not have enough understanding and knowledge to easily extract useful information. The old-fashioned way of processing textual information required a significant investment of money, time, and human resources, as human actions were needed to gather the information, analyse them, and visualise them. Because of all the stated benefits of BI solutions, a special place in the IT sector is reserved for data science, especially for data-mining techniques, which are helpful in obtaining meaningful patterns and making predictive customer-relationship models from numeric data.

2.1 Text mining tools and techniques

Although broadly used, data mining can be applied only to structured, numeric databases. However, a lot of business information is represented in the form of unstructured or semi-structured data. Therefore, the innovative technologies used to extract useful information from large textual documents and integrate scrappy information into business intelligence databases are necessary. Thus, in those cases, text mining plays a key role. Although these terms are closely related, it is important to understand the difference between business

intelligence, data science, data mining and text mining. In Table 2, the main differences between business intelligence and data science including their main definitions are indicated. While in Table 3, the comparison between data mining and text mining is described, together with the main definitions.

During the last decade, text mining has been recognized as an important research area. According to the "Survey of text mining", conducted in 2013, 80% of the World's formal and informal documents are stored in the non-structured text (Ramanathan & Meyyappan, 2013). With new platforms, social media, and other online channels of communication, it can be assumed that this percentage is even bigger today.

Table 2: Business intelligence vs data science

Business intelligence	Data science
BI looks at business's historical data to discover patterns and trends to make better, informed business decisions.	Data science looks at both past and present data and uses that data to create as much impact as possible for your business.
BI takes an analytical approach to develop decision-support applications.	Data science uses predictive and prescriptive analysis to generate insights out of data
BI tools present their analytical findings in the form of dashboards, graphs, reports, charts, etc.	Data science is about using complex tools and statistics to predict future trends.

Source: Khillar (2020).

Table 3: Text Mining vs Data Mining

Text mining	Data mining
Text mining is the processing of text-based data from documents.	Data mining simply means knowledge mining from data.
Text mining is the process of extracting meaningful insights from unstructured text data.	Data mining is the sorting and extraction of meaningful information or data from large datasets.
Common data sources include social media, emails, messages, forums, news articles, and library databases.	Common data sources include data warehouses, databases, the World Wide Web, and other information repositories.
Applications include risk management, fraud detection, customer care, business intelligence, healthcare, spam filtering etc.	Applications include financial analysis, intrusion detection, spatial data mining, machine learning, soft computing, etc.

Source: Khillar (2020).

From the information written by Jeff Schultz (2019), only social media was producing 1,209,600 new data each day. Having that numbers in mind, these data are huge potential advantage if inspected and analysed correctly.

To process and understand these unstructured data, additional techniques are needed, in order to gain precious information, which will lead to potential knowledge. Knowledge discovery from textual databases, known as text mining, is the process of extracting nontrivial and interesting patterns or knowledge from unstructured documents (Hassani, Beneki, Unger, Mazinani, & Yeganegi, 2020). Text mining has a huge commercial potential because text is the most natural form of storing information. It is a multidisciplinary area which involves information retrieval, information extraction, text analysis, categorization, clustering visualization, machine learning, database technology, and data mining (Tan, 1999). To analyse customers' experience, in the past, hotels were placing cards in the rooms expecting the comment, conducted surveys of satisfaction, and according to gathered data, hotels would establish follow-up measures. Despite these efforts, the problem continued, as guests did not always want to share their experiences and provide feedback to the hotels. These days, guests have the possibility to share their experiences via social media, forums, blogs, travel sites and other online platforms (Berezina, Cobanoglu, Miller, & Kwansa, 2012a). To understand well their customers and improve hotel operations and general performance, hotel managers must utilize customer reviews. All the online materials can be extremely voluminous, and because of that, it would be hardly possible to peruse all the documents by hand. This is the moment when text mining and its analysis would be a great help. Text mining review analysis gives a greater and better picture, but it also ensures that each relevant hotel's aspect can be examined with the appropriate analysis. Using text mining in the right way, especially in the hotel sector, can lead to better business opportunities, higher customer loyalty and greater customer satisfaction (McGarrity, 2017).

Common steps of text mining are (Kwartler, 2017):

- Defining a problem statement and a text mining goal
- Collecting the unstructured data
- Doing text pre-processing and organising data
- Extracting data, which are then structured
- Analysing data
- Providing insights, recommendations, and analytical output

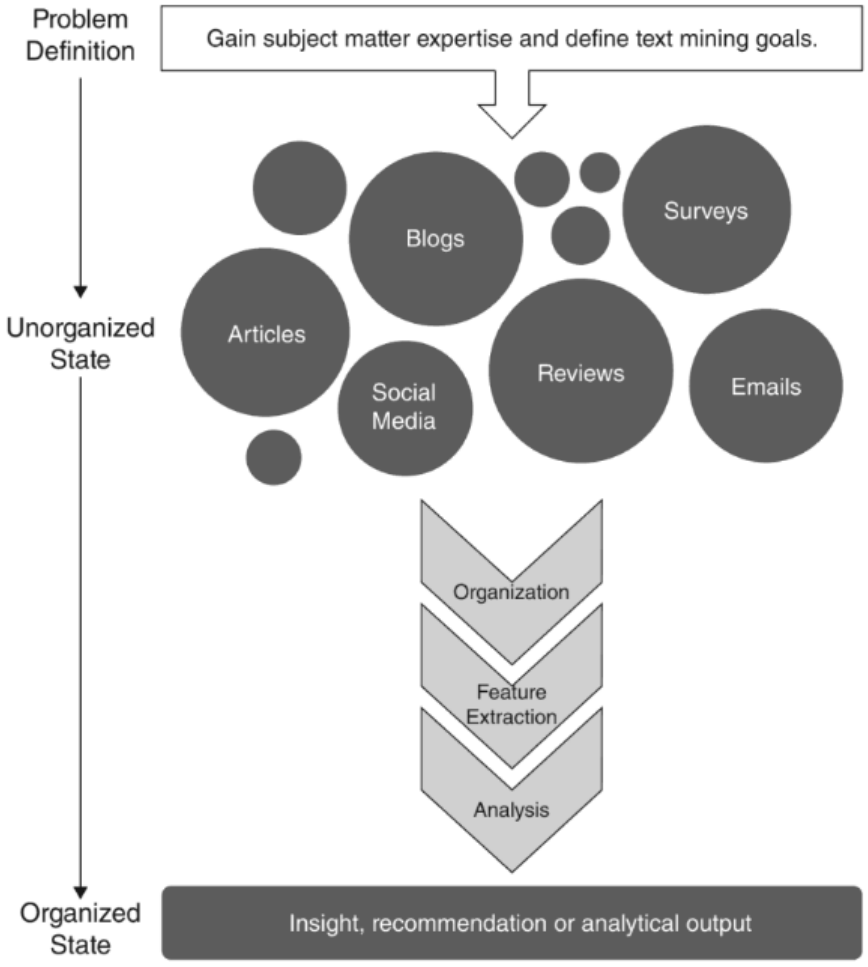
Figure 1 presents how the process or the basic workflow of text mining looks like. In the left part, the arrows indicate phases, while on the right side, the actions and outputs of the phases are described.

There are many tools and different ways and techniques of data retrieval which are used to extract knowledge and useful information from the unstructured text. Besides that, there is a

lot of analyses that can be done with the help of text mining. Some of the main processes and often used analyses are briefly described below.

Information Retrieval (IR) can be described as a software system that deals with the storage, retrieval, organization and information evaluation from document repositories, especially textual information. It is the activity of obtaining material that can usually be documented as unstructured nature. For example, Information Retrieval is the action when a user enters a query into the system. Google search engines are the most famous IR systems. Documents associated with a set of given words stored on the internet are recognized by Google search engines. Being assessed as an extension to document retrieval, it can be described as a place where the returned documents are processed with the aim of extracting the useful information critical for the user (Sagayam, Srinivasan, & Roshni, 2012). Information retrieval deals with a wide range of information processing, from information retrieval to knowledge retrieval (Inzalkar & Sharma, 2015). Because of the growth of WWW and more sophisticated search engines, information retrieval has gotten more attention in terms of research.

Figure 1: Processing of Text Mining



Source: Kwartler (2017).

Information extraction helps to identify and extract useful information from semi-structured or unstructured text. The crucial information such as name, organization, location, or gender is extracted without an appropriate understanding of the text. As information extraction is concerned with semantic information extraction, it could be described as the construction of relevant pieces of information taken from the text.

Categorization, also known as text classification and text tagging, assigns the documents to a predefined set of topics to sort the document into specific groups according to the document topic. Today, automated text categorization is used in many contexts, from traditional text indexing to text genre detection, automatic metadata generation, personalized commercials deliveries and many more (Agrawal & Batra, 2013). By using Natural Language Processing (NLP), a text classifier automatically analyses the text and, after analysis, assigns a set of pre-defined categories or tags based on its content. Text classifiers with NLP are an excellent alternative to structure textual data in a cost-effective, fast, and scalable way. Nowadays, text classification is allowing businesses to easily gain an understanding of data and build automated business processes. Stated below are the most common cases and examples of automatic text classification usage:

- Sentiment Analysis is the process which helps in exploring if text about the given subject is written in a positive or negative way (usually used for tracking the reviews, for brand monitoring purposes etc.).
- Topic Detection helps in identifying the right theme or topic of a particular text (for example, when analysing customer feedback, the information what is the correct review's subject can immediately be defined).
- Language Detection helps in detecting the language of a specific text.
- Clustering is characterized as one of the most important and most interesting in text mining. The main purpose of clustering is to find intrinsic structures in information and arrange them into smaller groups for further analysis and study (Dang & Ahmad, 2014). Therefore, clustering can be defined as an unconquered process which helps classify objects into groups known as clusters. The main challenge is to group collection without labels into the relevant cluster with no prior information available. Each label which is associated with objects is obtained uniquely from the data. To assist in data retrieval, it creates links between related documents. Furthermore, this allows retrieving related documents. Clustering can be used in many areas from biology to business intelligence, data mining and web search (Gupta & Lehal, 2009). In text mining, clustering can be used as a pre-processing step for algorithm which are operating on the detected clusters, or as a freestanding tool used to achieve data distribution.
- Summarisation can be very helpful for big companies, where people do not have time to read all documentation related to specific tasks. Thus, they decide to summarise and highlight the most important things in the text. It can be considered that parts of different documents or main points taken from the specific text contain many useful and important information related to the specific topic. Therefore, it can be said that text summarisation

represents the process of creating one version of the text by combining other texts, which then results in having more useful information for the end-user (Dang & Ahmad, 2015). Text summarisation includes various methods that use text categorization, for example, decision trees, neural networks, semantic graphs, regression models, swarm intelligence and fuzzy logic. The development quality of classifiers is variable, and it depends on the type of summarized text, which implies that all the stated methods have similar and frequent problems. (Dang & Ahmad, 2014).

2.2 Sentiment Analysis for analysing online hotel reviews

As information and different sources on the internet are mostly free and as almost 60% of the whole world population has an internet connection, people usually make decisions based on the things stated on the internet (Li, Meng, Jeong, & Zhang, 2020) Hence, it is nothing new that businesses, in general, want to know what other people think about their products, service, or campaigns. Therefore, opinion mining, also known as sentiment analysis, should be introduced in this part of the thesis. With the huge amounts of raw data on the internet, the popularity and significance of sentiment analysis have intensified in recent years (Bisio, Oneto, & Cambria, 2016). Nowadays, customers do online research on the service or the product they want to invest in, and according to comments, they can decide. Because of that, by conducting sentiment analysis is much easier to find a way of extracting meaningful information from data.

Collecting public opinion is significantly important in many areas such as politics, governance, business, tourism, scientific research, etc. Political parties might be considered as the ones who need to pay special attention to gathering and correctly measuring public opinion. A great example which explains this is the time when elections are about to happen. During that time, parties invest a lot into opinion polls, which in turn gives them the possibility to select the correct strategy (Chauhan, 2016). If some party does not know the public opinion about their objectives, that can lead to a disadvantage in the elections because other parties could do better research and be more informed. Measuring personal opinions with such accuracy was definitely not possible in the period before the internet, when researchers used surveys and polls. Having the possibility to conduct sentiment analysis has changed the world. It changed the way of doing every kind of business, starting from politics and sports to other industries such as energetics, tourism and others. Therefore, hotels are able to gather the information about their service and analyse the gathered information faster, cheaper and more accurately than ever before, they just need the right and qualified workforce.

The process of identifying positive or negative sentiment in the text is called sentiment analysis or opinion mining. It often can be used by companies to discover sentiment in social data, measure brand reputation, and realize what customers like. Since consumers can express their feelings and thoughts on the internet without boundaries, sentiment analysis is

starting to be an essential tool for understanding that sentiment. By analyzing customer feedback automatically, companies could learn what makes their clients upset or happy, and in that way, companies can tailor services and products in order to meet their clients' needs (Berezina, Cobanoglu, Miller, & Kwansa, 2012). For example, the manager can use sentiment analysis if he or she wants to automatically analyse 10.000 reviews about the hotel. This could bring manager specific knowledge about the guest preferences and favourite meal or favourite amenities that the hotel offers. It can also bring managers precious information about what customers do not like and what could be improved. Sentiment analysis focuses on polarity (positive, negative, neutral), but it can also focus on emotions and feelings (happy, angry, excited, sad, etc.), urgency (urgent, not urgent), and even intentions (not interested or interested). Usually, when a hotel review should be analysed, managers would like to know which specific features or aspects people are mentioning in a negative, neutral, or positive way. That is where aspect-based sentiment analysis can help. For example, in the review: "The room was very clean, and breakfast was great", an aspect-based classifier should determine that the review expresses a positive opinion about the aspects room (stating: the room was very clean) and food/drinks (stating: breakfast was great). In Figure 2, it is presented how do the sentimets towars aspects can be defined.

Figure 2: Sentiment analysis of hotel review

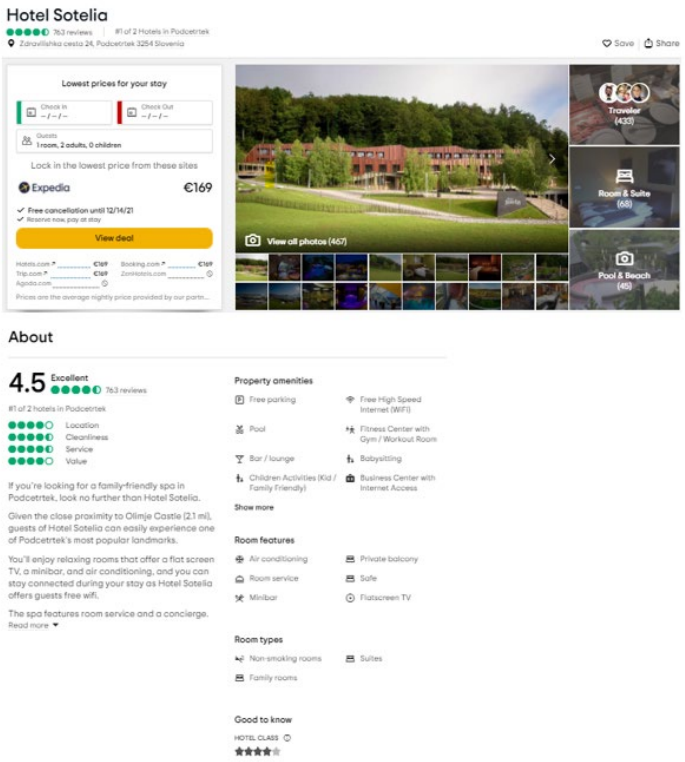
Source: Dabhade (2021).

Sentiment analysis is crucial for the hospitality industry, especially in the hotel sector, where guests express thoughts related to their experiences. Without this model, managers would go manually through reviews and read them one by one. Therefore, by conducting sentiment analysis, businesses save time and money. The positive side of sentiment analysis is that it can be done in real-time, which allows companies to predict if an angry customer is about to churn. Sentiment analysis could find potential gaps in the business and, in that way, get better insight information and improve accuracy. Figure 3 represents the picture on the official Tripadvisor website, where hotel Sotelia is searched. Tripadvisor already has pre-defined four aspects, which are: Location, Cleanliness, Service and Value and these recognized aspects are just rated with stars based on the reviews. These stars can be useful for a general picture of the hotel while future guests are searching for the next trip, but if the hotel wants to improve business or its services, these kinds of ratings will not help. Thus, using sentiment analysis would be a better choice.

There are many different types of sentiment analyses, which are used depending on the way of interpreting customers' feedback and opinion. Below are mentioned some of the most popular sentiment analyses:

Fine-grained Sentiment Analysis is used in case the polarity of the sentiments is very important to understand the client's needs. Here polarity can be defined from very positive to very negative or, in other words, from strongly positive to strongly negative.

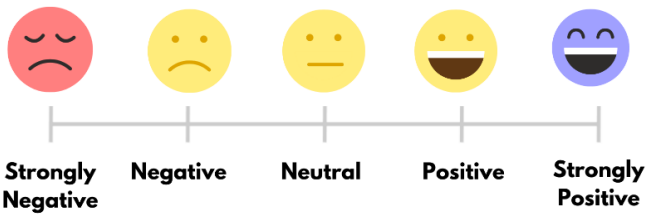
Figure 3: Tripadvisor website, hotel Sotelia



Source: Tripadvisor (2021).

Figure 4 presents how sentiments polarity can be defined in the fine-grained sentiment analysis.

Figure 4: Fine-granted Sentiment Analysis



Source: Rao (2019).

Usually, fine-granted Sentiment Analysis can be used to interpret hotel reviews in the different way than rates. For example:

- 5 stars review rating would be equal to strongly positive
- 1 star review rating would be equal to strongly negative

But it should not be forgotten that customers can give five stars review and write something negative about food, and thus, that would not be a strongly positive comment. In other words, many methods are based on star rating classification and detection, but when the analysis of review is done, the rating does not always show the accurate number of stars and, therefore, does not provide an accurate sentiment measure. This can also be stated as one of the barriers of using text mining while focusing on review rating (Maks, I., & Vossen, P., 2013).

Emotion detection can be used to detect emotions towards a specific product, service, or news. A wide variety of emotions can be reviled, but some of the most common are happiness, sadness, anger, frustration, and others. Many systems for emotion detection use a defined list of the words and defined emotions, also called lexicons or they use complex machine learning algorithms.

Aspect-based sentiment analysis is the sentiment analysis (ABSA), which aim is to extract the most important aspects of a specific product/service or entity and, at the same time, predict the polarity of each separate aspect in the review. Bing Liu, in his work from 2012 called “Sentiment analysis and opinion mining”, defines sentiment analysis as “the field of study that analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes” (Liu, 2012, p.7). Sentiment analysis is usually categorized into three levels which are document, sentence, and phrase or aspect (Liu, Xu, & Zhao, 2012). The first two categories are not able to provide enough information to do the decision making. Nevertheless, aspect-based sentiment analysis can be used to obtain information for decision making (Madhoushi, Hamdan, & Zainudin, 2019). From the reviewer's point of view, the feedback will be given for a specific aspect of the product or service which made the biggest impression on him or her. Thus, it cannot be stated that the reviewer likes or dislikes that service or product totally. Even though the reviewer’s overall opinion on the service or product can be negative, neutral, or positive, the reviewer typically writes all three in one review – negative, neutral, and positive towards various aspects of the service or product. For example, the reviewer can give five stars to review, meaning this overall opinion is positive, but at the same time write about room and cleanliness in a positive manner, while mentioning that food could be better and therefore characterize food aspect as negative. Moreover, the reviewer can state in the same review that he or she expected that the wi-fi connection would be better in the lobby of the hotel, but that connection in the room was great, meaning that wi-fi would be recognized by the program as neutral. Nevertheless, the overall rate would be five stars or be characterized as positive in general (Madhoushi, Hamdan, & Zainudin, 2019).

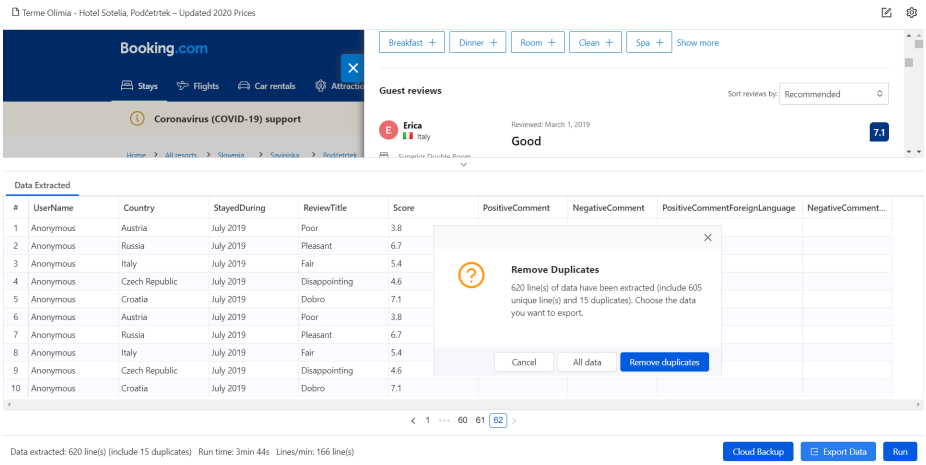
3 METHODOLOGY

Chapter 3 is dedicated to describing the methodology used in this research and the presentation of hotel Sotelia. In chapter 3.1, besides the methodology used, it is explained how the data for this research were gathered, but also the flow of the interview with the manager of hotel Sotelia is presented, while the hotel and its characteristics is presented in chapter 3.2.

3.1 Research methodology

As stated in the introduction, in terms of methodological approach, the thesis is of qualitative and quantitative nature. The rationale for choosing a mixed research methodology is that data posted online are, most of the time, trustworthy and accurate measurements of customer satisfaction. As the main point of the master thesis is about text mining, it was inevitable to provide strong theoretical background but also to conduct quantitative research, including several analyses. Concerning quantitative methods, as partially mentioned in introduction part, data were gathered from different online travel-related websites dating from January 2009 to January 2020. The text mining approach was put on to inspect key insides from the reviews. Moreover, these insights were analysed in terms of their aspects. Tripadvisor.com, Booking.com and Google.com were used as a source from which the data were gathered. Data from Tripadvisor were gathered with the help of the Python programming language, which is used for web scraping. Once gathering was done, data were extracted into an excel file. It is worth mentioning that review translation was done using code, so the reviews which were originally, for example, in the Italian language, were automatically translated into the English language. Data from Booking.com are gathered with the help of Octoparse. An illustration of how the data extraction with the help of Octoparse looked is presented in Figure 5.

Figure 5: Octoparse final extraction



Source: Own work.

Octoparse is the software for web scraping, mining, and data analysis, which turns unstructured or semi-structured data into structured data, and if using it, almost no prior coding knowledge is needed. The interesting thing is that software offers templates for the data extraction, but it is better to create a specific template to get the needed data. The software is user-friendly, and it is relatively easy to learn the basic steps which are necessary to understand how web scraping can be done.

Octoparse works based on XPATH, which uniquely identifies the HTML element on the page. The most attractive feature of Octoparse is its ability to scrape the web on a huge scale at the same time using distributed computing. Octoparse works well with both static and dynamic websites, including those that use Ajax. Ajax is short for Asynchronous JavaScript and XML (Extensible Markup Language), and it represents the set of web development techniques which are using several web technologies on the client-side to build asynchronous web applications (Turc, 2019). Web applications that use Ajax can transmit and get data from a server asynchronously without interfering with the existing page's appearance and behaviour. Important to mention is that Ajax allows web pages and web applications to change content dynamically without having to reload the entire page. Site Booking.com works on the same principle, and with the help of the Octoparse tool, web-scraping was done.

Moreover, data cleansing was done based on several criteria, in a similar way as for data from Tripadvisor.com. For instance, all reviews were translated into the same language, which is English, with the aim to reveal the important words and not the whole grammatically correct sentence, duplicates were removed, and a standard rate was established. Regarding translation in general, in some situations when translation is not as good as it should be, the slang words or irony were not recognised, which could cause a potential discrepancy between the original review and translated review. Such misalignment can cause problems to the software, in a way that the software cannot recognise the right meaning of the specific word and, due to that, can make mistakes in the determination of the right aspect polarity. This can also be seen as a limitation in text mining, and such a situation must carefully be taken into consideration while interpreting the overall results. These kinds of limitations were presented in chapter 5.2, where the answer to the second research question was provided.

Working with the Octoparse, as the additional method of data gathering, was chosen to demonstrate a potential way out of a situation where a hotel or any other company does not have enough resources to hire a good analyst or does not have employees with programming skills, but still decides to let its managers analyse the reviews. With the help of free web scraping tools, the entire process can be done faster.

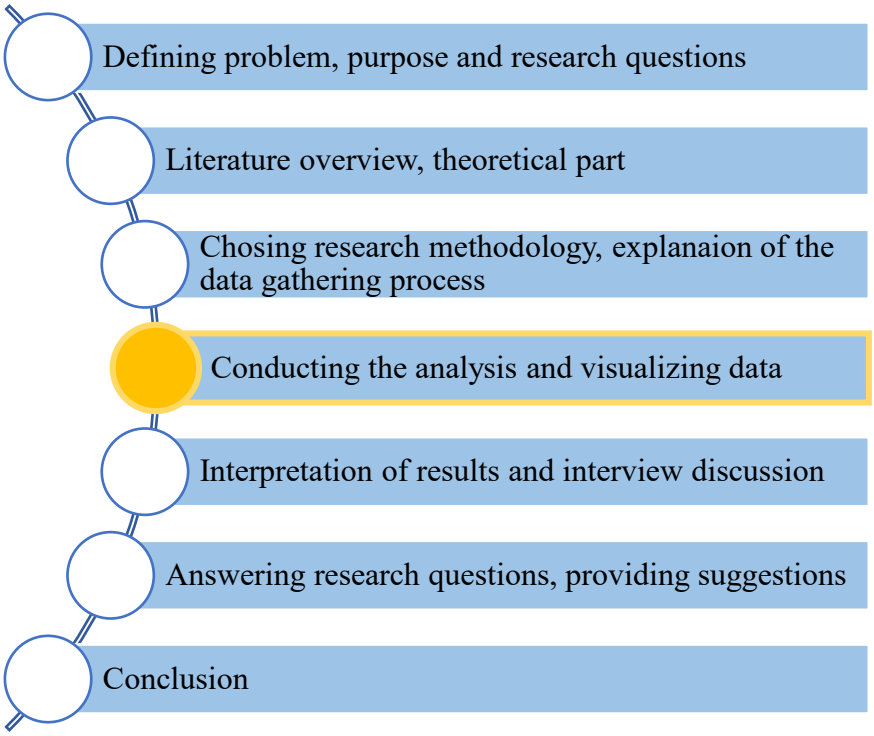
Once all data were collected in the same file, they were analysed using RapidMiner and Microsoft Excel, including general data analyses but also statistical analysis described in the form of a contingency table.

The qualitative method was mainly used for the theoretical part, where the primary resources were scientific journal articles. However, other sources such as e-books, websites and conference proceedings were used to obtain additional information and thus complete the theoretical part. The qualitative method was also required to describe all results of the analyses, interpret the results from the interview, and provide potential improvements.

The one-hour-and-a-half hour-long online interview took place in January 2022, and it was conducted with one of the representatives from the top management of the hotel Sotelia. The interview was divided into four parts. In the first part, the general questions about the hotel were discussed. The second part discussed why customer satisfaction is essential for hotel Sotelia and whether management puts much energy into investigating each review. The third part of the interview was reserved for describing how managers are currently managing reviews and how they approach potential issues regarding negative comments. The last part of the interview was based on questions about the aspect-based sentiment analysis that was used in this thesis, after which conclusions with recommendations were given to the representative. After all, results of randomly selected reviews are compared with the original review posted on travel site to reveal whether the analysis was reliable, and if it these types of analyses are worth conducting in hospitality industry.

In Figure 7, the visual representation of the process related to thesis can be seen. As the first three processes were already mentioned formerly, the phase of analysing data can be introduced.

Figure 6: Research flow



Source: Own work.

Data analyses were conducted using RapidMiner Studio, where several different operators were used. These operators will be examined in detail in chapter 4.1., where the explanation of the process in RapidMiner will be described. Once analyses in RapidMiner were done, the data were extracted into a Microsoft Excel file, where the rest of the analysis and statistical tests were made. The main three inputs needed for the further analyses were:

1. Rate (star rating scale)
2. Date when the review was posted on the website
3. English translation of the original review.

3.2 Description of Hotel Sotelia – Podčetrtek, Slovenia

Hotel Sotelia is a four-star hotel located in Slovenia in a small place called Podčetrtek. Hotel is surrounded by green areas, and it is brightened by sunlight coming from Obsotelje and Kozjansko hills. This place is known for its good local service, and it is an excellent choice for day trips. Destination Podčetrtek joined the Green Scheme of Slovenian tourism in 2017. Creating five-star boutique tourist experiences and following the national tourism development trends allows visitors a healthy and active stay in the "green" ambient, achieving personal fulfilment and inner peace (OECD, 2020). The hotel Sotelia can be highlighted as the brightest pearl of the destination Podčetrtek, as it is characterized as a hotel made for wellbeing.

Figure 7: Hotel Sotelia

Source: Terme Olimia (n.d.).

Because of the modern design, it is easily spotted, and because of its facilities and good service, it is well known among the guests, who are often visiting wellness and spas. As a result, it has been recognized a few times as one of the best hotels in Slovenia. Wellness hotel Sotelia is different from others as it is nature friendly, modern but still calming and equipped with expressive interior pieces. Moreover, the hotel is connected to Wellness Orhidelia, pools at Hotel Breza and Family wellness Termalija via an underground hallway. To enter the SPA hallway and wellness centre, guests are handed electronic bracelets,

allowing them to move around SPAs and enjoy infinite thermal and physical pleasures. Besides these amenities and facilities, the hotel offers a variety of wellness amenities and a few restaurants in the complex. For entering rooms, guests use electronic key cards.

According to the hotel's description, it can be concluded that the hotel's mission is to provide a relaxing and unforgettable experience along with a great food offer. This means that aspects related to food, room amenities, and facilities must be commented widely. If so, by conducting the analyses, it can be revealed if guests are mentioning the most critical aspects positively or negatively.

The hotel's reputation is very high on the most popular travel-related websites, and according to the rate, it can be said that the management of the hotel is doing an excellent job. However, some text mining analyses must be done to ensure that guests are satisfied.

4 ANALYSIS AND INTERPRETATION OF RESULTS

In chapter 4, detailed analyses based on collected data are presented, along with the interpretation of results and potential limitations of such analysis. Additionally, the most commented aspects are discovered and examined. Therefore, it was examined if the hotel is currently working according to their mission and description stated on the website. The purpose of the analyses was to look at the bigger picture of implementing text mining solutions for review checking and reveal if this approach can lead hotels to potential competitive advantage and if managers can rely on these types of analyses and, consequently, rely on results.

4.1 Explanation of the processes done in RapidMiner

RapidMiner, which is a software/platform that offers an integrated environment for data preparation, deep learning, text mining, machine learning and predictive analytics (RapidMiner.com), was used for detecting the aspects and finding their sentiments,

Users can enter raw data, such as databases and text, into the program, which is then analysed automatically and intelligently on a huge scale. Using template-based frameworks, which deliver results fast, the errors are reduced, and there is no need for programming. Nonetheless, some basic programming logic is needed to understand the whole process better. However, there is no need for coding if conducting similar analyses such as aspect-based sentiment analysis.

RapidMiner has a graphical user interface for creating and executing analytical workflows. These workflows are called "Processes," and they are made up of numerous "Operators." Within the process, each operator completes a specific task, and the output of one operator becomes the input of the next (Norris, 2013). The potential of RapidMiner can be increased by purchasing additional plugins from the RapidMiner Marketplace. In addition, developers

can use the RapidMiner Marketplace to design data analysis algorithms and share them with the community.

For this research, it was vital to use the Aliyan extension from the RapidMiner marketplace, as the primary research is based on the Aspect-Based Sentiment Analysis, and with this extension, it was much easier to proceed with text analysis. The reason why the fundamental sentiment analysis was not chosen is that it was not enough to provide just a polarity of review in the case of analysing hotel reviews. However, it was needed to have the operator which explores the sentiment and polarity towards specific aspects. Thus, aspect sentiment-based analysis was chosen. Defining the main aspects with the program's help was crucial, as these are vital information for understanding customers' opinions about a particular hotel and for building the knowledge about what customers typically mention in reviews.

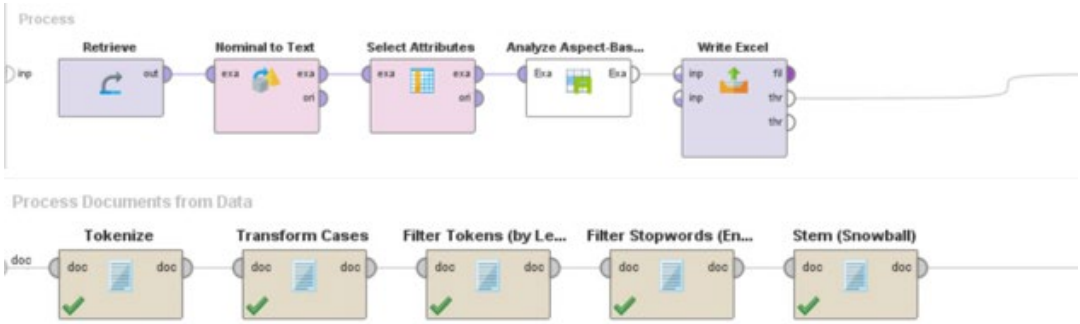
To get the aspects to which specific sentiment will be assigned, several different operators were used:

- Retrieve, as the operator which is used for accessing the stored information and loading them into the process.
- Nominal to Text, as the operator responsible for converting all nominal attributes to text. All nominal values are used as strings values of the new attribute. If there are missing values in the nominal attribute, the new value will also be missing.
- Tokenization was also done where non-letters were tokenized, cases were transformed (lower cases), tokens were filtered by length (min=4, max=25), filter was used to eliminate stop words and stem (snowball).
- Select Attributes operator was used to select attributes of the dataset and eliminates the other not needed attributes.
- Analyse aspect-based sentiment operator was used for analysing product or services or in this case reviews from websites, by analysing the sentiment of the review in relation to each of the aspects of the product or service mentioned in a review. For using such operator, it was necessary to create a new connection and selected full review as an input attribute and hotel as business domain
- Write excel operator was used for creating new dataset in an Excel file.

In Figure 8, the first process (data preparation and tokenization) that enabled the collection of aspects is presented. Having this process ready enabled future analysis.After the aspect sentiments were collected, the preparation phase for the following process began.

In the file gotten from the first process, the critical column consisted of numbers, semicolons, and aspect sentiments for each review were added. Then, aspects were separated with a comma sign. The excel formula (concatenate) was used to put the stated data into the same column (ex. 3; cleanliness: positive, food/drinks: positive, room amenities: positive). This step is important because the new process began with the create example set operator, where the generator type was comma separated text.

Figure 8: First process in RapidMiner (data preparation and tokenization)



Source: Own work

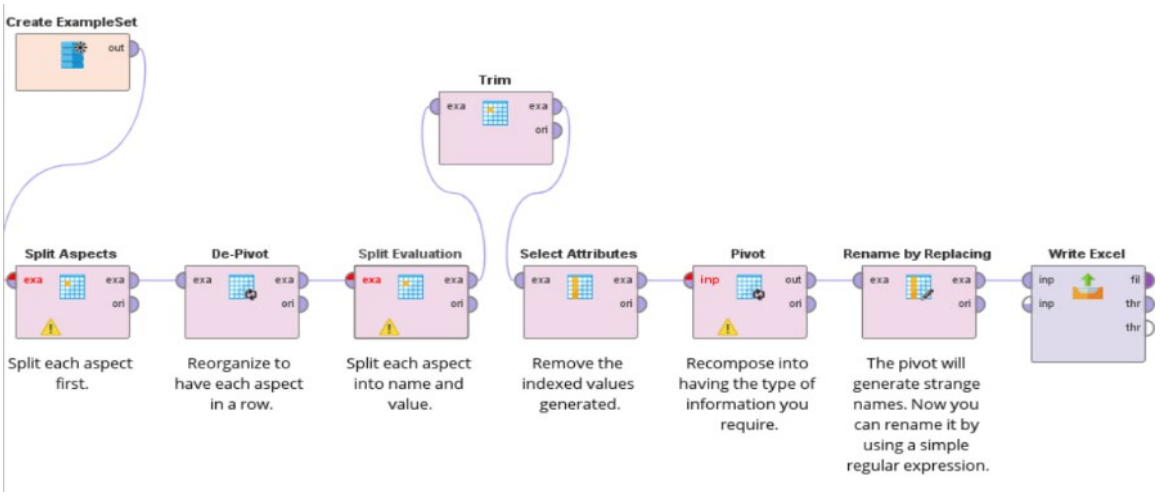
The second process aimed to get the file, where the names of the columns were aspects, and the rows were sentiments. In order to create such a file, nine new operators were used:

- Create ExampleSet was used to create dataset with user-specified characteristics and examples.
- Split operator was used for creating new attributes from the chosen nominal attributes. This was done by splitting the nominal values into parts in accordance with a specific criterion. In this case, operator split is used for splitting the aspects.
- De-pivot - operator was chosen to convert the examples of the picked attributes into examples of a single attribute, and in that way transforms the dataset.
- Split operator was used for splitting each aspect into name and value.
- Trim operator was used for removing spaces from the values of the chosen nominal attributes.
- In this case, operator Select Attributes was used for removing the generated indexed values.
- Pivot - operator was used to recompose data into the type of information required. Pivot operator can make a pivot table, which summarize the data in a bigger table by restructuring it into groups and calculating averages, sums and other statistics for each group.
- Rename by Replacing operator was used to replace parts of the attribute names with the unique replacement and in that way rename attributes.
- Write excel operator was used again to creates new dataset in the excel file, where the further analysis will take place.

After the execution of the second process, the further analyses were done in excel, where it was meant to find the most positive and most negative aspects in reviews commented by guests and build the further research.

The second process in RapidMiner can be seen in Figure 9.

Figure 9: Process of getting aspects in RapidMiner



Source: Own work.

The final file contained aspects mentioned in reviews and their polarity (positive, neutral, negative). Each aspect had its own column in the file, which was very convenient for conducting further analysis in excel. Several types of data were collected from reviews.

The main columns from the file were:

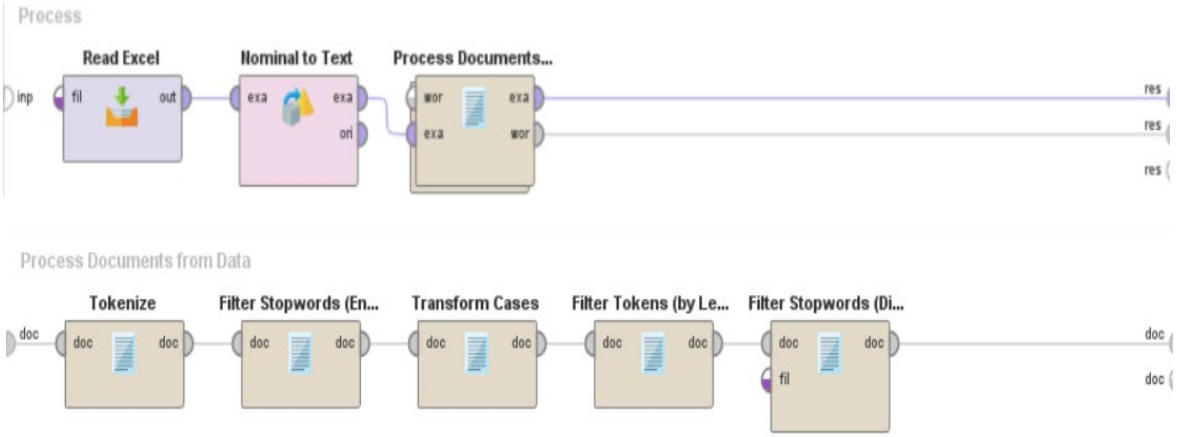
- Id, which was in the form of number from 1 to 1095
- Date when the review was posted
- Site from which the review was taken
- English translation of the original review
- Rate given by guest who stayed in hotel Sotelia
- Aspects relevant to the specific review
- 15 aspects recognized by RapidMiner, divided into separate columns, with specified polarities for each aspect.

RapidMiner identified 15 different aspects which could be found in reviews, and those were: Beds, Cleanliness, Comfort, Customer Support, Design, Facilities, Food/drinks, Location, Payment, Quietness, Room amenities, Staff, Value, View, and Wi-Fi.

One more analysis, which was done is RapidMiner was word frequency. The words from the whole dataset were counted, which enabled the identification of the most frequently used words in reviews posted on the mentioned sites. In Figure 10, the process of getting the most frequently used words can be seen.

From the results, it was evident that words with positive connotations were used in most cases. These words were primarily adjectives such as good, nice, excellent, clean, and other.

Figure 10: Word Frequency analysis done with RapidMiner



Source: Own work.

However, there were also many nouns and verbs connected to the hotel's mission and core business, such as room, staff, breakfast, pools, saunas, relaxation, eating, swimming, etc. In Table 4, the most frequently used words recognized by RapidMiner can be seen.

Table 4: Word Frequency Analysis

Word	Total Occurrence	Document Occurrence
Good	495	366
Room	494	347
Staff	487	436
Food	425	371
Nice	390	298
Rooms	374	335
Excellent	371	286
Clean	362	317
Orhidelia	346	263
Dinner	325	268
Breakfast	322	294
Great	318	259
Beautiful	301	246
Pools	301	227
Wellness	296	222
Saunas	272	205
Buffet	271	235
Sotelia	258	211

Source: Own work.

Additionally, a word frequency cloud was made to inspect the validity of the analysis done in RapidMiner.

From the Figure 11, it can be concluded that RapidMiner recognized the right aspects in terms of words occurrence, showed in the reviews.

Figure 11: Word Cloud

Source: Own work.

The standardise review rate was established as part of the cleansing for further research. Since these websites have different rate scaling systems, it was necessary to determine the rates in the same way, defining rates from one to five. Therefore, if the website had a defined rating scale from one to ten, the standardised rates were defined as follows:

- One and two were defined as rate one
- Three and four were defined as rate two
- Five and six were defined as rate three
- Seven and eight were defined as rate four
- Nine and ten were defined as rate five.

The logic behind scale starts with the lowest rate, which is one and ends with the highest rate, five. When it comes to descriptive rates, considering the new standardised rate scale, they were categorised into positive, neutral and negative, where:

- Reviews rated with standardized rates one and two were defined as negative
- Reviews rated with a standardized rate tree were defined as neutral
- Reviews rated with standardized rates four and five were defined as positive.

In Table 5, the rating scale is presented for easier navigation and a better understanding of the whole rating system established by different websites, but also the newly defined standardized rates and descriptive rates.

Table 5: Explanation for rate standardization

Standardized Rate	Tripadvisor.com	Booking.com	Google Reviews	Descriptive rate
1	1	1 - 2	1	Negative
2	2	3 – 4	2	Negative
3	3	5 – 6	3	Neutral
4	4	7 – 8	4	Positive
5	5	9 - 10	5	Positive

Source: Own work.

4.2 Analyses performed in Microsoft Excel

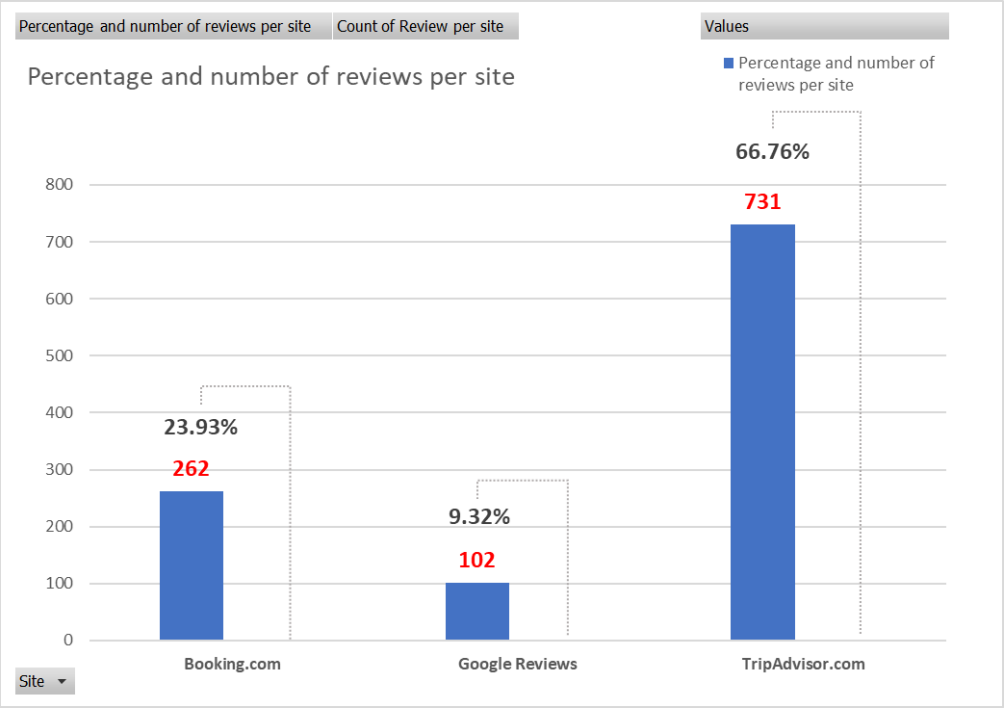
Before diving deeper into analyses, it must be mentioned that only reviews with textual explanation and rate given were counted and analysed for aspect-based sentiment analysis. All reviews, 1095 in total, were used to analyse aspect sentiments and polarity, make contingency tables and conduct other analyses. Precisely, 262 reviews were collected from Booking.com, 102 from Google.com, and 731 from Tripadvisor. Therefore, the highest impact on the analysis, in general, had the reviews from Tripadvisor.com, as they were written in the best suitable way. From these statistics, it can be easily spotted that people are a bit more engaged in writing reviews on Tripadvisor.com and Booking.com than on Google.com.

Nevertheless, if all reviews are considered, meaning reviews with textual explanation and rate and reviews having only rate, the difference between a number of reviews on Tripadvisor.com and Booking.com was not very big. The reason is that reviews from Booking.com in the vast majority of cases, contained only rate, which was not very helpful for analysis such as aspect-based sentiment analysis, as it was not possible to discover hidden patterns in reviews. However, these rates were helpful to inspect on which travel website hotel Sotelia has the highest rate and compare rates on all three websites to see if the hotel's reputation is consistent on different channels.

While checking the reviews, it was noticed that the management of hotel Sotelia invests much energy in replying to the reviews on Tripadvisor.com and Booking.com, which is the parameter showing the hotel is taking care of their guests and shows that the hotel is interested in guest's experience.

In Figure 12, a graphical representation of the number of reviews and percentage of the total number of reviews containing textual descriptions can be seen.

Figure 12: Total number of textual reviews per site and their percentages



Source: Own work.

Moving forward to the simplest analysis, reviews having just a few words, from which it was not possible to determine the aspects, are presented in Table 6. These reviews are gathered from Google.com and numbers in Table 6 indicate how many reviews were rated from one to five on Google.com.

Table 6: Number of Google Reviews with only a few words and rate provided

Rate	Number of reviews
3	1
4	4
5	35
Grand Total	40

Source: Own work.

It can be concluded that on the Google.com, people were in general satisfied with their stay and described it as excellent or wonderful. To be more precise out of 40 people 35 gave the excellent rate and only one person rated the hotel with three stars. To get the impression how do these reviews looked like, examples are provided in Table 7.

The overall hotels' grade on Google.com is high 4,6 out of five.

Table 7: Examples of Google Reviews with a few words and rate provided

English Translation	Original Rate	Descriptive Rate	Standard Rate
Well it was good	3	Neutral	3
Greaaaat! Love it	4	Positive	4
Fairy tale.	5	Positive	5
Fantastic	5	Positive	5
Great	5	Positive	5

Source: Own work.

The analysis for Booking.com was done in the similar way, for the reviews without any specific textual explanation towards their satisfaction. According to the analysis, there are 345 reviews on Booking.com, which did not contain long descriptions. These reviews had only one or two words, and rate given (example: "Very good", "Relaxing stay" etc.). It can be concluded that also on the Booking.com, most of the travellers were more than satisfied with hotel. Only few of them rated hotel with rate two and three, and the rest of the guests gave only positive rates, four and five. Nevertheless, as mentioned, since there were no additional descriptions available in these reviews, it was hard to understand why the customer rated the hotel with such rates. In Table 8, the number of reviews from Booking.com is provided and divided by the given rates. Also in Table 9, there are examples of the Google reviews provided, containing only a few words.

Table 8: Number of Booking.com reviews with a few words and rate provided

Rate	Number of reviews without aspects per rate
2	4
3	9
4	72
5	260
Grand Total	345

Source: Own work.

Table 2: Examples of Booking.com reviews with a few words and rate provided

ID	Translated review	Descriptive rate	Standard rate
1096	Poor	negative	2
1100	Disappointing	negative	2
1222	Fair	neutral	3
1360	Pleasant	neutral	3
1365	Exceptional	positive	5
1367	Awesome👍👍👍	positive	5

Source: Own work.

Nevertheless, when the full description of customer satisfaction with a product or a service is available, it is much easier to understand which aspect makes the customer satisfied and with which aspect the customer was not satisfied. Thus, as stated in the previous chapters, the company offering a product or service can improve customer satisfaction by listening to customers' wishes and reading reviews. If the reviews with descriptions were added to the previous calculation, this percentage would be even higher, and according to Booking.com in 87% of all cases, guests who stayed in the hotel claimed that their experience was very positive and that they were satisfied. This short analysis was not needed for reviews from Tripadvisor.com as the reviews were almost all suitable for aspect-based sentiment analysis. Having a final file with all aspects listed, by using pivot tables and statistics in excel, the main analysis was done. In these analyses, it was inspected how many times the aspects are displayed in the dataset. Using function count enabled counting the overall number of negative, positive, and neutral appearances of specific aspects. Additionally, the percentage of polarity (positive, negative, and neutral) is calculated for each aspect according to the number of total appearances of inspected aspects and the overall number of reviews. Also,

Table 10 is indicated in how many reviews each aspect is characterized as positive, negative, or neutral, whereas the total indicates how many reviews these aspects are shown in file (column Total). Also, in percentages, how often these aspects were mentioned in reviews (considering all reviews with text and aspects, thus all positive, negative, or neutral aspects).

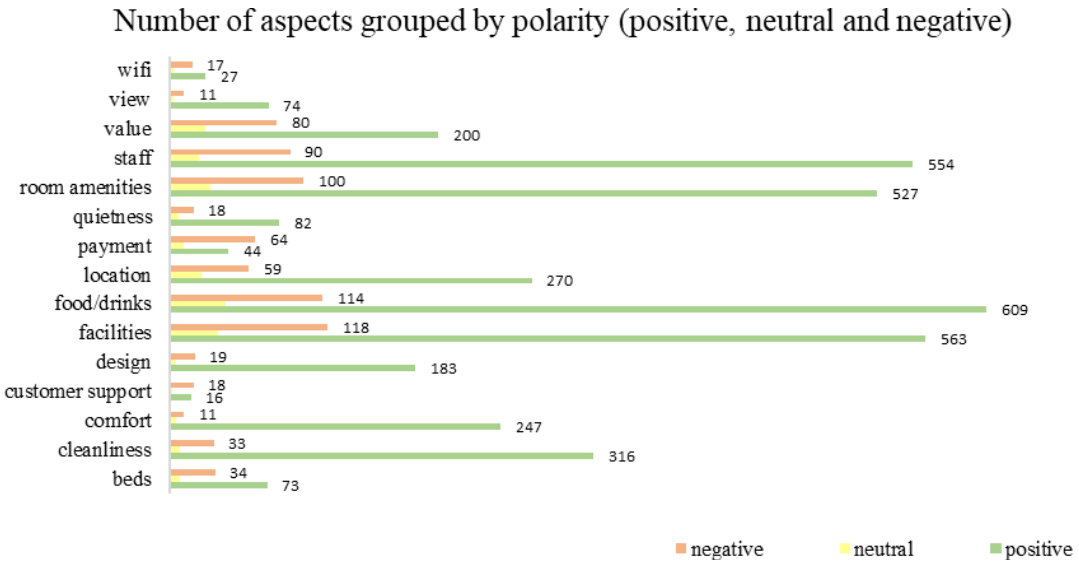
Table 10: Tripadvisor.com numbers and percentages of aspects and their polarities

Aspects:	Negative	Neutral	Positive	Total	% out of Total 1095	% of negative	% of neutral	% of positive
Beds	34	8	73	115	10.50%	3.11%	0.73%	6.67%
Cleanliness	33	8	316	357	32.60%	3.01%	0.73%	28.86%
Comfort	11	5	247	263	24.02%	0.95%	0.43%	21.26%
Customer Support	18	2	16	36	3.29%	1.64%	0.18%	1.46%
Design	19	4	183	206	18.81%	1.74%	0.37%	16.71%
Facilities	118	36	563	717	65.48%	10.78%	3.29%	51.42%
Food/drinks	114	42	609	765	69.86%	10.41%	3.84%	55.62%
Location	59	24	270	353	32.24%	5.39%	2.19%	24.66%
Payment	64	11	44	119	10.87%	5.84%	1.00%	4.02%
Quietness	18	7	82	107	9.77%	1.64%	0.64%	7.49%
Room amenities	100	31	527	658	60.09%	9.13%	2.83%	48.13%
Staff	90	22	554	666	60.82%	8.22%	2.01%	50.59%
Value	80	27	200	307	28.04%	7.31%	2.47%	18.26%
View	11	3	74	88	8.04%	1.00%	0.27%	6.76%
Wi-fi	17	3	27	47	4.29%	1.55%	0.27%	2.47%

Source: Own work.

Also Figure 13 is provided to demonstrate how many times each aspect is characterized as positive, negative, or neutral.

Figure 13: number of aspects by polarity, Tripadvisor.com data



Source: Own work.

It can also be stated that people who visited hotel Sotelia were focused more on the aspects, where the percentage of mentioning of aspect mentions is higher than 30%. The special focus was put on four below-mentioned aspects, as they are recognized in more than 60% of the reviews:

- Food/Drinks
- Facilities
- Staff
- Room Amenities.

In Tables 11,12,13, and 14, polarity of sentiments is indicated in the first column, followed by number of each polarity related to specific aspect considering the whole dataset. The third column present the percentage of polarity toward specific aspect considering the total number of appearances only for that specific aspect. In the last column, the percentage toward specific aspect, but considering the whole dataset (1095 reviews), is provided.

Table 11: Sentiments towards food/drinks indicated in numbers and percentages

Polarity	Count of food/drinks	% of food/drinks	% out of Total reviews
negative	114	14.90%	10.41%
neutral	42	5.49%	3.84%
positive	609	79.61%	55.62%
Grand Total	765	100.00%	69.86%

Source: Own work.

Table 12: Sentiments towards facilities indicated in numbers and percentages

Polarity	Count of facilities	Count of facilities2	% of Total reviews
negative	118	16.46%	10.78%
neutral	36	5.02%	3.29%
positive	563	78.52%	51.42%
Grand Total	717	100.00%	65.48%

Source: Own work.

Table 13: Sentiments towards staff indicated in numbers and percentages

Polarity	Count of staff	% of staff	% of Total reviews
negative	90	13.51%	8.22%
neutral	22	3.30%	2.01%
positive	554	83.18%	50.59%
Grand Total	666	100.00%	60.82%

Source: Own work.

Table 14: Sentiments towards room amenities indicated in numbers and percentages

Polarity	Count of room amenities	% of room amenities	% out of Total reviews
negative	100	15.20%	9.13%
neutral	31	4.71%	2.83%
positive	527	80.09%	48.13%
Grand Total	658	100.00%	60.09%

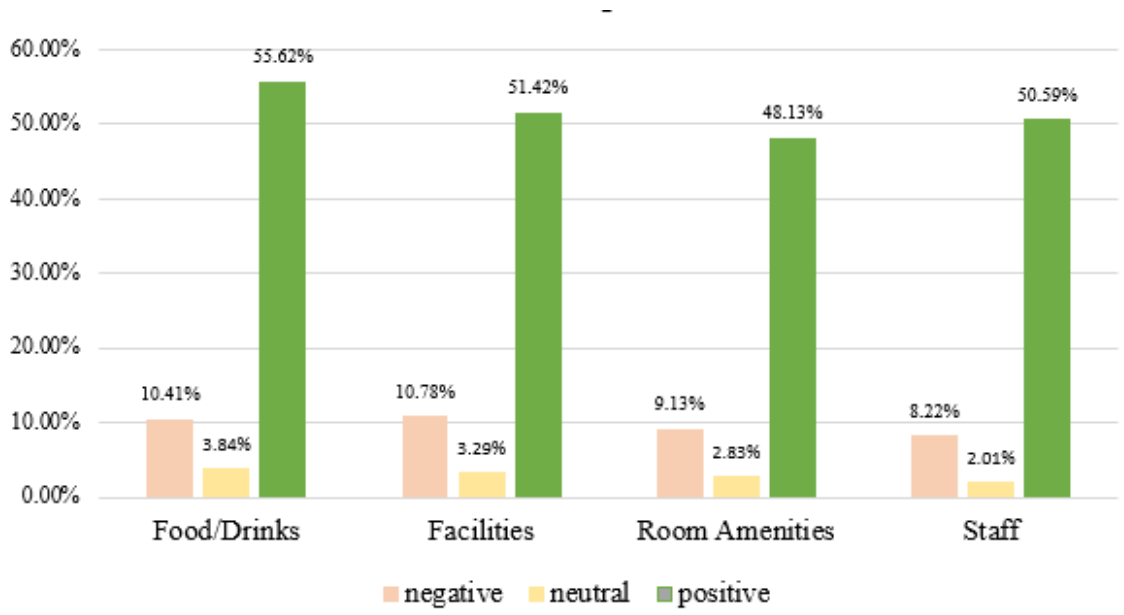
Source: Own work.

The percentage toward specific aspect, considering the whole dataset (1095 reviews), is provided, showing only data about main four indicated aspects. For easier understanding, these percentages can be found in the last column of tables 11,12,13, and 14 (indicated as: % out of Total reviews).

While analysing rates per site, the highest number of reviews is posted on Tripadvisor, containing 66.76% of reviews from the whole data set. The rating scale on Tripadvisor is defined from one to five. Rating with just one star was very rare and found in only 4 cases, presenting 0.55% from the mentioned website. Rate with just two stars was found in only 7 reviews, representing 0.96%. Then neutral rate, reviews rated with three stars, were found in 41 reviews out of 731, which means that 5.61% of reviews were rated as neutral. The reviews with mainly positive rates (four and five) are more likely to be seen on Tripadvisor. More specifically, reviews with a rate of four can be found in 230 reviews, which is 31.46% out of all reviews from Tripadvisor. Moreover, 449 reviews on Tripadvisor.com have the maximum rate, which represents 61.42% of reviews from Tripadvisor.com. From this, it can be concluded that most reviews posted on Tripadvisor have very high rates (four or five),

which means that the customers or guests who visited the hotel are in general satisfied with this hotel.

Figure 14: Polarities of most occurred aspects, compared to the whole dataset (in %)



Source: Own work.

A similar conclusion can be written when discussing the Google reviews rates, with few exclusions. Google is the site where the lowest number of reviews is posted, and reviews rated with rate two does not even exist. Only 102 reviews where sentiments towards specific aspects could be extracted and identified are extracted from Google.com, out of which 4 were rated with one, representing 3.92% of reviews from this site. Again, the highest number of reviews were rated as positive, 19 of them with 4 stars and 73 of them with 5 stars, which is 71.57% of all reviews from Google.com. Furthermore, as previously said, reviews where aspects were not available were mostly rated as positive. Therefore, reviews from Google are mostly positive, with just a few exceptions, where guests left negative reviews.

Last but not least, only 2 reviews from Booking.com were rated with the lowest rate, 5 reviews with rate 2, which means that in total, only 7 reviews were posted with a negative context. Then again, higher rates, meaning positive comments, were also present many more times than negative. To be precise, 96 reviews with a rate of four and 132 reviews with a rate of five, which means that 87.02% of reviews posted on Booking.com were positive. It can be concluded that people on Booking.com prefer posting very short reviews, which can be explained by numbers. Only 262 reviews out of 607 could be used for the aspect-based sentiment analysis. The rest of the reviews had only rates with a few words.

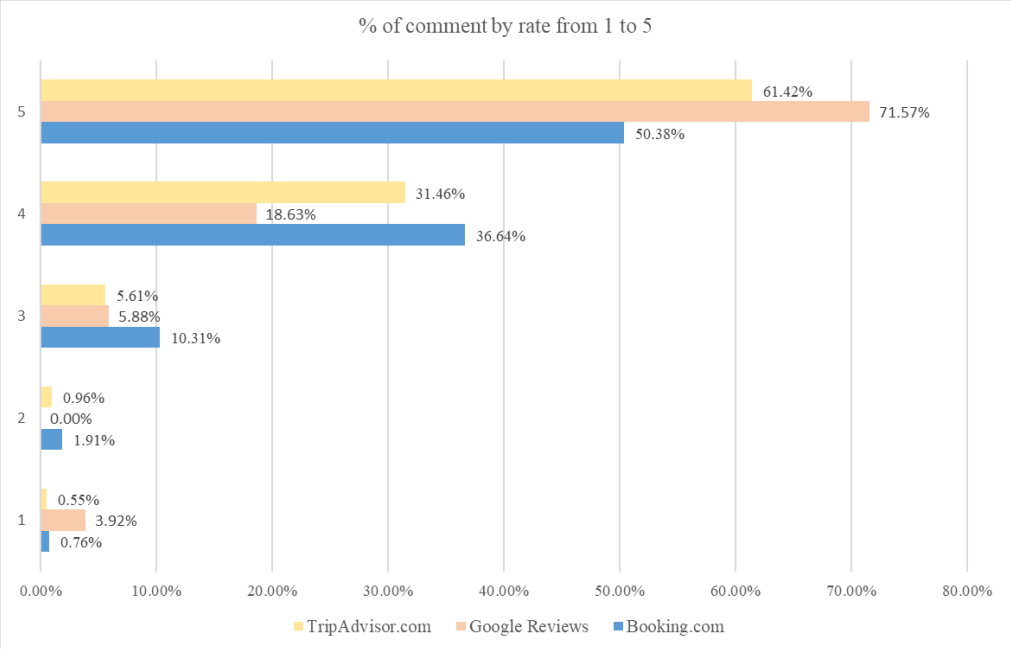
In Table 15, the detailed overview of the rates given on each site, expressed in percentages, can be found:

Table 15: Overview of given rates per each site expressed as a percentage

Rate	Booking.com	Google Reviews	Tripadvisor.com
1	0.76%	3.92%	0.55%
2	1.91%	0.00%	0.96%
3	10.31%	5.88%	5.61%
4	36.64%	18.63%	31.46%
5	50.38%	71.57%	61.42%

Source: Own work.

Figure 15: Percentage of comment by rate from one to five, depending on the site



Source: Own work.

In order to get the root cause of the reviews rated as negative with one or two stars, the aspects were taken. As 22 reviews were negative, it was inspected which of 15 aspects is mentioned in the vast majority of negative reviews. In 10 reviews out of 22 (only negative), the aspect defined as food and drinks was used in the negative context. Also in 10 reviews, the staff is mentioned in a negative way, in 8 reviews facilities is characterized as negative, whereas in 8 reviews, guests were not satisfied with room amenities. Thus, food/drinks and staff are mentioned in 45.55% of negative reviews, while facilities and room amenities are mentioned in 36.36%, considering only negative reviews. Reviews were read to make sure that the right sentiments were determined. Therefore, in Table 16 can be found a few examples of negative reviews from the

mentioned sites, but also the comment on the right site, stating if something is missing regarding the aspects or if any aspect is determined wrongly.

Table 16: Negative reviews and their aspects recognized by RapidMiner

Review	Aspect	Rate	Comment
The same every day; Habitually and not clean especially the old sauna area very unkempt and smells very strict	cleanliness: negative, location: negative	2	All aspects and polarities defined correctly.
The breakfast was good. The staff are rude - at the restaurant, the lady who serves food and snacks. Very cold, ugly attitude towards guests. For this amount of money for two nights we expected a lot more, with only one pool included in this price, where there is a soaring and poorly ventilated. It's all for extra pay. It's not worth that much at all - unfortunately we've had a bad experience.	facilities: negative, food/drinks: negative, value: negative, staff: negative	1	Aspect food/drinks is not defined correctly, it should be positive as the reviewer described breakfast as good. But the sentence is not very clear, and this made a confusion.
I booked a room overlooking the pool and paid for it. We were staying in a room with a view of nothing (roof).	facilities: negative, payment: negative, room amenities: negative, view: neutral	2	Aspect view is not defined correctly, it should be negative and not neutral.
Am not so enthusiastic - the rooms are so expensive; I will not come back; Nothing special	room amenities: negative, value: negative	2	All aspects and polarity defined correctly.
The staff was very friendly and helpful. Especially the lady who harvest the garden and herbs. Depending on the room rate with breakfast (245e), the price is too high, according to the offer. We had to pay extra to get into other pools. So I think we will not be here next time.	room amenities: positive, value: positive, food/drinks: positive, facilities: negative, payment: negative, staff: positive	2	Aspect payment and value are not defined correctly, it should be negative as the guest wrote that price is too high.
Not enough towels; rude staff.;	staff: negative	2	All aspects and polarities defined correctly
The boss of reception at hotel Sotelia has no human feelings!	staff: negative	1	All aspects and polarities defined correctly

table continues

Table 16: Negative reviews and their aspects recognized by RapidMiner (continued)

Review	Aspect	Rate	Comment
So far, a very known address in Slovenia. Unfortunately, the hotel was not worth a recommendation on our last stay. At dinner sometimes no free tables available, then after waiting - a dirty table. The noise level, no matter if breakfast, dinner or even partly at night in the corridors of a spa, which is actually not for recreational purposes, not worthy. Never again.	cleanliness: negative, facilities: negative, food/drinks: negative, value: negative	1	All aspects and polarity defined correctly, but aspect quietness is not recognized, which
For Terme Orhidelia we had to pay € 18.00 pp, when we booked there was no question of that. Only Terme Termalija was included, and this is catastrophic. Thermal water supposedly 36 degrees Celsius, but too cool. The staff is unfriendly, and the evening meal must be paid with 15.00 € pp despite half board. I do not recommend it - wasted money.	staff: negative, payment: negative, food/drinks: negative	1	All aspects and polarities defined correctly except payment. Payment is not mentioned and software recognized payment and not value as it should. Value should be negative.
Facility is nice. Food is tasty. Rooms are comfortable. But the people who works here they do not like foreigners. We felt so unwelcoming here. Specifically, lady at the cafeteria. Finn sauna was the best, but you never was able to step in because of those sessions. People were pack there like sardines in the can. How even by the safety law they allow to do this? Another big problem, I agree with other people, you have to be naked go get there. Such until sanitary, especially in Turkish sauna, where you cannot even bring a towel to sit on it. You can get some serious disease. This place is definitely not the best. They have so much to improve. Another annoying spot: they should have a towel station at the pool, where you leave or take new towels, not to drag all day alone, like in upscale resorts.	comfort: negative, food/drinks: negative, room amenities: negative	1	Aspects and polarities defined correctly, but some of them are missing, for example: Cleanliness should be negative and Staff should be negative as reviewer was not satisfied with these aspects. Aspect food/drinks should be positive as reviewer stated Food is tasty.

table continues

Table 16: Negative reviews and their aspects recognized by RapidMiner (continued)

Review	Aspect	Rate	Comment
Hotel rooms very nice (4 stars)! Food, sorry a single disaster! I'm not a gourmet type of person but what is offered in this hotel (buffet) is under all ...! If I were a director in this hotel, I would replace the kitchen staff immediately! Without thinking! Unloving service, absolutely tasteless, completely cooked food! Breakfast. Well (unfortunately no yogurt?) Hotel gladly again, but because of the food not more!	food/drinks: negative, room amenities: positive	1	All aspects and polarities defined correctly.
We are right here in the Sotelia Hotel and unfortunately have to rate it negatively! The system is great, but more than one star loses it to the following: The hotels sold all rooms to the guests with increased prices to the guests (including us), but it was not indicated that guests without a room reservation also came to the thermal baths. So it was packed and no staff! It was also no longer beach chairs for hotel guests. In addition, there was only one (restaurant) where you could eat something in the thermal bath, unfortunately the queue was something from Lang as we had to wait 1:55 hours on the first day. The staff is very inexperienced, and complaints are not noticed! We were in a bar that closes at 2:00 p.m. and had no ice cream for our son even though we were there at 1:57 p.m. and it meant that we couldn't get anything anymore. The bar staff simply ignored us when it said that it didn't have 3 minutes to close. Nice area but unfortunately no service! Pity!	room amenities: negative, value: negative, facilities: negative, location: negative, food/drinks: negative, staff: negative, customer support: negative	1	All aspects and polarities defined correctly.

Source: Own work.

From Table 16, it can be seen that in most of the cases, the software recognized the proper aspect, sentiments, and polarity for aspects. However, there are some errors in recognizing, such as determination between value and payment, and in some cases, the aspect is not recognized at all, for example, cleanliness or quietness. Nevertheless, if all negative reviews are considered, out of 22 reviews and 63 mentioned aspects, only 9 aspects were wrongly reported or not reported.

As there are fifteen different aspects and five different rates, it was also meaningful to inspect the association between different aspects and different rates to see if there is a relationship

between them. In statistics, a measure used to quantify a relationship between aspects and rate (in this case), or any other two or more variables is called the measure of association. Association can be determined by many analyses, including regression and correlation analysis, contingency table, and others. Which method will be used to determine the association depends on data characteristics. Data used for this analysis were polarity of aspect (positive/negative) and rates of review (positive/negative, therefore reviews rated with 1,2,4 or 5 stars). These data are dichotomous (two qualitative variables), so the contingency table was conducted. The entries presented in the tables are frequencies/counts showing how many times the inspected aspect was found. Keeping in mind that there are fifteen different aspects and that it is necessary to inspect the relationship between each aspect and rate, a contingency table is created for each of them separately. Since each aspect is inspected separately, the table involves only two variables, and therefore they are classified as two-dimensional contingency tables. In Figure 16, a two-dimensional contingency table if both variables are dichotomous is presented.

Figure 16: Two-dimensional contingency table

		Variable A		
		Category 1	Category 2	Total
Variable B	Category 1	A	B	A+B
	Category 2	C	D	C+D
	Total	A+C	B+D	N = A+B+C+D

Source: Raič (2019).

For calculating, the coefficient related to Figure 16, the calculation for Cramer's coefficient (V) is shown in Equation 1.

Equation 1: Calculation for Cramer's coefficient (V)

$$V = \frac{|AD-BC|}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \quad (1)$$

More information is given by the predicted values, indicated in the Equation 2:

Equation 2: Calculation for Cramer's coefficient (V), predicted values

$$V = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \quad (2)$$

– If $\phi > 0$, if the first variable takes the first value, the second variable tends to take the first value; on the contrary, if the first variable takes the second value, the second variable will also take the second value. In this case, the variables are connected and associated.

– If $\phi < 0$, if the first variable takes the first value, another variable will take the second value; on the contrary, if the first variable takes the second value, the second variable will take the first value. Therefore, in this case there is no association between two variables.

The scale with qualitative definition of association according to which the results are interpreted is rather subjected, thus the following range is defined:
– From 0.0 to 0. 2: insignificant/loose association
– From 0.2 to 0.4: slight association
– From 0.4 to 0.6: moderate association
– From 0.6 to 0.8: strong association
– From 0.8 to 1: very strong association

Following this logic, the association is inspected for each aspect and positive/negative reviews. Therefore, variable B was aspect polarity, category one was positive, and category two was negative. Moreover, variable A was rate, and here variable one was positive rates (4 or 5 stars), and variable 2 were negative rates (1 or 2 stars). In Table 17, the template for calculating the association between stated aspects is shown:

Table 17: Calculating aspects association

Aspect/Reviews	Positive rate (4 or 5)	Negative rate (1 or 2)	TOTAL
Positive	A	C	A+C
Negative	B	D	B+D
TOTAL	A+B	C+D	**SUM**

Source: Own work.

Having done all the association calculations for different aspects and positive and negative rates, the results and associations were ready. In table 18, the association coefficients can be found. Table 18 is created to overview the association coefficients between each aspect and the review rate. For example, if one of the aspects is mentioned as positive, inspect if there is a chance that the whole comment would be positive and vice-versa. These coefficients were calculated in the same way as indicated in Equation 2.

From Table 18, it can be concluded that there is a strong association between beds and reviews, whereas the association is moderate between customer support and reviews. Furthermore, there is a strong association between cleanliness and reviews, meaning that if the review is rated as positive, there is an extremely high chance that aspect cleanliness is mentioned positively and vice-versa. Moreover, there is a slight (very close to moderate) association between food/drinks and review, between staff and review and between room

amenities and review. However, for the aspect of Wi-Fi, it was not possible to calculate the association with review, as Wi-Fi is not mentioned in any of the reviews rated with 1 or 2. Therefore, as the number cannot be divided by 0, it was impossible to obtain the result.

Among the 15 aspects identified from the text mining exercise, the study finds that the following aspects have a significant probability of increasing the ratings if in the review they are mentioned in the positive way, meaning they are strongly connected and associated:

- Cleanliness/ Review = 0.877
- Beds/Review = 0.674

Table 18: Cramer's coefficient

Aspects	Cramer's coefficient
Facilities/Review	0.293
Food&drinks/Review	0.315
Staff/Review	0.369
Room amenities/Review	0.334
Value/Review	0.298
Beds/Review	0.674
Cleanliness/Review	0.877
Comfort/Review	0.239
Customer support/Review	0.545
View/Review	0.298
Design/Review	0.268
Location/Review	0.278
Payment/Review	0.235
Quietness/Review	0.283
Wi-Fi/Review	#DIV/0!

Source: Own work.

5 DISCUSSION

After all theoretical and practical parts of the thesis were done and explained in detail, it was time to compare the best practices with practices used in hotel Sotelia. In chapter 5, the discussion about the analysis and results was provided together with the interview interpretation and answers to the research questions. As mentioned in chapter 3, the interview helped understand if the provided analyses were valuable for the hotel and its management and helped reveal what is currently the main challenge for hotel Sotelia. Moreover, the interview was helpful to check whether the software recognized the right aspects as the main ones, thus if provided improvements are meaningful.

5.1 Interview and interpretation of results

As data can sometimes be invalid and staged, having in mind that it is well known that guests can intentionally write a negative reviews or extremely positive reviews, it was meaningful to have the interview with the sales representative of the hotel. The aim was to compare the results from the analysis with the hotel's current operations and connect the theoretical part provided in the first two chapters with the practical part provided in chapter 4. Then, on top, propose suggestions and make conclusions based on the analysis. Before the interview took place, little research was done, where the number of comments on other Slovenian hotels was taken into consideration. Hotel Sotelia, with circa one thousand reviews, is expectedly among the hotels with the highest number of posted reviews in Slovenia. Still, this number of reviews is not very high compared to other hotels within the same category (4*) worldwide. There is no specific amount of data per dataset defined to conduct text mining, but it is always better to have a bigger dataset. The reason for this is that the quantity of available information is dictated by the size of the sample, which in turn, defines the precision or degree of confidence in any sample estimations. Of course, there is no need to implement text mining techniques if the dataset consists of few rows and columns. However, text mining is advantageous when somebody is faced with a few hundred or a few thousand data, which need to be processed and inspected to gain knowledge from the dataset. Although using text mining has more opportunities than limitations, it is still a new technique, and it is essential for a person with experience in conducting the analysis. Surprisingly, not many hotels in Slovenia use this technique. Many of them still believe more in the old-school paper documentation and surveys. As the number of posted reviews for Slovenian hotels is not so significant, hotels and their management usually do not spend money on hiring analysts to explore the reviews. Instead, they spent more time reading reviews one by one. There are many reasons for such behaviours, but the general reason for not using it in Slovenian hospitality is a lack of qualified labour and, consequently lack of knowledge and experience for using text mining for conducting similar types of analyses, as presented in chapter 4.

The manager of hotel Sotelia expressed that finding a way to engage guests into commenting, so the sample could be big enough in terms of size and good enough in terms of quality is currently the biggest challenge for the whole Sotelia team. This engagement is indicated as the important one from the management perspective because it would give potential guests a sense of competency while booking their accommodation. Even though hotel Sotelia has one thousand online reviews, it cannot be forgotten that many people just write one or two words, which is not valuable to the management as the review with the longer description are. However, it is hard to find a way to motivate guests to write such comments, and on the other hand, management cannot force their guests. Therefore, the burning issue in hotel Sotelia is how to come to a situation where management will have better online reviews in terms of quality, which will serve to conduct more complex analyses and find the hidden patterns behind the reviews. In hotel Sotelia, the employees analyse reviews manually,

without any text mining technique. The reason for not using text mining tools, the manager mentioned both lack of skilled labour and lack of knowledge, confirming the statement from the beginning of chapter 5.1 regarding the labour-market situation in Slovenia. In Sotelia, every Tuesday, all management board members go through all guest reviews from the previous week, and once in a month, the hotel checks the overall ratings on the travel-related websites. After that, management selects a responsible person and sets a deadline to eliminate the shortcoming that led to the poor assessments. These shortcomings are then reviewed, and the best corrective and preventive measures are built based on the review that unsatisfied guests posted after he or she left the hotel. The hotel does not consider only online guest reviews from Tripadvisor.com, Booking.com, and Google.com) as the important ones but also responses on the receptions, at the bars, in restaurants, or reviews received in the hotel's online surveys. For hotel Sotelia, the online survey is a convenient way of getting information about the customers' preferences. It gives the hotel an accurate picture and guests' opinion, which allows the hotel to react or even fix the impression before their guests leave.

In general, managers of hotel Sotelia think that they do not face any significant problems with the analysis itself, except the time that must be spent on the manual work. He added that, even though the employees are trying to provide the answers to all online and offline reviews, unfortunately, there is a lack of labour and time. However, it is mentioned and proven by analysis that hotel Sotelia receives much more positive feedback than negative. Sometimes it is hard to spot a room for improvement because guests usually share good impressions. This can also be seen by checking reviews on the websites, where the vast majority are rated with a very high rate, followed by a short positive description. As evidence, the hotel is currently rated 4.6 out of 5 on Google.com. On Tripadvisor.com data showed that the hotel is rated 4.5 out of 5, whereas on Booking.com hotel's rate is 8.7 out of 10.

When it comes to the most important aspects for hotel Sotelia, it was discussed that the hotel does not focus on building a reputation based on one specific aspect. Oppositely, the management of the hotel always strives to ensure that all aspects are equally good and that guests are satisfied with everything the hotel has to offer. Nonetheless, the representative stated that based on the type of hotel's business named thermal resort, cleanliness and aspects related to the well-being of guests, such as room amenities or food and drinks, may bring more attention in the eyes of the customers than example technical equipment (wi-fi or conference rooms), or location. This was also proven in chapter 4, where the percentages regarding the four most commented aspects were provided in the tables numbered as 9,10, 11 and 12. Out of four aspects, the aspect that was mentioned in most of the reviews was food/drinks. This aspect was mentioned in 756 reviews out of 1095, which means that in 69.86% of reviews, it was found that people were talking about food and drinks in positive, neutral, and negative ways.

People mentioned positive things related to food and drinks in 609 reviews, in 114 reviews it was mentioned in the negative context, while in the rest of the cases, it was neutral. Keeping in mind that aspect food/drinks was mentioned in 55.62% as a positive, in can be concluded that the percentage of positively characterised comments related to food is high, which means that the hotel management can be satisfied with how the restaurant currently operates. It was also an indication that the hotel's business is going in the wanted direction. The food is characterised as a crucial factor on the official website, which greatly contributes to the guests' satisfaction. Nonetheless, as a significant factor for hotel, management should focus even more on making the food more appropriate for all types of populations in terms of guests' eating habits and preferences.

Guests mentioned words related to hotel facilities in 717 reviews, in 527 reviews facilities is characterised as positive, and in 118 reviews as negative, while the rest is characterised as neutral. Taking the same data for staff, out of 666 reviews where staff is mentioned, 554 reviews are commented in a positive way, 90 with a negative connotation and the rest is commented in a neutral way. On the fourth place are room amenities, which are in total mentioned 685 times, 527 times in the positive way, 100 times in the negative way and the rest is stated as neutral. The manager's statements confirmed the percentages and numbers related to the most commonly mentioned aspects. He stated that guests visiting hotel Sotelia are eager to pay attention to the things that are highly connected to the hotel's mission – to relax and forget about time in the oasis with thermal water and a beautiful food assortment. If these conditions cannot be met, some guests will rate their stay at the hotel, or at least a specific aspect, as negative. The manager then connected these results with the fact that some things management can change and improve, but he mentioned that some aspects are unchangeable, such as location. As an additional explanation, he added that hotel Sotelia, a highly categorised hotel, received numerous architectural awards, which confirms the architectural originality, the dynamic and positioning in the natural environment. Therefore, amenities such as location, wi-fi, and bed comfort may be self-evident for Sotelia's guests. Meaning that guests may imply the quality of the things, and therefore, they may not be focusing on mentioned aspects while reviewing their stay.

He added that it is much harder to ensure good cleanliness and well-being of guests in general than a good wi-fi connection during the pandemic. The manager added that it is not surprising that these four aspects recognised as the most important ones for guests could also be found in the reviews rated with rates 1 and 2, as these are the aspects with the highest number of mentions. He continued by saying that employees are aware of all negative reviews, but management can only work on their prevention by listening to guests but cannot impact someone's subjective opinion. Nevertheless, these kinds of negative reviews represent only a small percentage, only 2% (22 out of 1095) out of the whole analysed dataset, but still, this percentage should not be taken for granted. Moreover, it should not be forgotten that the software determined all of these aspects and translated reviews from other languages into English. This means that mistakes are possible as software cannot precisely

determine the double negation or sentences containing exaggerated words, sarcasm, or any types of the valence shifters. Therefore, these negative reviews had to be also inspected manually, as done in Table 16.

Considering only negative reviews from Table 16, it can be concluded that in 85.72% of reviews, the software recognised the right aspects, which is a high percentage for this kind of analysis. In 7.93% of cases, aspects were not defined correctly, and in 6.34%, aspects were not recognised. Although negative reviews are not very welcome, they can be useful to see what management should improve. The manager confirmed this, as he said that even though it hurts to receive a negative review, sometimes it can be more helpful than the positive one. In that way, management directly receives the signal that the guest was not satisfied during his/her stay. Then, the hotel can make concrete changes, which would be beneficial to future guests and make their stay even better. Of course, sometimes the competition, which is situated in the same region, can write a negative review or pursue the unknown person to post rude text to tarnish the hotel's reputation, and in that sense, shift the focus to their hotel.

Fortunately, hotel Sotelia did not have many fake reviews in the past, but management is facing exaggerated reviews written by guests which were not satisfied with Sotelia's services. The hotel faced less than ten reviews in the past that were obviously fake. However, as everything connected to the internet and technology is improving daily, these websites are often updating the policy. They are checking if guests had visited the hotel, tracking the level of rude words and the truthfulness of the whole review. In two out of ten fake reviews that Sotelia faced in the past, it was expressed a bad intention to lower Sotelia's reputation. This situation was caused because guests did not get free services after the complaint. In cases that employees manage to recognise this kind of intention, they inform Tripadvisor and show them evidence, including emails, in which these guests threatened the hotel. Nevertheless, Tripadvisor removed the negative reviews in both mentioned cases, and in recent years, hotel Sotelia did not face any negative intentions.

After the questions regarding the main challenges, dealing with fake reviews and main aspects connected to the mission were answered, the attention was paid to the connection between the aspects. The analysis that showed how important the connection between the aspect and review rate is, was provided at the end of chapter 4 (table 16). Results obtained via the contingency table indicated that if the cleanliness is reviewed as positive, it is most likely that the whole rating would be positive, meaning that the association between the cleanliness and review rate was strongest out of all associations between other aspects and rates. In the interview, the representative confirmed this by stating that the hotel's general business orientation and the nature of its services must always involve high cleanliness standards, especially in the pandemic period. He continues that this kind of result can also be expected in the future, as the Covid – 19 introduced tremendous changes in people's behaviour. In the last two years, people have been more sensitive to cleanliness and health in general, and hotels should do their best to adapt to new circumstances. Again, as a high-

category hotel, it is not strange that the contingency table showed these results, considering that all these aspects are strongly connected to the hotel's mission. Therefore, the manager just confirmed that the stated combination seemed rational and true according to his knowledge, experience, and logic. He said that he did not have the opportunity to see this kind of analysis conducted in hotel Sotelia in the past. However, he concluded that this analysis could be incredibly beneficial, especially when the hotel has many reviews. The manager added that he would be happy to re-check it in the post-covid times and environment.

At the end of the interview, as the reason for having the highest number of quality reviews on Tripadvisor, the interviewee stated that hotel Sotelia does not force their guests to evaluate the hotel on the online sites. Nonetheless, the only thing that boosted the number of Tripadvisor reviews is that in 2017 online booking provider called Phobos made an API, and hotel Sotelia is now directly connected to Tripadvisor for making reservations. Thus, from 2017 on, guests had an option to make reservations directly on Tripadvisor, which is the trigger for receiving many more reviews on Tripadvisor, than on other online platform. However, in the period from 2020 to 2022, a decrease of reviews posted on all travel-related websites was spotted because of the travel restrictions introduced during the Covid-19 pandemic. As mentioned in chapter 1.2, the covid-19 stopped the growth of the hospitality industry, and consequently, the number of reviews on the websites decreased. However, according to the manager's words, the number of reservations started to increase again, which will lead to more profit and more reviews on websites in the future.

From the interview, it was concluded that initiatives and suggestions from guests are vital to the management of hotel Sotelia, as the offers are defined based on guest preferences and reviews. The manager mentioned that the hotel generally works on continuous improvement related to its services to offer the best possible experience to its guests. To emphasise the importance of online shared reviews, the manager said that these reviews could also show where exactly the hotel stands in the market in terms of position, meaning that reviews could be management's way to compare the hotel's success with the success of similar resorts. He continued that in that way, management can also see which improvements are needed and in which direction they should develop new offers, to be better than competitors. Consequently, that could lead to a better position on the market, higher prices, and a better reputation among current and future guests. From the results obtained from the analysis, it was concluded that hotel Sotelia is an excellent hotel, in which guests' satisfaction is in the first place. Moreover, from the main aspects occurrence and the other provided results and analysis, it can be concluded that hotel Sotelia operates according to its mission, where the employees create the well-being and unforgettable experiences for guests who can indulge in endless thermal culinary pleasures.

Nevertheless, the interviewee also underlined that after seeing these analyses from the thesis, exploring what customers think about the main facilities and functionalities of the hotel seems to be much easier with text mining. The manager then stated that using text mining is

very good and welcome for time saving and a better overview of guest satisfaction. Unfortunately, until now, management did not have the opportunity and knowledge to implement it into their activities. However, after this research, management is more aware of the benefits and opportunities of using text mining techniques and will strive to implement them into regular business activities in the upcoming months.

5.2 Opportunities of using text mining in hospitality industry

After interview is done and after all the analyses and results are interpreted, the answers on the research questions can be provided. Therefore, in chapter 5.2, the benefits of using text mining will be presented.

RQ1: On which analysis can hotel managers rely on while using text mining techniques for analysing reviews in the hospitality industry? In other words, what are the opportunities of using text mining in hospitality industry?

To briefly summarise the answer to the first research question, the opportunities of using text mining are tremendous since the hospitality sector is a customer-focused business that collects many data via different systems. These systems can be property management systems (PMS), central reservation systems (CRS), guest loyalty program databases, point-of-sale (POS) and others. As a result, applying text mining techniques may help managers improve guest experiences, design marketing plans, increase retention and loyalty, and maximise profit. Moreover, text mining enables hotels to filter through massive data sets in search of meaningful relationships, allowing them to anticipate rather than react to client needs. Likewise, it was stated that among the most significant opportunity of using text mining is the possibility of analysing specific aspects, saving time by fast data extraction, and translating text quickly. With the stated benefits, it is easier to understand if the guests are satisfied and spot changes in their behaviours and inspect their willingness to return to the hotel. Furthermore, there is no limit in terms of the dataset size, which gives the same opportunities to small and big hotels. Managers have many different analyses available for analysing different business segments. Each of the analyses stated in the previous pages can bring specific knowledge, depending on the aims and goals of the research. Nevertheless, it is essential to understand that failure or success of the analysis is frequently determined not just by the ability to collect data but also by translating and transforming that data into information that will assist in better managing the hotel. In other words, it can be stated that managers can rely on text mining techniques if they get to have a good analyst and if they have enough knowledge to understand and read analysed results. The crucial step that hotels or managers must ensure prior to implementing any text mining tools and techniques into regular business is to find a provider who is:

- Experienced in extracting the knowledge from the information
- Experienced in creating specific predictive models

- Able to collect the correct data, select the adequate analysis and tools
- Eager to constantly refine the process to make it more transparent and improve it, making the analysis better each time.

Among the most used analysis in the hospitality sector are related to deviation detection, forecasting, clustering, classification, association rules and sentiments. Without text mining, marketing insights about the guest's characteristics and buying habits may stay entirely unexplored. Research showed immense potential in this field of study, and it can be expected that these benefits will only grow in the future. From the interview, it can be concluded that the manager recognized and confirmed the importance of all the analyses while focusing on the aspect-based sentiment analysis and the contingency table. As mentioned in chapter 5.1, he confirmed that these kinds of analyses are beneficial for the business unit, not only for a deeper understanding of the customers but also for an easier and faster understanding of the main drivers for customer satisfaction. On the other hand, he added that analyses provided an excellent overview of what needs to be changed to achieve even better results in the future. At the same time, the contingency table seemed to be helpful for the managers to inspect the association between different aspects of the hotel's business. It was also confirmed that if data gathering and data cleansing are done correctly, managers can rely on mentioned analyses in the hospitality industry and use the results for further research.

5.3 Limitations of using text mining

Up to now, the focus has been on the positive sides of using text mining. However, even though data mining and text mining can bring exceptional advantages to the organization, some crucial things need to be considered once leaders are ready to use text mining solutions in their regular daily activities. Therefore, in chapter 5.3, research question two related to the limitations of using text mining is answered.

RQ2: Are there any pitfalls in analysing reviews while using text mining tools and techniques? If so, what actions should be undertaken to resolve them?

In order to provide the answer to the second research question, the list of five main limitations is provided, stating the possible ways to solve these issues. The most common limitations of using text mining are:

- Lack of knowledge is among most common reasons for not using text-mining solutions. An essential step is finding the right workforce. A qualified person will ensure that all the steps, from web scraping, data cleansing, and analysis to the final interpretation of the results, are done correctly, without significant mistakes. This is important because if the analysis or any mentioned step is done incorrectly, the whole interpretation and results would be wrong, leading to mistakes in the decision-making process. In many cases, it is stated that not everyone can conduct these analyses. Data mining is a relatively new field in science, so there is a lack of workers. Consequently, companies should

invest even more money and time in hiring good analysts to have a suitable overview of the hotel's current market position and learn about potential improvements (RevolveAI, 2020). This was also confirmed by the manager of hotel Sotelia, as he mentioned that this limitation is the main reason why Sotelia did not implement text mining into regular activities.

- Language and translation can also represent the barrier of using text mining. For example, if the text that needs to be analysed is in Italian or German, there is a bigger chance that the programming language will not be able to provide a 100% accurate translation. Sometimes, the main point of the whole review may be missed because of the ambiguity, which leads to erroneous results. This happens mainly because many tools take English as the default language (Lee & Yang, 2003b). Nevertheless, as programming languages are upgrading daily and every minute new information about solving problems is available, it will not be strange that one-day society will face the situation where the programs will provide 100% correct translation. However, the translated text should be taken with particular attention and should carefully be checked prior to conducting further analysis. Therefore, the final results should be taken with a certain level of flexibility. This was also proven in the paper, as it had to be checked whether the comments were translated correctly so that the software could recognise the right aspects. Before conducting the analysis and receiving the aspects, it was inevitable to check and change the translation of some reviews, as the original translation was not meaningful. The software would not recognise the aspect correctly if the translation had not been changed. Even though this change was done, some aspects were still not correctly recognised because of the ambiguity, or any other reason related to the insufficiently clear review. Nonetheless, as it was concluded from Table 14, translation and aspects were correctly recognised in most cases. Although it is much faster to check only translated text than to go manually through all reviews, this can increase the time needed for the whole analysis.
- Sensitivity, ambiguity and understanding of the specific words, phrases, and contexts is another limitation which should not be forgotten. The programs and software cannot understand these kinds of text the same way as humans (Shaidah & Alfawareh, 2012). This means that if the comment, review, or any analysed text was written with an abbreviation, slang words, or maybe double negation, the program might not understand and might not interpret the results correctly. Of course, there are many options to minimise this misinformation, such as writing an algorithm for detecting and identifying the abbreviations, slang and more. However, it must be stated that this can only be done if the person has enough programming knowledge. If not, having not detected slang, abbreviations, double negation, diminutive and augmentative words can lead to totally different meanings and, for example, wrongly detected aspects in the aspect-based sentiment analysis. For example, two sentences were investigated: "Today will be a super ugly day" or "This bag is terribly beautiful". In the first sentence, two main words can be spotted: super and ugly. The program will most likely rate super ugly as positive if conducting the aspect-based sentiment analysis because the first word seen was super.

The same goes for the second example. Having the word "terribly" prior to beautiful makes the program think that the connotation is negative, and therefore this would probably be rated as negative. Because of the stated, while doing data cleansing, special attention should be paid to these kinds of scenarios where misleading words may occur.

- Privacy and legislation problems are sometimes stated as the two main limitations of text mining. Customer data is used in marketing tactics to reach the correct audience at the right moment. Personal information about users is a gold mine for marketers looking to sell their goods and services. Private data is utilised to analyse client behaviour better and identify similar buyer profiles. Because of the large datasets are available, many people get user-specific offers on the social media or other communication channels (MBA Knowledge Base, 2018). The especially sensitive information is connected to the medical sector, where it is imperative to secure the patient's privacy and if the private data leak, all the sides are endangered. When it comes to the travel-related website, the sites are already made so that the user needs to agree to terms and conditions. Depending on the things written in the terms and conditions, they can permit hotels to access posted data without legal consequences. The privacy and legal difficulties that may arise from data mining are the fundamental drivers of the developing conflicts. Many applications of data mining are causing privacy concerns. Every year, the government and businesses collect massive data about consumers and store it in online data warehouses. Part of the concern is who will access the data once it has been collected and kept in a data warehouse. Customers may be unaware that the information obtained on them is shared with more than just the person or site who collected it.
  With today's technology, data mining may be used to gather data from data warehouses, identify various information and relationships about consumers, and make connections based on this extraction, thereby jeopardising the privacy and information of customers. Data aggregation, which involves gathering data from many sources and combining it with being analysed further, is an example which shows that data mining needs data configurations that can encompass customer information, which may jeopardise privacy and secrecy.
- Data mining, particularly collecting data on people, has significant ethical issues (Bhan & Perpetua, 2020). When selecting whether to notify a person that his or her information is being preserved for future data mining, companies confront an ethical problem. Data mining and text mining may discriminate against people based on their ethnic, religious, and sexual orientation. If data mining is used in this way, this is considered unethical and illegal. Any illegal or unethical usage of people’s personal information needs to be avoided, and before any decision is made, people should be aware of how their data can be used, for which purpose and what part of the information they are sharing will going to be taken. They should also be aware if this exposure will have any consequences and, if so, are these long-term or short-term consequences. To ensure that individuals are secure, data must be used and examined appropriately.

5.4 Detecting hidden patterns with text mining

The previous few pages provided conclusions regarding the opportunities and barriers of using text mining. Finally, having the first two research questions answered, the answer to the third research question could be given.

RQ3: Are text mining techniques the right ones for detecting hidden patterns in customers' sentiments and can that action lead to business improvement and competitive advantage?

Parts of the answer can already be found in chapter 2, in the discussion, and in the answer provided for the first research question, where it was written what the advantages of using text mining are. Having the analysis done and having all the advantages already written, it is now the moment to unite these insights and conclude if the text mining techniques are helpful in detecting the hidden patterns. As seen in the analysis part described in chapter 4 and as proven in the interview, text mining techniques used in the thesis helped reveal aspects and their sentiments, which helped provide basic statistics and insights for further conducting the research. Not only for this kind of research but also for many more, uncovering the hidden patterns can lead to important innovations, inventions, and general business improvements. For example, with text mining, the pharmaceutical industry can reveal hidden knowledge and hasten the pace on the discovery of the new drug. In other product-oriented companies, real-time analysis of customer reviews can be conducted to discover product gaps and take immediate attention if needed. Moreover, some other applications of text mining are:

- Fraud detection: With text mining, it is possible to analyse large volumes of textual data and detect fraudulent transactions. These fraudulent claims can be uncovered by checking for frequently used keywords in descriptions of accidents.
- Customer service: By evaluating their content, text mining can automate the ticket tagging process and automatically route tickets to relevant geographic regions. It can also assist businesses in determining the urgency of a ticket and prioritising the most important ones.
- Business intelligence: Text mining makes it easier for analysts to review enormous amounts of data and quickly uncover useful information in business intelligence. Manual analysis is impractical if data are acquired from multiple sources. Text mining software speeds up the process and allows researchers to extract valuable data.
- Healthcare: Text mining is becoming increasingly useful in the healthcare profession, particularly for clustering data. Manual investigation takes a long time and is expensive. Text mining can be used in medical research to automate extracting crucial information from medical literature and in many more cases (Chang, et al., 2018).

  It can be concluded that in almost all industries and businesses, text mining can bring substantial knowledge, unhide hidden patterns and, with the proper research and analysis, lead to complete business improvement. Notably, in the case of hotel Sotelia,

text mining helped reveal the main aspects on which the customers are focused, their sentiments, word frequency, connection with other aspects, and their relations, which helped the hotel better understand their customers. It helped provide the essential aspects needed for further and deeper analysis. It also simultaneously showed that it is constructive to use text mining techniques to have real-time and most accurate information, especially if the dataset is extensive. According to the manager, the research showed a lot of room for improvement for micro-changes, which can lead to more satisfied guests. These micro-changes, such as placing more baskets for towels in the spa centre or making a wider food assortment for vegetarians, are not expensive but are very important for guests. Therefore, the manager confirmed and agreed that hidden patterns in customers' behaviour, which are connected to the aspect sentiments, were provided within this research. That can lead to the final conclusion that text mining can be used for detecting hidden patterns in customer's sentiments, which can lead to business improvement and competitive advantage.

CONCLUSION

Everything that has an impact on the business and can be evaluated is considered unstructured data (Tanwar, Duggal, & Khatri, 2015). Giving the form and structure to large, unstructured datasets is the first step toward obtaining meaningful insights. This entails translating the dataset from originally stored data (video, audio, picture, or text) into text (Antons, Grünwald, Cichy, & Torsten Oliver, 2020). It was learned that text mining is a natural language processing approach that provides generous help to analysts in gaining valuable insights from unstructured data. These techniques and methods are used to better analyse patterns of client's behaviour, uncover hidden data, and forecast new trends (Lee & Siau, 2001; Hoontrakul & Sahadev, 2008). Exactly one of these methods was used in this thesis. Case of hotel Sotelia had a purpose of finding the practical analysis that can be beneficial to hotels in general and was also used to uncover the benefits the limitations that can arise while using text mining tools and techniques. All reviews from the online travel-related website (Booking.com, Tripadvisor.com and Google.com) connected to hotel Sotelia were used for this work, dating from January 2009 to January 2020.

Together, the literature overview and case study contributed to the broader understanding of the text mining implementation, opportunities, and barriers. Moreover, the aspect-based sentiment analysis, contingency table and other accompanying analyses were conducted and inspected with the representative, which then confirmed the validity of the results. As one of the most important benefits of using text mining, it was mentioned the possibility to collect, analyse and visualise data very fast, which gives more time to focus on other activities related to business. Additionally, having the analysed and accurate results increases the possibility of competitive advantage over the other businesses and gives the hotel a direct sign of what is good in their business and what must be changed to make their guests even more satisfied and loyal. Furthermore, by using different analyses, hotels can inspect all the

aspects of their businesses, the relationship between two or more aspects, gain a complete overview of the online reviews, compare themselves and their competitors, try different analyses depending on a business niche such as aspect-based sentiment analysis, clustering, decision tree, categorisation, topic detection, summarisation and many more. As the possibilities are increasing with the rapid growth and improvements in the IT industry, it is also now possible to automatically translate text from foreign languages into one single language and, in that way, have one language data set (for example, all data in English), which allows easier analysing and bigger dataset. This language technique was also used in this thesis. Having reviews available in different languages, it was decided to translate them into English. It has also been argued throughout this work that, even though there are many positive and beneficial things if implementing text mining and modern business intelligence solution, there are also some limitations that must not be overseen. When it comes to the limitations, there are five main limitations associated with text mining, which were enumerated in the answers to research question two, described in chapter 5.2. In general, the most common limitation of text mining implementation is a lack of knowledge. It is very important to have a good analyst, otherwise, analysis can be done wrongly, which could lead to the wrong understanding of the market and business situation. To have a good analyst with specific knowledge on board, a company must invest a lot of resources and energy. This is also among the reasons why hotel Sotelia did not take into consideration the implementation of modern data science tools and techniques in the past.

The way to overcome this situation is to convince the business unit and board of directors that in the long-term, text mining will bring a lot of the advantages and new possibilities to business, and hiring a new, qualified workforce to conduct the analysis can only be beneficial. Another limitation is language translation. While using a translator, the sense of the sentence posted in the original language can be lost. Moreover, as stated during this research, these tools still cannot spot and correctly understand slang words, abbreviations, specific phrases, ambiguity, and other language barriers, which can lead to the wrong translation and consequently to wrong results. As explained and described in chapter 4, the way to avoid these situations would be very good preparation of data before conducting the analysis, detailed data cleansing and double-checking before the analysis is run. The other two mentioned limitations are widespread in the other industries, but it must be stated that in the hospitality industry, and especially on official travel-related websites, each user must agree to the terms and conditions of the particular site if he or she wants to post a review or even just browse the site. Meaning that, almost by default, all the reviews are already visible. Thus, in terms of privacy, users voluntarily allow a website where their data can be used. To connect the stated limitations with the work in the theses, at the very beginning, it was demonstrated that information technology is advancing at a rapid pace, and in the coming years, it can be expected that information technologies will contribute to the success of other industries even more, which will, in turn, contribute to the growth of the entire IT industry. Likewise, engagement on the internet will continue to rise, and the impact on eWOM is expected to be higher, reviews will be generated faster, especially in the post-covid era, when

people will have the opportunity to travel again. Having in mind the prediction that engagement on the internet will continue to grow, the fact that electronic word of mouth is extremely powerful in today's world and having already proven statements that online reviews have an impact on the decision making of the future guests, there is definitely a big potential in using text mining techniques during day-to-day activities, especially in the hospitality sectors. Likewise, it is logical to expect that the data science area will follow the steps of the IT industry in terms of predicted improvements, which will then contribute to improvements in the hospitality sector, as these two industries are closely related. These predicted improvements would consequently reduce the number of limitations and enhance even more the benefits in terms of sensitivity, accuracy, and precision.

When it comes to the evaluation of the current business of hotel Sotelia, analysis shows that guests of hotel Sotelia are eager to comment on the food and drinks, facilities, staff, and room amenities, which is not many different from guests' comments which can be seen in other hotels. These are the most commented aspects when it comes to the hotel industry in general, and it is not strange that guests of hotel Sotelia are also focused on these four. Luckily, these aspects are, in more than 80% of cases, mentioned as positive, which proves that hotel Sotelia is working in accordance with its main values and aims. It can also be concluded that hotel Sotelia is following the trends in the hotel industry but making specific offers to the targeted audience, promoting on the right channels to involve and attract all the potential guests, using contactless technologies, and digitalised services, and personalised welcoming messages. Moreover, it gives their guests the possibility to experience the local way of living and try local food. Last but not least, hotel Sotelia is following the latest trends in terms of environmental protection and sustainability, which showed as very important not only for nature but also for Sotelia's guests. Luckily, as hotel Sotelia is the best Wellness hotel in Slovenia, there are not many things, changes and actions that need to be undertaken, but a suggestion taken out of the analysis would be to avoid checking the reviews manually and hire an analyst, who will be responsible for following and understanding of guest reviews. In that way, management would have a good understanding and clear visualised results, which would then make it possible to build new strategies and enhance even more the customer experience.

Overall, this research has effectively answered the research questions that served as a foundation for digging a little bit further into a still relatively new and complicated field of study. The theoretical and case-study research required a lot of investigation and precision because the dataset was not that big, meaning that even a small number of incorrectly understood reviews could lead to inaccurate conclusions. Thus, this study was done very carefully to obtain the required results accurately. When some of the research and requisite capabilities were highlighted to the management, the hotel's awareness of the advantages of implementing text mining increased while challenging the hotel to dig further and explore which analysis would be the best for them. The recommendations given might be implemented in the future as the management is aware of the increasing online commenting

engagement trend to come and understands the advantages of using text mining. This thesis can also be helpful to smaller but also bigger hotels worldwide since text mining techniques can be used regardless of the size of the dataset.

Finally, as a relatively new field in IT, the topics stated in this thesis require constant and further research. The case study should be expanded, focusing more on the sensitivity of hidden pattern detection, for a better and deeper understanding of how to correctly spot and comprehend the fake reviews and not only fake but reviews are written with the intention to negatively impact the business' overall result. Furthermore, further research is needed in order to continue following the opportunities that text mining could bring in the future, but also to spot the limitation and to find new, faster and more accurate solutions how to overcome these limitations. Another limitation is mostly connected to translation, as mistakes made by software during the translation can make the results less reliable. Therefore, additional manual check is inevitable, and results should be taken with precautions. As mentioned in chapter 5.1, when the dataset is bigger, the probability of having better information in terms of quantity is higher, which in turn, increases to have more reliable and more accurate obtained results. Although text mining can be used regardless of the dataset, as a suggestion for future research, it would be proposed to inspect and gather reviews from more hotels in the same region, as this would allow the researcher to conduct more analyses, compare the results, and provide better overview of the Slovenian hospitality in general.

REFERENCE LIST

1. Agrawal, R., & Batra, M. (2013). A detailed study on text mining techniques. *International Journal of Soft Computing and Engineering, 2*(6), 118-121.
2. Analiytiks, L. (2019, November 17). Why Business Intelligence Is Important. Retrieved April 13, 2021, from https://analytiks.co/why-business-intelligence-is-important/
3. Antons, D., Grünwald, E., Cichy, P., & Torsten Oliver, S. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management, 50*(3), 329-351.
4. Awan, H., Siddiquei, A., Jabbar, A., Abrar, M., & Baig, S. (2015). Internal marketing and customer loyalty: A dyadic analysis. *Journal of Service Science and Management, 8*(2), 216-228.
5. Benfield, J. A., & William, S. J. (2006). Internet-based data collection: Promises and realities. *Journal of Research Practice, 2*(2), 1-15.
6. Berezina, K., Cobanoglu, C., Miller, B. L., & Kwansa, F. A. (2012). The impact of information security breach on hotel guest perception of service quality, satisfaction, revisit intentions and word-of-mouth. *International journal of contemporary hospitality management*, *24*(7), 991–1010.
7. Bhan, M., & Perpetua, N. F. (2020). Techniques and Issues in Text Mining. Journal of Computational and Theoretical Nanoscience, *17*(9-10), 4368-4374.

8. Bisio, F., Oneto, L., & Cambria, E. (2016). Sentic Computing for Social Network Analysis. In F. Pozzi, E. Fersini, E. Messina, & B. Liu, *Sentiment analysis in social networks* (pp. 71-90). Burlington, USA: Morgan Kaufmann.
9. Bowtell, J. (2015). Assessing the value and market attractiveness of the accessible tourism industry in Europe: a focus on major travel and leisure companies. *Journal of Tourism Futures, 1*(3), 203-222
10. Calheiros, A. C. (2015, December). Sentiment Analysis in Hospitalitiy Industry Using Text Mining: The Case of Portuguese Eco-Hotel. Lisbon, Portugal: ISCTE – University Institute of Lisbon.
11. Car, T., Stifanich Pilepić, L., & Šimunić, M. (2019). Internet of things (iot) in tourism and hospitality: Opportunities and challenges. *Tourism in South East Europe*, *5*, 163-175.
12. Chang, J., O'Reilly, C., Pontika, N., Haug, K., Owen, G., & Oudenhoven, M. (2018). *Text Mining 101 - What is text mining, how does it work and why is it useful?* Retrieved March 13, 2021 from https://www.fosteropenscience.eu/content/text-mining-101
13. Chauhan, P. S. (2016). Opinion Mining and Sentiment Analysis using Rapidminer. Vienna, Austria: Modul Vienna University.
14. Chittiprolu, V., Samala, N., & Bellamkonda, R. (2021). Heritage hotels and customer experience: a text mining analysis of online reviews. *International Journal of Culture, Tourism and Hospitality Research, 15*(2), 131-156.
15. Chuang, H.M., & Shen, C.C. (2008). A study on the applications of data mining techniques to enhance customer lifetime value — based on the department store industry. *2008 International Conference on Machine Learning and Cybernetics. 1*, pp. 168-173. Kunming, China: IEEE
16. Dabhade, V. (2021). *Conducting Social Media Sentiment Analysis: A Working Example*. Retrieved October 16, 2021 from https://www.expressanalytics.com/blog/social-media-sentiment-analysis/
17. Dang, S., & Ahmad, P. (2014). Text Mining: Techniques and its Application. *International Journal of Engineering & Technology Innovations, 1*(4), 22-25.
18. Dang, S., & Ahmad, P. (2015). A review of text mining techniques associated with various application areas. *International Journal of Science and Research, 4*(2), 2461-2466.
19. Dawar, N. (2013). When marketing is strategy. *Harvard business review, 91*(12), 100-108.
20. Deloitte. (2020). *The future of hospitality - Uncovering opportunities to recover.* Retrieved September 11, 2021 from https://www2.deloitte.com/content/dam/Deloitte/ca/Documents/consumer-industrial-products/ca-future-of-hospitality-pov-aoda-en.pdf
21. Dodds, R., & Butler, R. (2019). *Overtourism: Issues, realities and solutions* (Vol. 1). Berlin, Germany: De Gruyter.

22. Dogan, G., & Chi, C. G. (2020, July 5). Effects of COVID-19 pandemic on hospitality industry: review of the current situation and a research agenda. *Journal of Hospitality Marketing & Management, 29*(5), 527-529.
23. eBizMBA Inc. (2021). *Top 15 Best Travel Websites | October 2021*. Retrieved November 14, 2021, from http://www.ebizmba.com/articles/travel-websites
24. Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data.* New York, USA: Cambridge University Press.
25. Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management, 66*, 53-61.
26. Godnov, U., & Redek, T. (2019). The use of user-generated content for business intelligence in tourism: insights from an analysis of Croatian hotels. *Economic Research-Ekonomska Istraživanja, 32*(1), 2455-2480.
27. Gössling, S., Scott, D., & Hall, C. M. (2021). Pandemics, tourism and global change: a rapid assessment of COVID-19,. *Journal of Sustainable Tourism, 29*(1), 1-20.
28. Gupta, V., & Lehal, G. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence, 1*(1), 60-76.
29. Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing, 4*(1), 1-34.
30. Hoontrakul, P., & Sahadev, S. (2008). Application of data mining techniques in the on-line travel industry: A case study from Thailand. *Marketing Intelligence & Planning, 26*(1), 60-76.
31. HotelTechReport. (2021). *Digital Transformation in the Hotel Industry*. Retrieved June 13, 2021, from https://hoteltechreport.com/news/digital-transformation
32. Imhoff, C., & White, C. (2011). Self-service business intelligence: Empowering users to generate insights. *The Data Warehousing Institute (TDWI) best practices report*, 40.
33. Inzalkar, S., & Sharma, J. (2015). A survey on text mining-techniques and application. *International Journal of Research In Science & Engineering, 24*, 1-14.
34. Kääriäinen, J., Tihinen, M., Ailisto, H., Komi, M., Parviainen, P., Tanner, H., Tuikka, T. and Valtanen, K. (2016). The Industrial Internet in Finland: on route to success? *VTT TECHNOLOGY 278,* 1-88.
35. Khillar, S. (2020, 7 14). *Difference Between Data Mining and Data Science*. Retrieved June 15, 2021, from http://www.differencebetween.net/technology/difference-between-data-mining-and-data-science/
36. Kim, C. S., Bai, B. H., Kim, P. B., & Kaye, C. (2018). Review of reviews: A systematic analysis of review papers in the hospitality and tourism literature. *International Journal of Hospitality Management, 70*, 49-58.
37. Kwartler, T. (2017). Text mining in practice with R. New York, USA: John Wiley & Sons.
38. Laudon, K. C., & Laudon, J. (2004). *Management Information Systems: Managing the Digital Form (13th Global Edition).* London, England: Pearson Education Limited.
39. Lee, C.-H., & Yang, H.-C. (2003). A multilingual text mining approach based on self-organizing maps. *Applied Intelligence, 3*(18), 295-310.

40. Lee, S., & Siau, K. (2001). A review of data mining. *Industrial Management & Data Systems, 101*(1), 41-46.
41. Li, H., Meng, F., Jeong, M., & Zhang, Z. (2020). To follow others or be yourself? Social influence in online restaurant reviews. *International Journal of Contemporary Hospitality Managemen, 32*(3), 1067-1087.
42. Limberger, P., Anjos, F., Meira, J., & Anjos, S. (2014). Satisfaction in Hospitality on TripAdvisor.com: An Analysis of the Correlation Between Evaluation Criteria and Overall Satisfaction. *Tourism & Management Studies, 1*(1), 59-65.
43. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.
44. Liu, K., Xu, L., & Zhao, J. (2012). Opinion target extraction using word-based translation model. *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1346-1356). Beijing, China: Institute of Automation, Chinese Academy of Sciences.
45. Madhoushi, Z., Hamdan, A., & Zainudin, S. (2019). Aspect-based sentiment analysis methods in recent years. *Asia-Pacific Journal Of Information Technology And Multimedia, 7*(2), 79 - 96.
46. Maks, I., & Vossen, P. (2013). Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions? *International Conference Recent Advances in Natural Language Processing RANLP 2013*, (pp. 415-419). Hissar, Bulgaria: INCOMA Ltd. Shoumen.
47. Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, 30 (12), 3514-3554.
48. Masset, P., & Weisskopf, J.-P. (2021). *2021 Top Hospitality Industry Trends*. Retrieved February 12, 2022, from https://hospitalityinsights.ehl.edu/hospitality-industry-trends.
49. MBA Knowledge Base. (2018). *Ethical, Security, Legal and Privacy Concerns of Data Mining*. Retrieved April 16, 2021 from https://www.mbaknol.com/information-systems-management/ethical-security-legal-and-privacy-concerns-of-data-mining/
50. McGarrity, L. (2017). *What Sentiment Analysis Can Do for Your Brand*. Retrieved November 13, 2021, from https://www.marketingprofs.com/opinions/2016/29673/what-sentiment-analysis-can-do-for-your-brand
51. Norris, D. (2013). *Rapid Miner-a potential game changer*. Retrieved November 18, 2021, from https://www.bloorresearch.com/2013/11/rapidminer-a-potential-game-changer/
52. OECD. (2020). *OECD Tourism Trends and Policies 2020.* Retrieved November 23, 2021 from https://www.oecd.org/cfe/tourism/OECD-Tourism-Trends-Policies%202020-Highlights-ENG.pdf
53. Ramanathan, V., & Meyyappan, T. (2013). Survey of Text Mining. *International conference on technology and business management*, *80(4),* 508-514.
54. Ranjan, J. (2009). Business intelligence: Concepts, components, techniques and benefits. *Journal of theoretical and applied information technology, 9*(1), 60-70.

55. Rao, P. (2019). *Fine-grained Sentiment Analysis in Python (Part 1)*. Retrieved December 05, 2021 from https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4
56. Rather, R. A., & Sharma, J. (2017). The effects of customer satisfaction and commitment on customer loyalty: Evidence from the hotel industry. *Journal of Hospitality Application & Research (JOHAR), 12*(2), 42-60.
57. RevolveAI. (2020, March 10). *Predictive analytics advantages and disadvantages.* Retrieved May 21, 2021, from https://revolveai.com/predictive-analytics-advantages-and-disadvantages/
58. Rouse, M. (2019). *Big Data.* Retrieved June 27, 2021, from https://searchdatamanagement.techtarget.com/definition/big-data
59. Sagayam, R., Srinivasan, S., & Roshni, S. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal of Computational Engineering Research, 2*(5), 1443-1446.
60. Schultz, J. (2019). *How Much Data is Created on the Internet Each Day?* Retrieved August 17, 2021 from https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/
61. Shaidah, J., & Alfawareh, H. M. (2012). Techniq Techniques, Applications and Challenging Issue in Text Mining ues, Applications and Challenging Issue in Text Mining. *IJCSI International Journal of Computer Science Issues*, *9*(6), 431-436.
62. Statista (2020). *Number of user reviews and opinions on TripAdvisor worldwide 2014-2019*. Retrieved June 23, 2021, from https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/
63. Raič, M. (2019). Statistika. Koper, Slovenia: UP FAMNIT, Biopsihologija.
64. Roberts, C. (2021, May). *6 Key Tech Trends to Watch For In Hospitality Industry*. Retrieved December 07, 2021, from Apextech https://apextechinc.com/6-key-tech-trends-to-watch-for-in-hospitality-industry/
65. Tan, A.-H. (1999). Text mining: The state of the art and the challenges. *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases. 8*, (pp.65-70). Beijing, China: Citeseer
66. Tanwar, M., Duggal, R., & Khatri, S. (2015). Unravelling unstructured data: A wealth of information in big data. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, (pp. 1-6). Noida, India: IEEE
67. Taşkın, D. (2016). An Application of Text Mining to Capture and Analyze eWOM: A Pilot Study on Tourism Sector. In R. Sumangla, & P. Avinash, *Capturing, Analyzing, and Managing Word-of-Mouth in the Digital Marketplace* (pp. 168-186). Pennsylvania, USA: IGI Global.
68. Tatulli, K. (2019). *53 Travel Industry Statistics Planners Should Know*. Retrieved September 18, 2021 from aventri.com/blog/53-travel-industry-statistics-for-planners
69. Tepeci, M. (1999). Increasing brand loyalty in the hospitality industry. *International journal of contemporary hospitality Management, 11*(5), 223-230.

70. Terme Olimia. (n.d.). *Wellness Hotel Sotelia ****s*. Retrieved September 20, 2021, from https://www.terme-olimia.com/en/hotels/wellness-hotel-sotelia-s
71. Business Research Company. (2022). *Information Technology Global Market Report 2022.* Retrieved September 20, 2021 from https://www.thebusinessresearchcompany.com/report/information-technology-global-market-report
72. *Tripadvisor* (2020). Retrieved August 02, 2021, from https://www.tripadvisor.com/
73. Tripadvisor (2021). *Hotel Sotelia*. Retrieved December 12, 2021, from https://www.tripadvisor.com/Hotel_Review-g1057702-d1057018-Reviews-Hotel_Sotelia-Podcetrtek_Styria_Region.html
74. Tsang, A., & Prendergast, G. (2009). Is a "star" worth a thousand words? The interplay between product-review texts and rating valences. *European Journal of Marketing*, *43*, 1269 - 1280.
75. Turc, T. (2019). AJAX Technology for Internet of Things. The 12th International Conference Interdisciplinarity in Engineering. 32, pp. 613-618. Targu Mures, Romania: Procedia Manufacturing.
76. Ukhalkar, P., Phursule, D., Gadekar, D., & Sable, D. (2020). Business Intelligence and Analytics: Challenges and Opportunities. *International Journal of Advanced Science and Technology, 29*(12s), 2669-2676.
77. Vidhya, K., & Aghila, G. (2010). Text mining process, techniques and tools: an overview. *International Journal of Information Technology and Knowledge Management, 2*(2), 613-622.
78. Xindong, W., Xingquan, Z., Gong-Qing, W., & Wei, W. (2013). Data Mining with Big Data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97-107.
79. Zelia, B., Rui, C., Gorete, D., & Martins, A. (2020). eWOM of Guests Regarding Their Hotel Experience: Sentiment Analysis of TripAdvisor Review. In *Handbook of Research on Social Media Applications for the Tourism and Hospitality Sector* (pp. 295-308). Pennsylvania, USA: IGI Global.
80. Zhao, X., Wang, L., Guo, X., & Law, R. (2015). The influence of online reviews to online hotel booking intentions. *International Journal of Contemporary Hospitality Management*, *27*(6), 1343 - 1364

APPENDICES

Appendix 1: Povzetek (Summary in Slovene language)

Družbeni mediji in spletne strani, povezane s potovanji, kjer ima elektronsko ustno izročilo (eWOM) prvo prioriteto, imajo izjemen pomen v današnjem svetu, zlasti v sektorju gostinstva in turizma. Številni potencialni gostje se bodo odločali na podlagi recenzij oz. ocen, saj te veljajo za zaupanja vredne, objektivne in nepristranske. Lahko zgodi, da so hotelski managerji preobremenjeni s podatki, nimajo pa dovolj razumevanja in znanja, da bi iz recenzij izluščili uporabne informacije. Tako morajo hoteli za ohranjanje dobrega položaja na trgu slediti trenutnim trendom v sektorju gostinstva in turizma in na podlagi spletnih ocen in ocen na kraju samem spremljati želje gostov, zasedenost, glavne konkurente in zadovoljstvo. Zato morajo hotelski managerji najti najboljši način upravljanja hotelov z združevanjem informacijske tehnologije z drugimi managerskimi pristopi. Staromodni način obdelave besedilnih informacij je zahteval precejšnja finančna sredstva, veliko čas in človeških virov, saj je bilo za zbiranje informacij, njihovo analizo in vizualizacijo potrebno človeško delovanje. Danes je odkrivanje informacij iz velikih podatkovnih nizov s pomočjo podatkovnega rudarjenja veliko hitrejše. Če spletne vsebine ustrezno analiziramo, lahko prinesejo dragoceno znanje o strankah, pa tudi o ukrepih, ki jih je treba izvesti za izboljšanje poslovanja.

Znan slovenski hotel Wellness hotel Sotelia je bil vzet kot primer, da bi prikazali praktične možnosti analize recenzij, ki lahko prinesejo koristi hotelu. Raziskava je bila izvedena z namenom preveriti, ali lahko uporaba tehnik podatkovnega rudarjenja, kot je analiza razpoloženja, pomaga ne le pri zbiranju in analizi podatkov, temveč tudi pri preverjanju, ali lahko te tehnike izboljšajo poslovanje in podajo pravo sliko mnenja strank, ki ga delijo na potovalnih spletnih straneh. Torej, da bi lahko razkrili, ali je mogoče tehniko rudarjenja po besedilih uporabiti za raziskovanje skritih vzorcev v vedenju gostov, in da bi nakazali, kakšne so prednosti in omejitve uporabe rudarjenja po besedilih. Poleg tega je bil namen preučiti, ali lahko uporaba rudarjenja po besedilih privede do splošnih izboljšav in konkurenčne prednosti, če so analize pravilno izvedene.

Ocene s spletnih strani Tripadvisor.com, Google.com in Booking.com iz obdobja od januarja 2009 do januarja 2020 so bila zbrana s pomočjo programskega jezika Python in orodja imenovanega Octoparse. Potem so bila analizirana s programom RapidMiner. Ta program je bil uporabljen za analize rudarjenja besedila, med katerimi je bila za to raziskavo najpomembnejša analiza sentimenta za posamezne vidike. Ko so bili rezultati na voljo, je bil opravljen intervju z managerjem hotela Sotelia, da bi preverili veljavnost rezultatov in na koncu ugotovili, ali se lahko managerji zanašajo na to vrste analiz.

Glavni rezultati so pokazali, da gostje hotela Sotelia želijo več pozornosti nameniti stvarem, povezanim s hrano in pijačo, objekti, in osebjem. Ti štirje vidiki so se pojavili v skoraj vsaki drugi oceni, v več kot 78 % pa so bili označeni kot pozitivni. Ta visok odstotek je dokaz, da hotel Sotelia deluje v skladu z glavnim poslanstvom hotela, kar je bilo potrjeno tudi v intervjuju. Po besedah direktorja so bili glavni vidiki, ki jih je prepoznal program, točni.

Poleg tega je bilo ugotovljeno, da je na podlagi hotelskih ocen vodstvo lahko zadovoljno s trenutnim poslovanjem, saj je hotel v skoraj 60 % primerov ocenjen z najvišjo oceno, ob upoštevanju mnenj s treh omenjenih spletnih strani. Raziskava je tudi pokazala, da je korelacija med čistočo in ceno zelo močna, kar pomeni, da če je čistoča ocenjena pozitivno, obstaja velika verjetnost, da bo tudi splošna ocena pozitivna, in obratno. Na razgovoru je bilo potrjeno, da so gostje občutljivi na čistočo, kar je dokaz, da se lahko v prihodnosti uporabi kontingenčna tabela za preverjanje povezave med različnimi vidiki. Po predstavitvi vseh analiz je manager potrdil, da bo moral management spremeniti način pregledovanja in analize mnenj. Hotel bo poskušal zamenjati ročno pregledovanje mnenj z uporabo rudarjenja po besedilih in s tem pospešiti postopek, kar bo povečalo znanje o gostih, njihovih željah in namerah, s čimer se bodo povečale možnosti za konkurenčno prednost.

Kot prednost uporabe besedilnega rudarjenja je bila ugotovljena možnost zelo hitrega zbiranja, analiziranja in vizualizacije podatkov, zaradi česar je več časa za osredotočanje na druge dejavnosti, povezane s poslovanjem. Poleg tega lahko natančni rezultati, pridobljeni z analizami, vodijo do konkurenčne prednosti, hkrati pa dajejo hotelu neposreden znak, kaj lahko ohrani tako, kot je, in kaj mora spremeniti, da bodo njegovi gostje bolj zadovoljni in zvesti. Še več, hoteli z uporabo različnih analiz lahko pregledajo vse vidike, povezane z njihovim poslovanjem, razmerja med dvema ali več vidiki, imajo popoln pregled nad spletnimi ocenami in trendi, lahko primerjajo sebe in svoje konkurente, preizkusijo različne analize, kot so analiza sentimenta za posamezne vidike, razvrščanje v skupine, odločitvena drevesa, kategorizacijo, odkrivanje tem, povzemanje in številne druge. Vse navedene analize lahko pomagajo pri odkrivanju skritih vzorcev glede sentimenta strank. Kljub temu pa poleg vseh prednosti še vedno obstajajo nekatere omejitve, ki jih je treba upoštevati pri uporabi rudarjenja po besedilih, na primer pomanjkanje znanja, jezikovne ovire, vključno z napačnim prevodom, dvoumnost besedila, sposobnost programske opreme, da prepozna ironijo, dvojno zanikanje in sposobnost razumevanja stavka na pravi način. Dodatno vprašanje je, kako se lahko raziskovalci z uporabo besedilnega rudarjenja soočijo s težavami glede zasebnosti in zakonodaje, pa tudi z etičnimi vidiki tovrstnih analiz. Za vsako od navedenih težav je bila navedena možna rešitev.

Pregled literature in študija primera sta skupaj prispevala k širšemu razumevanju izvajanja rudarjenja po besedilih, priložnosti in ovir. Na splošno je ta raziskava odgovorila na raziskovalna vprašanja, ki so služila kot osnova za poglobitev v še vedno razmeroma novo in zapleteno področje preučevanja. Teoretična raziskava in raziskava primera sta zahtevali veliko dela in natančnosti pri tem, saj nabor podatkov ni bil tako velik, kar pomeni, da bi že majhno število nepravilno razumljenih pregledov lahko privedlo do netočnih zaključkov. Zato je bila ta študija izvedena zelo previdno, da bi pridobili ustrezne rezultate. To magistrsko delo je lahko v pomoč tudi manjšim, pa tudi večjim hotelom po svetu, saj se tehnike rudarjenja po besedilih lahko uporabljajo ne glede na velikost nabora podatkov.

Omejitve te raziskave so večinoma povezane s preverjanjem mnenj, saj so bili neizogibni dodatni ročni pregledi, kar je podaljšalo čas, potreben za analize. Zaradi možnih napak, ki

jih je programska oprema naredila pri prevajanju besedila, je treba rezultate obravnavati previdno. Druga omejitev izhaja iz analize zgolj enega primera, saj bi lahko analiza več primerov omogočila boljši vpogled v problematiko, s tem pa tudi posplošitev ugotovitev.

Appendix 2: Interview questions

This interview is conducted with one of the representatives of hotel Sotelia. The aim of this interview is to inspect if text mining tools and techniques can easily and reliably be used for analysing reviews as stated in research and if hotel can take further actions in accordance with results. Moreover, does hotel Sotelia already use some of the similar methods and if the results stated in the thesis can be described as correct and valuable. The name and surname of the participant on Interview will stay anonymous.

1. What is your strategy in handling many reviews on travel sites?
2. Do you have employees especially for that segment of business or you are analysing reviews with help od text mining techniques?
3. If you are analysing reviews manually (reading one by one), what is the reason for doing so? Lack of knowledge, lack of qualified labor or any other reason_
4. If you are analysing reviews with help od text mining techniques, can you please:
   - state on which analysis you rely on?
   - explain the steps in your process (briefly)
   - state three most important benefits of using it
   - how often are you doing analysis?
5. In your opinion, could analysing reviews lead to potential competitive advantage and how important analysing review to your hotel is?
6. On which aspect of business, you are focused the most according to your mission and vision? Cleanliness, Relaxation, Facilities, Room Amenities?
7. Do you think that using text mining techniques in analysing hotel reviews is reliable way of getting information about aspects (for example about cleanliness)?
8. What is the biggest challenge for hotel Sotelia in analysing reviews and how do you overcome those challenges?
9. I have noticed that you are answering on reviews on TripAdvisor.com, which is, according to many studies, especially important action for keeping the quests and maintaining their satisfaction. How do you recognize a fake review, and have you noticed many of them in the last 5 years?
10. According to the research I have made, people who stayed in hotel Sotelia are most likely to comment the following: facilities, room amenities, staff, food/drinks. They are less likely to comment other aspects such as: location, wi-fi, comfort of beds. Does this seem correct to you and can we say that hotel Sotela is working in accordance with its three main quotes on the site:
    - “Created for well-being”
    - “Pleasures behind closed eyes”
    - ” Health too goes through the stomach”

11. Do you think that hotel managers could gain substantial knowledge about the customers and in that way shape hotel's offers according to what other quests had said?
12. While doing research, I have noticed that quests are posting the highest number of reviews on TripAdvisor.com and the lowest number of reviews can be found by typing hotel Sotelia directly on Google.com. As you are doing marketing promotions, are you focus on promoting hotel on special site or it there any other reason for such differences? Below you can find statistics:

Website	Reviews per site in numbers	Reviews per site in percentages
Booking.com	262	23.93%
Google Reviews	102	9.32%
TripAdvisor.com	731	66.76%
Grand Total	1095	100.00%

Rate	Booking.com	Google.com	Tripadvisor.com
1	0.76%	3.92%	0.55%
2	1.91%	0.00%	0.96%
3	10.31%	5.88%	5.61%
4	36.64%	18.63%	31.46%
5	50.38%	71.57%	61.42%

13. The association correlation shown that if cleanliness is commented in the positive context, then the whole review would have high rate (4 or 5) and vice-versa.

Aspects	Cramer's coefficient
Facilities/Review	0.293
Food&drinks/Review	0.315
Staff/Review	0.369
Room amenities/Review	0.334
Value/Review	0.298
Beds/Review	0.674
Cleanliness/Review	0.877
Comfort/Review	0.239
Customer support/Review	0.545
View/Review	0.298
Design/Review	0.268
Location/Review	0.278
Payment/Review	0.235
Quietness/Review	0.283
Wi-Fi/Review	#DIV/0!