NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

How to monitor and generate intelligence for a DMO from online reviews

Mário Monteiro Andrade Filho

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

HOW TO MONITOR AND GENERATE INTELLIGENCE FOR A DMO FROM ONLINE REVIEWS

By

Mário Monteiro Andrade Filho

Dissertation presented as the partial requirement for obtaining a Master's degree in Information Management, Specialization in Marketing Intelligence.

Advisor: Diego Costa Pinto, Phd.

Co-supervisor: Claudio Gottschalg Duke, Phd.

August 2020

DEDICATION

I would first like to thank my wife Marisa, for the support given and my children, as well as my mother Maria Helena. I would also like to thank my professor advisor Mr. Diego Costa, my co-advisor Mr. Cláudio Duque and also to Mr.Said Mohammad - NRCC.

ABSTRACT

Social media and customer review websites have changed the way the tourism sector is managed. Social media has become a new source of information, due to the large amount of UGC / e-Wom generated by consumers An information that is "available" but at the same time noisy and of great volume, which makes it difficult to access and analyze. This study investigates and verifies the possibility of using data present in content reviews of a Content Web Site Review - TripAdivsor - to generate actionable information for a Destination Management Organization. With a focus on negative reviews, tourist attractions of Lisbon and using the “R code” and its packages, the study shows that with the correct technique chosen and the action of an intelligence analyst, data can be extracted and provide substrate for actions, strategy and intelligence generation – which is Social Media Intelligence. The findings prove that the flood of web 2.0 data can serve as a source of intelligence for the Destination Management Organization (DMO). By monitoring sites like TripAdvisor, a DMO can hear what tourists talk about attractions and thereby generate insights for intelligence and strategy actions. A DMO can even, analyzing this data, make your attractions more desirable, and even act in adverse situations, reducing risky situations.

KEYWORDS

Destination Management Organization; Monitoring; Social Media Intelligence; Tourism; TripAdvisor .

INDEX

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

CASM - Centre of Analysis of Social Media
CRW- Content Review Web Site
DMO - Destination Management / Marketing Organization
eWOM - Electronic word-of-mouth
GDP - Gross Domestic Product
HTML - Hypertext Markup Language
ICT -Information and Communication Technology
KIT - Key Intelligence Topics
NLP -Natural Language Processing
NRC -National Research Council Canada
SA - Sentiment Analysis
SOCMINT -Social Media Intelligence
SMM - Social Media Mining
TA- TripAdvisor
UGC - User Generated Content
UNWTO - World Tourism Organization
WOM - Word-of-mouth
WTTC- World Travel & Tourism Council

1- INTRODUCTION

With the growth in social media (SM) usage in recent years, people have the opportunity to share their ideas, feedback, opinions, and interests more than ever. As a result, according to several authors, like Agichtein, Castillo, Donato, Gionis, and Mishne (2008) and Patin , Pitta, and Quinones (2012), social media has become a new source of information. From the data generated by several other consumers who have already used a particular product or service, other interested consumers have the opportunity to get to know the product better, its positive and negative aspects, to obtain more updated, reliable and relaxed information, thus reducing the risks and possibility of regret (Gretzel & Purifoy, 2008).

Besides that, Web 2.0 provided for consumers a new communication platform similar to that of the word of mouth that also empowers consumers- social media networks - and changing the way information is being produced, transferred, and consumed (Bindra, Kandawal, Singh & Khan, 2012).

Gretzel (2006) brings an interesting point of view when he states that the media allow to tell stories on a "24/7" basis, reinforcing even more its ubiquity. Social media (SM) also removed spatial and time constraints that were inherent in traditional methods of communications provided online tools that enable one to many sharing of multimedia content and employ easy to use interfaces that will allow even non-specialists to share and connect. In addition to interactivity and communication, due to the characteristics and functionalities of SM, these can also be extremely useful tools and assume an essential role in the management, strategic planning, development, and promotion, as well as offer opportunities for innovation processes.

The broad adoption of social media by users has generated an exponential increase in data and content that offer, although not in full, opportunities to be treated and transformed into information and knowledge for organizations.

Newton (2008) states that Information and Communication Technology (ICT) is fundamental to the competitive organizational strategy. As SM was provided by the evolution and progress of ICTs, they can be fundamental as a source of information for generating strategies. The knowledge provided by the extraction of information from social media can increasingly be recognized as essential to the competitive advantage of any organization.

It can be added several factors to explain why social media data can generate value for companies:

- A significant amount of information associated with customers, competitors, industries, and technology can be gathered from social media.

- the rapid growth of social media users around the world allows organizations to access a wide range of consumer thoughts easily, opinions, and behaviors on time (Chen, Chiang, & Storey, 2012);

Regardless of all the possibilities and benefits of social media as a source of information, the unstructured nature and volume of this information, which is distributed across a variety of social media sources, make the task of extracting useful information a challenge (Dai, Kakkonen, & Sutinen,2011).

Agarwal and Ylliasi (2010) argue that different issues affect the quality of social media data. The exponential growth of SM data can also create the problem of information overload, and finally, the use of slang, which is sometimes intensified within the cultural context, is another challenge to extract meaningful information (Töllinen, Järvinen, & Karjaluoto,2012). To summarize, SM data is not structured and of uncertain credibility, and this requires special treatment to make them meaningful and ready for business use.

However, monitoring, storing, analyzing, and transforming information automatically into organized knowledge is considered difficult because of the large and complex amount of information and data on many different social media platforms that cannot be discovered or easily identified using conventional database management systems, and therefore require the use and creation of tools, new methods and scientific techniques for collecting and analyzing social media data, which are still scarce (Stiegltiz & Dang-Xuan, 2012).

It is important to emphasize that the monitoring of social media content should not be restricted to metrics and the collection of information and experiences of users. They need to be interpreted, analyzed, and studied to obtain relevant information, identify opportunities, failures, needs, expectations, experiences, desires, and critiques of users and consumers (potential and real).

Information and knowledge can be obtained from data through techniques of interpretation, annotation, classification, grouping, summarization, among others, allowing the association and correlation of other information, many of them being part of a set of techniques, tools, procedures, and algorithms known as Data Mining (Santos, 2009). So, data mining becomes an essential tool for collecting information, especially from the accumulation of material posted on social networking sites and content on the Internet (Chen et al., 2012). As a consequence, it is imperative to look for data collection alternatives that allow for cost reduction and assistance in monitoring the impact of the activity.

A successful organization should have the ability to monitor the external environment, capture and analyze all available information (e.g., customer reviews, product and service reviews), generate intelligence and identify what happened and try to predict which may occur in the future. As many companies are unfamiliar with social media as a source of information to monitor the business environment (Dai et al., 2011) and there is also a lack of sufficient understanding of the social media mining process, the proposal would be to show how data social media obtained through monitoring can serve as a substrate for generating Intelligence.

The advent of Web 2.0, mainly in the format of media and social networks, has caused significant changes and transformations in tourism, especially about the production of content by users and the sharing of information and content between users and consumers, gaining popularity among online activities of travellers (Xiang & Gretzel, 2010).

Concomitantly, the Internet's interactivity, customization, and vast information resources provided tailored search and content to users, being able to cover almost any personal preference. As a natural consequence, the Web became one of the most effective means for potential tourists to search for information (Werthner & Klein, 1999).

When tourists decide to travel, they find difficulties when assessing the quality provided if they have not visited these locations before (Kim, Lehto, & Morrison, 2007). Tourism products are experiential - intangible and even impossible to be physically evaluated before purchase. As a result, their purchases are considered risky, and information-intensive in terms of their decision-making process (Mcintosh, Goelder & Ritchie, 1995). Although they are experienced products related to tourism, tourist attractions have fewer characteristics that can be made tangible. So, in the case of tourist attractions, reviews play a vital role because attractions cannot be assessed by some standard aspects, such as hotels and restaurants.

Tourism organizations – Destination Management Organization can take advantage of the vast amount of information generated on websites, such as TripAdvisor, to generate intelligence and analyze what customers are talking about their destinations.

TripAdvisor is one of the most recognizable consumer-generated content sites or consumer review websites (CRW). Consumer review websites (CRW) are social media applications that enable users to upload product-related reviews and ratings. TripAdvisor, among the leading CRWs, hosts more than 630 million reviews for more than 7 million businesses and has 455 million unique monthly visitors (TripAdvisor 2017). Researchers have already recognized TripAdvisor as being the most famous such site among tourists (Xiang & Gretzel, 2010).

In order to focus on research, we narrow our field of study for the Destination Management Organization (DMO) – VisitLisboa (THE LISBON TOURISM ASSOCIATION) - a nonprofit Private Association. It is the Regional Tourism Promotion Agency for the Region of Lisbon since 2004, maintaining international tourism promotion as its main activity.

Why choose Tourism Sector? Because the tourism sector in Portugal is responsible for 20% of employees and 17,8% of the Gross Domestic Product - GDP (WTTC 2018), that is an important generator of employment, income, and wealth.

Thus, it becomes a rich source of data for satisfaction analysis with destinations and tourism products. Another important feature is that it enables methodologies for extracting data and transforming it into essential information to assist public and private managers in decision-making.

Since the advent of the media, Tourism and Hospitality have been revolutionary industries for the adoption of online criticism as a means to get feedback from customers (Lehto, Park, Park, & Lehto, 2007). However, how does DMO use these reviews?

The big question - important to repeat - is: despite this large amount of data, how can organizations turn it into actionable information and knowledge?

I - How can DMO - VisitLisboa conduct media monitoring to capture and analyze media data?

II - How can DMO - VisitLisboa identify the strengths and weaknesses of tourist attractions using media information? Would it be possible to identify the strengths and weaknesses of a DMO's attractions and position them effectively in the market using media information?

And to help answer these questions, we will focus on negative comments. Several authors (Knowles, 2019; Faed & Forbes, 2010; Patel, 2018; Krause, 2018; Gin, 2016) state that through the analysis of negative comments the company can perceive problems previously hidden, learn more and improve faster.

In addition to answering the previous questions, we aim to create a social media intelligence framework to support a DMO to use external intelligence derived from monitoring social media, specific to your business, to improve service quality, innovation and decision- making.

The present study aims to link the monitoring of social networks to generate actionable information - intelligence for a DMO. The first is linked to the area of communication and advertising and the second is more linked to the hospitality sector. Almost nothing addresses the issue of tourist attractions, points that define a destination, and consequently the work of a DMO. We believe that through the use of them we will be able to acquire more relevant information that may lead to a more forceful and effective action by DMO VisitLisboa.

2 - LITERATURE REVIEW

Systematic literature reviews are a key component of much academic research. The systematic literature review has been carried out in several different phases to overcome some of the weaknesses and limitations of traditional literature reviews. Firstly, we identified the research question(s)/objective and identified relevant literature. Literature was selected based on their relevance to online reviews in the hospitality and tourism fields. Secondly, we made decisions about what research to include OR exclude.

Much of the work in the area – as searched on B-on, Web of Science, and Google Scholar- utilizing the words "TripAdvisor " **and** "social media mining " **and** "destination management organization " or "DMO" or " destination marketing organization" is concerned with two aspects: generating content and analyzing content. The first is linked to the area of communication and advertising and the second is linked to the hotel and hospitality industry.

This the search in Web of Science:

TS=(social media mining AND TripAdvisor AND sentiment analysis AND social media intelligence AND destination management organization OR DMO or destination marketing organization) – 2008 to 2020.

Refined in: HOSPITALITY LEISURE SPORT TOURISM OR MANAGEMENT OR BUSINESS OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR SOCIAL SCIENCES INTERDISCIPLINARY) AND SOURCES : (TOURISM MANAGEMENT OR ADVANCES IN HOSPITALITY AND TOURISM RESEARCH AHTR OR JOURNAL OF DESTINATION MARKETING MANAGEMENT OR JOURNAL OF TRAVEL RESEARCH OR JOURNAL OF TRAVEL TOURISM MARKETING OR CURRENT ISSUES IN TOURISM OR INTERNATIONAL JOURNAL OF TOURISM RESEARCH OR INTERNATIONAL JOURNAL OF CONTEMPORARY HOSPITALITY MANAGEMENT OR INTERNATIONAL JOURNAL OF TOURISM CITIES OR TOURISM AND HOSPITALITY RESEARCH OR TOURISM MANAGEMENT STUDIES OR JOURNAL OF SUSTAINABLE TOURISM OR JOURNAL OF TOURISM FUTURES OR WORLDWIDE HOSPITALITY AND TOURISM THEMES OR ANNALS OF TOURISM RESEARCH OR ASIA PACIFIC JOURNAL OF TOURISM RESEARCH OR TOURISM MANAGEMENT PERSPECTIVES OR SCANDINAVIAN JOURNAL OF HOSPITALITY AND TOURISM OR INTERNATIONAL JOURNAL OF CULTURE TOURISM AND HOSPITALITY RESEARCH OR JOURNAL OF HOSPITALITY TOURISM RESEARCH OR TOURISM RECREATION RESEARCH OR INFORMATION TECHNOLOGY TOURISM OR JOURNAL OF HOSPITALITY MARKETING MANAGEMENT OR JOURNAL OF TOURISM AND SERVICES OR TOURISM REVIEW OR MARKETING AND MANAGEMENT OF INNOVATIONS OR PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON TOURISM RESEARCH ICTR 2018 OR INFORMATION AND COMMUNICATION TECHNOLOGIES IN TOURISM 2010 OR JOURNAL OF HOSPITALITY AND TOURISM MANAGEMENT OR TOURISM ANALYSIS).

After stipulating the parameters seen above, we had 540 results.

To analyze them initially, we used the Vosviewer software version 1.6.14. VOSviewer is a popular freely available software tool for visualizing bibliometric networks. The functionality of the VOSviewer is especially useful for displaying large bibliometric maps in an easy-to-interpret way.

To interpret the results of Vosviewer map is important to focus on two major things :

- Size of the circle or word -The larger a term, the higher the frequency of occurrence of the term

- Distance: In general, the smaller the distance between two terms, the higher the relatedness of the terms, as measured by co-occurrences.

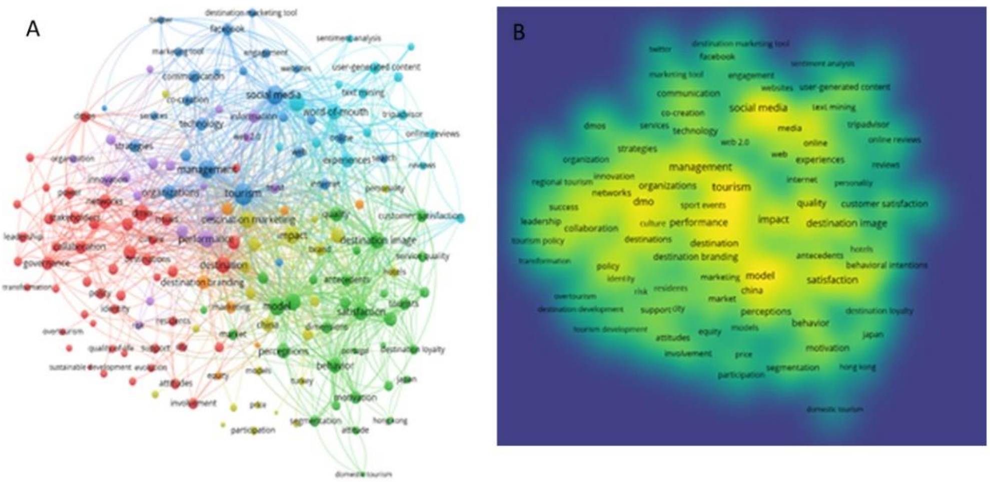

Figure1 - A - Map of results keywords from Web of Science B - Density of the terms

First of all, is possible to see (fig.1A)seven clusters of words from the search. Through the size of the circles or the words, it can be seen that the most prominent and frequent items are Social Media, Tourism, and Management.

According to the figure above (fig.1B), is possible to state that both the terms TripAdvisor and Strategies are low in color density, which shows the low frequency of these keywords in bibliographic research. It is important to deepen the analysis and for that, we will check the main keywords related to this work: TripAdvisor, Social Media Mining, Strategy, Destination Management Organization / DMO and check their position on the map and the strength of their links.

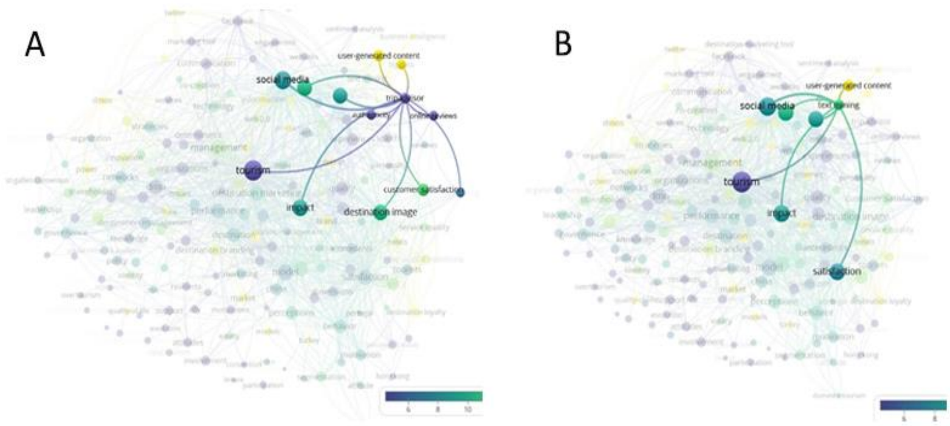

Figure 2- A – TripAdvisor and links B - Text Mining and Links

As already shown in figure 2A, we can see that the term TripAdvisor is peripheral in relation to the center of the map, which shows us that it is still a little-explored topic. The strongest links are Tourism and Destination Image.

The term “text mining” (fig.2B) is also peripheral and has stronger links to Tourism, Social Media, Impact, Satisfaction, and UGC – User Generated Content. As will be shown in the paper, the techniques of text mining, social media mining and sentiment analysis have been developing a lot and will make it possible to extract unstructured texts and transform them into structured and analyzed regarding sentiment.

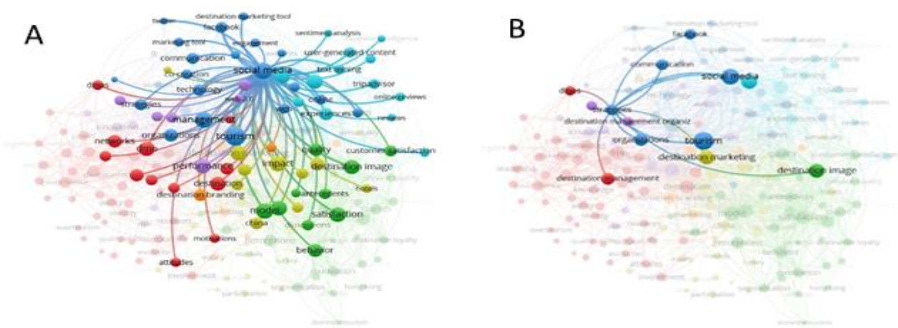

Figure 3- A – Social Media and Links B - Strategies and Links

The term “social media” belongs to the blue cluster, where it has stronger links with “Tourism”, “Management”, “Marketing Tools” and “Communication” (fig.3A) .

The term “strategies”, also belonging to the blue cluster, has the strongest connections social media, destination marketing, tourism, communication, DMOs and destination management. The focus of this term is closer to the issue of Marketing or Communication than a strategic process for the entire DMO (fig. 3B).

A survey was also carried out using the B-on website (https://eds.a.ebscohost.com/) with the following terms “social media mining tourism AND TripAdvisor reviews AND TripAdvisor AND sentiment analysis” and the following:

Journals: - tourism management perspective - international journal of tourism research - international journal of information management- tourism review - Cornell hospitality quarterly - worldwide hospitality & tourism themes - online information review - journal of hospitality & tourism technology - journal of travel & tourism marketing - journal of hospitality marketing & management - international journal of contemporary hospitality management - journal of travel research - information technology & tourism - sustainability

Subjects: - Tourist attractions: - social networks: - hospitality - data analysis - text mining - natural language processing - opinion mining - consumers' reviews – TripAdvisor - strategic planning - content analysis - tourism—management - hospitality industry - data mining - online reviews - big data - sentiment analysis - tourism - social media

The search returned 87 articles that were also analyzed via VosViewer.

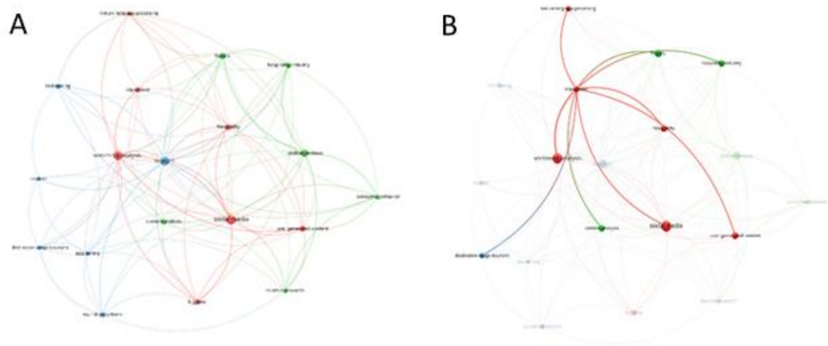

Figure 4 - A – social media mining tourism AND TripAdvisor reviews AND TripAdvisor AND sentiment analysis B – TripAdvisor and links

The figure above (fig.4A) shows the 3 clusters of keywords. It is important to note that the term “attraction” does not appear and only the terms “hospitality” or “hotels” are highlighted. This means that maybe there is a lack of the use of “attraction” as an object of research.

The keyword "TripAdvisor" has participants in its cluster the terms "sentiment analysis", "social media", "natural language processing", "user-generated content" and "hospitality". In a nearby cluster, we have hotels and the hospitality industry, reinforcing the issue of the focus of most papers on hotels and not on tourist attractions (fig.4B).

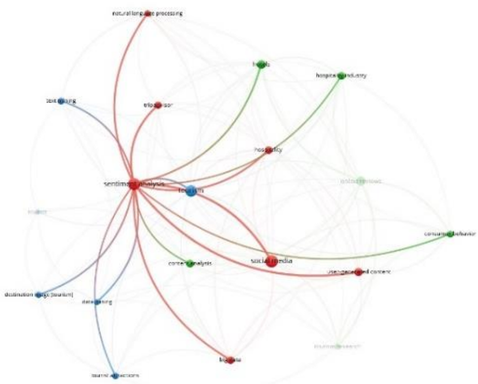

Figure 5 – Sentiment Analysis and links

Figure 5 shows the cluster of which “sentiment analysis” is part. Despite not being part of the initial search terms in B-on, it appears prominently, which means the term is frequently used in the keywords of the papers.

Another B-on survey was conducted with the bellow-designated parameters and returned only five texts - none with the effective use of negative reviews as an information source. The reason for choosing negative online reviews will be discussed later in this work. This search returned six papers.

TS=(social media mining AND TripAdvisor AND negative reviews)**Tempo stipulate:** 2008-2020. **Índices:** SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC.

Finally, we used PRISMA - a systematic review based on the "preferred reporting items for systematic reviews and meta-analyses" in an effort to systematically evaluate the quality and quantity of tourism research.

All findings were exported to the RefWorks reference management software for more analysis. After the removal of duplicate references, the remaining 594 records were analyzed according to the literature selection criteria. To continue the analysis - the abstracts were read first, and the full text of the articles was evaluated when further analysis was required. In addition, the studieswere discarded due to the methodological analysis technique used, as well as articles that did not have a sufficient focus on tourism were similarly eliminated. The screening process resulted in 128 records.The texts were carefully reviewed for eligibility in the final analysis. Only 20 studies were identified as eligible.

Figure 6 describes the number of studies analyzed and excluded at different stages of the literature review. The flow chart of the reports was in accordance with the PRISMA statement with some adjustments (Moher, Liberatti,Telzlaff & Altman, 2009).

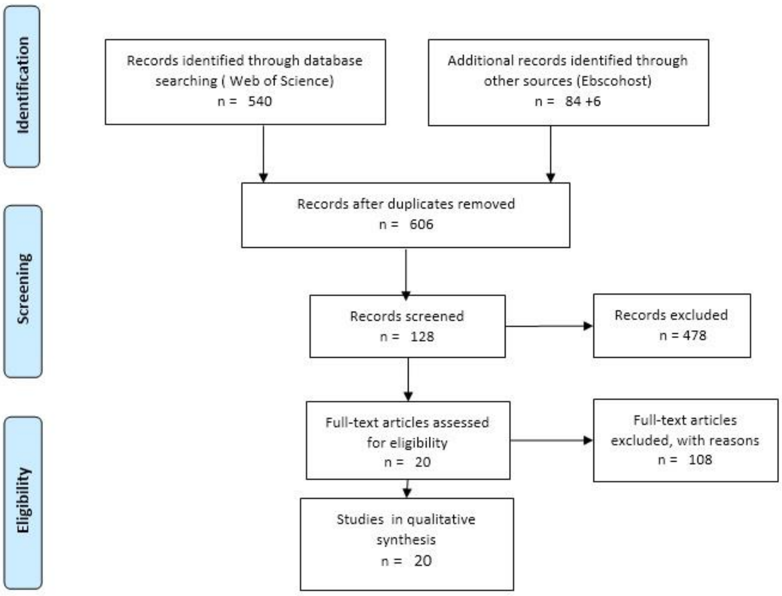

Figure 6 – PRISMA – Adapted from Moher et al. (2009)

From all the research carried out, we selected texts and works that could adhere more to the theme of the work and stipulated 20 of them to highlight, based on keywords and findings – Amadio and Procaccino (2016); Berezina, Bilgihan, Cobanoglu and Okumus (2015); Buhalis and Foerste (2015); Chang, Ku and Chen (2019); Dai, Sutinen and Kakkonen (2010); Dey, Haque, Khurdiya and Shroff (2011) ; Dickinger and Mazanec (2015); Franzoni and Bonera (2019); Gémar and Jiménez-Quintero(2014); Hu, Zhang, Gao, and Bose (2019); Leung, Law,van Hoof and Buhalis (2013); Moro, Rita and Coelho (2017); Marine-Roig and Clavé(2015); Schuckert,Liu and Law (2015); Sheehan,Vargas-Sánchez, Presenza and Abbate (2016); Taecharungroj and Mathayomchan (2019); Thomaz Biz, Bettoni and Pavan, (2015); Thomaz, Biz, Bettoni, Mendes-

Filho and Buhalis (2017) ; Tsao, Chen, Lin and Ma (2019); Vecchio, Mele, Ndou and Secundo (2018). All the findings and keywords of these texts are available in Appendix I.

Now we can discuss the research gap - a research question or problem which has not been answered appropriately or at all in a given field of study. A research gap also implies a lack of empirical studies, either from a certain theoretical perspective and or methodological approach.

Based on the maps generated by Vosviewer, in all the works read, it can be said that **USING THE TRIPADVISOR TO MONITOR AND GENERATE INTELLIGENCE FOR A DMO - focus Negative Reviews** is an unexplored topic and with the right tools, we believe it can be effective, as we will show in the evolution of this work. This is a topic that may fill the study gap. This article aims to contribute to the discussion on DMOs, suggesting a shift from the prevailing view (as oriented towards marketing) to a new one, based on the interpretation of the DMO as an intelligent agent.

To continue the literature review, we define important terms related to our theme based on concepts extracted from the bibliographic researches carried out: Web and Tourism, User Generated Content , e-WOM , Consumer Review Website (CRW), TripAdvisor (TA), Monitoring Social Media in Tourism, Negative Online Reviews, Data Mining, Natural Language Processing, Sentiment Analysis, Social Media Mining (SMM), Destination Management Organization (DMO) and Social Media Intelligence.

2.1 - WEB 2.0 , TOURISM AND UGC/ e-WOM

Web 2.0 has enabled customers to have a stronger capacity and impact on “information production and distribution” (Xiang & Gretzel, 2010) than companies provide, since all can be transformed into creators and/or recipients of content (Buhalis & Law, 2008; Xiang & Gretzel, 2010).

The Internet, and in particular Web 2.0, provides consumers with a new communication platform similar to word of mouth that also empowers consumers (Pan, MacLaurin, & Crotts, 2007). The traditional WOM (word of mouth) has been reinforced with the electronic WOM (e-WOM). Verhagen, Nauta and Feldberg (2013) defined e-WOM as any positive or negative comments made by potential, current or former customers regarding a brand, product or service that is available to other customers and/or organizations through the Internet. Mendes, Matos & Valle, 2012; Rezabakhsh, Bornemann, Hansen and Schrader (2006) state that eWOM (electronic word-of-mouth, i.e. analyses, evaluations, online recommendations) is increasingly being used by consumers to share their experiences about a product or service.

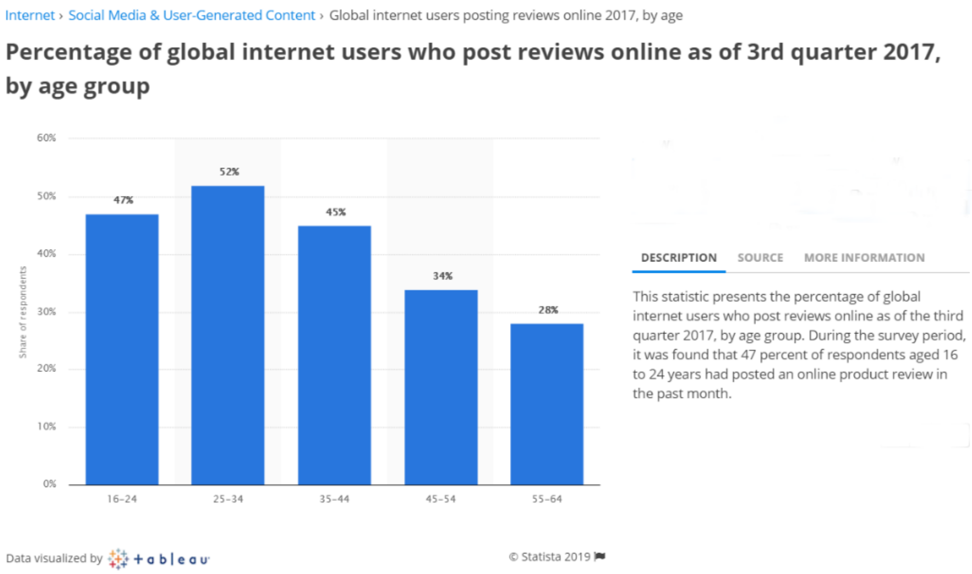

Figure 7- Percentage of global internet user who posts reviews online – SOURCE – STATISTA 2017


Internet › Social Media & User-Generated Content › Trust in online customer reviews 2014-2017

Do you trust online customer reviews as much as personal recommendations?

	2014	2015	2016	2017
Yes, always	-	8%	18%	19%
Yes, if I believe the reviews are authentic	22%	31%	27%	25%
Yes, for some types of businesses, no for others	34%	22%	19%	20%
Yes, if there are multiple customer reviews to read	26%	19%	20%	20%
No, I am often skeptical about online reviews	-	-	12%	13%
No, I don't trust reviews at all	17%	20%	4%	3%

Showing entries 1 to 6 (6 entries in total)

DESCRIPTION SOURCE MORE INFORMATION

This statistic compares consumers trust in online reviews compared to personal recommendations. During the October 2017 survey it was found that 25 percent of respondents trusted online reviews if they believed them to be authentic.

Data visualized by +ableau © Statista 2019

Table 1 – Trust online reviews – Source: Statista 2017

Besides the e-WOM, there are other similar terms used, such as digital word of mouth, user-created content and user- generated content (UGC). Although some authors (Thao, & Shurong, 2020) make a slight distinction between e-WOM and UGC, for the purposes of this work we will assume as similar.

Web 2.0 and UGC have been, and probably will be, increasingly changing the way people search, find, read, collect, share, develop, and consume information (Sigala, 2014). Interesting is the definition of UGC by Blackshaw and Nazzaro (2006) - UGC is "a mixture of facts and opinion, impression, and feeling, founded and unfounded, experiences, and even rumors".

The content generated and shared among users ends up arousing desires, expectations, and perceptions in other users. They also end up influencing the decision-making process for the purchase of products, since they allow potential consumers to gain the desired knowledge from different sources, as well as benefit from the experiences of obtaining information and advantages for their own experience (Yoo & Gretzel, 2008b, Roque, Fernandes, & Raposo, 2012).

These analyses are seen as reducing risk for buyers (Yoo & Gretzel, 2008a). Buyers want to read about previous buyers' experiences with a product or a service without having to try it themselves, and by seeing them as references, consumer actions are influenced.

Logically, the advent of Web 2.0 has caused significant changes and transformations in Tourism, especially with regard to the production of content by users and the sharing of information and content between users and consumers, gaining popularity among online activities of travelers (Xiang & Gretzel, 2010). As a result, social networks are becoming increasingly important in travel planning, mainly due to their role as vital sources of information that provide access to other travelers' experiences (Chung & Buhalis, 2008).

Because it is "vital" for tourists, Goldenberg, Libai, and Muller(2001) stress that other consumers' word-of-mouth influence the decision-making processes of other consumers, in this case tourists.

One of the main problems tourists face when deciding to travel is assessing the quality of places they have never seen before (Kim et al., 2007). Since they are considered "intangible products", vacation travel can be considered complex due to the experimental nature of the vacation product, involving risks and therefore requiring extensive information research (Sirakaya & Woodside, 2005). According to Chung and Buhalis (2008), consumers look for information from a variety of sources, with social networks being the most widely used to minimize the risks of making wrong travel decisions (Leung, Schuskert ,& Yeung, 2013).

In seeking information, consumers rely on other travelers' experiences as a means to increase reliability and reduce uncertainty and increase the usefulness of exchange (Litvin, Goldsmith & Pan, 2008; Yoo, Lee, & Gretzel, 2007; Fotis, Buhalis & Rossides, 2012). Thus, data from other consumers who have already experienced the specific product, and are willing to provide information, is considered the most preferred and influential source in the context of travel-related decision-making (Crotts, 1999).Gretzel and Yoo (2008) point out, although the analyses provided by other travelers are often considered by readers to be more up-to-date, enjoyable and reliable than the information provided by travel service providers.

Therefore, the information is more critical and the need for information research becomes even more intense (Bronner & de Hoog, 2013; Mudambi & Schuff, 2010; Pan, MacLaurin, & Crotts, 2007).

The technological advance of information and communication, with the consequent changes in values and lifestyles, has led to the emergence of new customers for tourism products. The new customers for tourism products are those who are more informed, more independent, more individualistic, and more involved (Gretzel, Fesenmaier & O'Leary -2006; Buhalis, Costa, & Ford, 2006). Thus, it may be crucial to identify the most relevant new sources of information to consider, both for monitoring and for creating advertising content (Munar & Jacobsen, 2014). Thus, with the advent of "Travel 2.0", the tourism industry has become an information intensive industry, as it allows travelers to easily access information through the Internet (Qu & Lee, 2011).

Most studies on social networks in tourism are focused on the consumer, suggesting that the use of social networks by travelers or the influence of social networks on tourist behavior has been studied frequently (Denizci, Kucukusta, & Liu, 2016).

Litvin et al. (2008) concluded that if a tourist is recognized by his online peers as experienced and reliable, his opinions can have a significant influence on purchasing decisions made by travelers to other tourist destinations. Thus, these comments become valuable commercial assets, i.e. e-WOM, are valued by their peers.

In addition, Zhu and Zhang (2006) stressed that comments generated by online users are useful for both consumers and online retailers. Similarly, Dellarocas (2003) indicated that online word-of-mouth could have important implications for managers in terms of brand building, product development, and quality assurance.

However, in the context of tourism, the impact of online consumer criticism on the performance of tourism businesses remains largely unknown. Vermeulen and Seegers (2009) conducted an experimental study with 168 participants to determine the impact of online criticism on travelers' attitudes towards hotels and revealed that exposure to online criticism has raised awareness in hotels and that positive criticism has improved travelers' attitudes towards hotels. Despite all the influence of the content review on tourists, how can one infer the importance of commenting on tourist information and how the industry has been handling this wealth of data?

2.2 - CONSUMER REVIEW WEBSITE (CRW) AND TRIPADVISOR

Consumer review websites (CRW) are social media applications that enable users to upload product or service-related reviews and ratings. CRW can offer a wide range of features, such as uploading comments or even pictures.

Generally, exist two main categories of CRW sites: sites with open systems, where consumers can go onto the website and post a review, and closed systems, where only a confirmed buyer of the product or service can submit a review.

There are several sites linked to CRW, but some of the best known are linked to tourism and hospitality. Consumer reviews were found to be used throughout the stages of the travel planning process, increasing travelers' confidence about decision-making, reducing risk, and assisting trip planning, mainly in accommodation selection (Gretzel&Yoo,2008). CRW is not the only source of information in the pre-trip stage, but also in the post-trip stage focusing on the motives that drive travelers to post reviews (Papathanassis & Knolle,2011).

TripAdvisor (TA) is one of the most recognizable consumer-generated content sites or consumer review websites (CRW). TA is a CRW dedicated to tourism, and hospitality social media have grown tremendously in the last decade, with the advent of interactive Web 2.0 technologies (Hays, Page, & Buhalis 2013; Munar, Gyimóthy & Cai 2013; Xiang & Gretzel 2010).

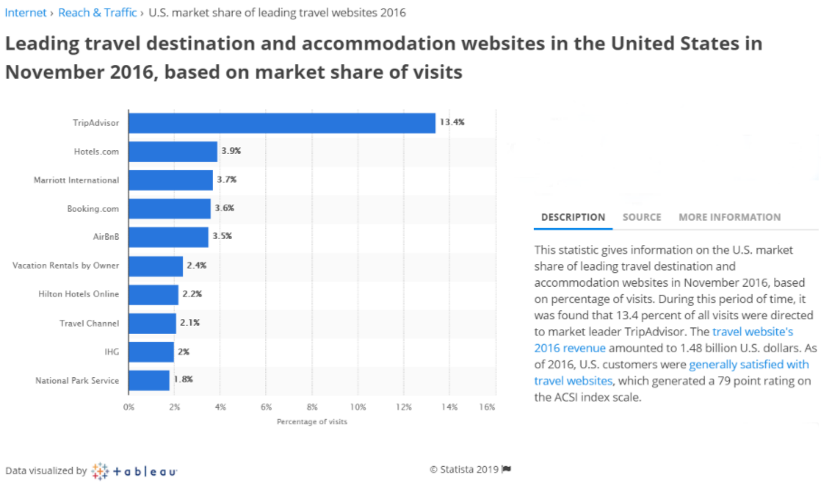

Figure 8 – Leading Travel destination websites USA 2016 – Source: Statista

As seen in figure 11, Tripadvisor is a present clear leader (Munar et al. 2013; O'Connor 2010). The interactive capacity of the platform enables users from virtually all over the globe to engage in a range of communicative activities such as information searches, rating products, and services, and initiating and participating in tourism-related discussions (Litvin et al.,2008; O'Connor 2010).

Figure 9- Number of user reviews and opinions on TA 2014-2018– Source: Statista

In August 2018, the number of unique visitors was 455 million, 702 million reviews and opinions, and the number of attractions was nearly 1 million (Statista 2017). With this scale of user's reviews and opinions, Tripadvisor offered the most comprehensive platform for investigating tourism social media communities.

According to Munar et al. (2013), TA aims to promote social interaction in the tourism industry, encouraging users to share their experiences with the various hospitality fields. This network has created a voting utility system regarding opinions already expressed. It can be seen as merit- and skill-based recognition of users and reference to other travelers concerning the expressed opinions.

Tripadvisor's architecture allows two primary forms of interactive user platforms, namely, reviews and discussion forums. Based on the information given by TA, new topics are posted every day to the TripAdvisor forums, and, further, more than 90 percent of questions posted to TripAdvisor's English-speaking forums are replied to by other travelers within 24 hours. TA's fact sheet explains that forums enable members to ask for and share their opinions, advice, and experiences in interactive discussions with the community.

In particular, as seen in literature review, there is a lack of research in tourism studies on how DMOs or attractions can use content reviews with a focus on improving management.

The interaction between users on TripAdvisor generates an extensive database of attractions and tourist destinations, turning the site into a reference for information on the web. User reviews and opinions on TripAdvisor not only have quantitative elements (ratings, number of comments, etc.), but also qualitative information, allowing a DMO, for example, to identify what tourists are saying about the attractions and, in the event of receiving a negative rating, to correct their flaws more quickly and efficiently. For Torres (2013), if the reviews of these customers are processed efficiently, they can allow companies to improve the quality of their products and services, turning this information into a great asset for those who manage it.

Extracting data from TripAdvisor is crucial, as it helps to capture constantly expanding user-generated content, which can be used by the travel and hospitality industry. These contents, once processed, can serve to understand product/service evaluations, feedback, complaints, and be a source for brand monitoring, brand analysis, competition analysis, trend observation, and much more.

The Tourist destination, with capital as a critical protagonist of the planned experience, lived and remembered, is not considered by TripAdvisor as a "valuable" subject. While it is true that we can review the attractions (“Torre de Belém”, “Mosteiro dos Jerónimos”, “Praça do Comércio”, “Tram 28” and “Castelo de S. Jorge”), as tourists cannot give a rating to Lisbon or Portugal.

Therefore, it is pertinent to know the perception that tourists have of each of these attractions. And while these agencies use various studies and analyses to determine the positioning of their tourism brand, the fact remains that there are still few cases of DMOs using online reputation tools for what is said in forums, social networks, and CRW.

2.3 -MONITORING SOCIAL MEDIA IN TOURISM

The significant adoption of social media by individuals and organizations has resulted in a wealth of data and information that, when treated and analyzed, offers opportunities to be transformed into knowledge for organizations. That has resulted in the need for organizations to monitor the content, as well as the performance and results of strategies adopted in social media (He, Zha, & Li 2013).

According to Zeng, Li, and Duan (2012), monitoring in social media is a strategy that aims to extract relevant information from unstructured content, requiring continuous monitoring and refinement of information to achieve good results.

Paine (2011) points out that organizations invest resources in communicating and relating to consumers and states that social media monitoring provides opportunities to listen to consumers and, consequently, to better communicate with them. Silva (2012) adds that monitoring on social media is the act of transforming data into knowledge.

He et al. (2013) also claim that content mining in social media has become a critical need for organizations and aims for further analysis and also supports and facilitates decision- making.

Concerning tourism, social media monitoring offers opportunities to identify tourists' opinions and feelings about destinations, services, and tourist attractions, monitor events, and everyday situations to identify weaknesses, strengths, opportunities and critical situations using this information to help in the management of the tourist destination, to define priorities, direct investments, create public policies, training courses, redefine marketing strategies and promote tourism, positioning, among others. All of this information collected and monitored in social media can be used strategically to achieve the stated objectives, identify what is working and define the next actions of marketing, planning, and management of the tourist destination.

Although many tourism organizations already recognize the importance of using social media as a source of information, communication, and interaction, it is necessary to establish the needs and methods of use (Munar, 2012).

Munar (2012) also presents that managers can monitor trends and perceptions of tourists about a particular tourist destination, identify the increase or decrease of their interest in relation to specific tourist attractions or types of tourism, identify positive reports and stories published by users about the destination to use them in advertising campaigns and marketing strategies and even identify and examine negative content for destination quality management.

Besides, identifying negative and positive opinions in social media can help organizations in their strategies for change and strategy creation, users and customers to decide on the purchase of a product, service or destination for their vacations and government organizations to improve services, launch campaigns, among others (Ku & Chen, 2007).

However, through the use of automated computing techniques for data mining, it is possible to develop applications to collect and extract useful information and knowledge from many textual documents - such as in social media - in order to subsidize the process decision of an organization (Silva, 2012; He et al., 2013).

Among the techniques that have been investigated to extract information and knowledge in SM, data mining techniques and content analysis have been highlighted by providing resources to analyze large complex and dynamic textual data sets, characteristics of social media (Santos, 2009).

A point to highligh is that is many organizations are not familiar with using and analyzing social media to gain competitive intelligence and further claim that organizations do not have enough knowledge of the data mining process in social media (Dai et al., 2011).

2.4-SOCIAL MEDIA – CRW - TRIPADVISOR AS A SOURCE OF INFORMATION

The broad adoption of social media by users has generated an exponential increase in data and content that offer, although not in full, opportunities to be treated and transformed into information and knowledge for organizations.

Schukert, Liu, and Law (2015) highlight the ease of use of online analysis by customers and thus generate a considerable mass of data. With websites and web 2.0, writing an online review is a more common way to comment on your satisfaction and rate a site or service. That way, expressing feelings or filing a complaint is just a click away.

Young (2016) clarifies that “customer feedback data is simply too valuable and powerful to ignore”.

Overall, Young (2016) says that once a company begins to leverage this information and use it to make crucial decisions, it can improve its performance, from customer retention to improved products and services.

For businesses, using information from an online customer review can provide:

- Improvement of the decision-making – Several authors state that customer feedback data provides company managers with a wealth of pertinent information that can improve decision-making (Chen at al.,2012; Devika & Surendran, 2013; Young,2016). For them, extracting data from shared information on SM has become an indispensable tool that helps managers of companies and organizations during the decision-making process.

- Improvement of the company's staff - is the point that emphasizes Pantelidis (2010).

- Support for strategic planning – Moe and Schweidel (2014) also see opportunity in using online opinions. For them, with this information, organizations can better discover valuable ideas hidden in social networks, and use them to nurture their strategy.

Young (2016) goes further and states that the analysis of this information can even highlight and show the possibility of opportunities and growth in some areas.

As noted, online reviews/feedback provide a real and fast channel for assessing additional information on service delivery, quality, and customer demand, specifically from negative reviews with low ratings and grades, as they are more likely to reflect real problems.

Given the fact that all competing organizations have equal chances of accessing the same information from their external environment, the company's ability to exploit intelligence seems to be more important than finding and collecting it (Rollins, Bellenger, & Jonhston, 2011).

Although the information on TripAdvisor is available for everyone to read, it becomes impossible for a single person to read all of them. It is a big data problem, which can only be solved with the use of computational techniques, in addition to the use of data mining, text mining and sentiment analysis tools.

2.5 - WHY NEGATIVE ONLINE REVIEWS?

Bill Gates said: “Your most unhappy customers are your greatest source of learning” (Gates & Hemingway, 1999).

If we follow this statement and choose the negative online reviews of Lisbon attractions for the data source and intelligence generation for DMO Visit Lisboa?

Faed and Forbes (2010) state there is a growing recognition that negative customer feedback can be regarded as a strategic tool in companies.

Patel (2018) affirms that negative feedback is an opportunity to improve. It’s difficult to get customers to complete surveys. That means that getting voluntary feedback is invaluable. It can help you uncover and resolve key consumer pain points. This, in turn, can help a business grow. If you dig deeper into those reviews, you will be able to find actual business ideas. Customer complaints are a powerful tool to uncover internal problems on time and fix them before they go out of control.

Krause (2018) corroborates this view and adds: “If you want to improve customer service, monitoring negative sentiment comments can inform the right solution to solve the problem.”

Gin (2016) points out that there are several actions that every organization can take to extract as much value as possible from negative feedback:

1. Create competitive intelligence - when reviewing evaluations, especially negative ones, the organization can get a sense of how its products and services compare to competitors and see if customers have similar complaints. It is important to also make a careful analysis of your competitors, as this can also help you determine your weaknesses and opportunities to win over your customers.

2 - Benchmark performance in relation to the sector - Direct customer feedback can be a great way to shape the evolution of the products and services that the organization provides. And while big changes in products and services can be expensive, small changes in the service process can make a big difference in defining customer expectations and increasing satisfaction.

3-Making customer service aware of frequent customer complaints and how to deal with them can be fundamental to retain customers. The formalization of this process, creating a simple cycle of training and feedback for the team, can have a significant impact on the image and the result of the business.

4- Look for improvements. This is true for most data-driven companies, but as changes are made to your training, products, services and marketing, carefully monitor how it affects your comments.

2.6 - DATA MINING CONCEPTS

Data mining is a multidisciplinary subject and, therefore, the definitions about the term vary according to the field of action of the authors being: statistics, machine learning, artificial intelligence, pattern recognition, database and storage systems data retrieval, information retrieval, visualization, high-performance computing algorithms and several other application domains (Han, Kambe, & Pei, 2012).

Barbier and Liu (2011) argue that the primary goal of data mining is to find new information in a data set that is hidden or latent and that data mining can help individuals and organizations better understand large data sets.

Furthermore, it is essential to emphasize that due to the emergence and advances of new information and communication technologies (ICT) such as the Internet and social media, data are currently represented in different types and formats such as unstructured, spatial and temporal, multimedia, text, web, social media, among others. Thus, data mining technologies were adapted to exploit this data (Rezende, 2003). That has resulted in several other new processes and variations of data mining, such as Text Mining, Web Mining, Web Content Mining, Social Media Mining, Opinion Mining, Sentiment Mining, among others.

For the development of this dissertation, the focus will be Web Content Mining, Opinion Mining, Sentiment Mining, and Social Media Mining.

Han et. al (2012) define Web Content Mining as “the process of extracting knowledge of the content of documents and their metadata”.

This approach mainly covers the extraction of knowledge of the content of textual documents (text pages, HTML or other formats, emails, mailing lists, user groups, blogs, etc.) as well as multimedia data mining on the Web (images, videos, and audios) using or not associated textual data (Santos 2009; Han et al., 2012).

According to Ahmad (2013), the data available on the web are classified as: structured, semi-structured, and unstructured. In TripAdvisor reviews, we consider UNSTRUCTURED DATA. Unstructured data is the data in the form of standard text documents and is related to text mining, natural language processing, machine learning, and web-question-answering.

The growing use of the Internet and social media has given researchers a new and growing source of data on human behavior, and every interaction of users like the use of search engines, media, and social networks generate data that allow them to document and analyze the online behavior of users. Therefore, the “social web” and content mining in social media began to attract and receive the attention of several researchers (Han et al., 2012; Kapland & Haelenlein, 2010).

Although most of this data remains inaccessible to researchers, some services such as Google, Twitter, Facebook, and others allow access to behavioral data through application programming interfaces (APIs) and other data such as publications or comments on commercial platforms are publicly available and can be retrieved by automated scripts (Jungheer & Jurgens, 2013).

Malik and Rizvi (2011) state that there are several ways to extract digital information from social web sites, but these authors highlight three outstanding ways: web usage mining, semantic annotation, and web scraping.

Devika and Surendran (2013) simply define web scraping as “web data extraction is the reverse process of page generation”. For Mitchell (2015), Web Scraping consists in developing and running a script to efficiently download web pages’ contents and to extract the needed information.

The main advantage of this method is that it allows unstructured web data, usually in HTML format, to be transformed into a well-structured database that can be parsed (Vargiu & Urru, 2012).

Marres and Weltevrede (2013) detail that the scraping process consists of a series of steps in which the formatted data is extracted from an “informative mess.” In their words, “scraping is building a chain from the relatively unformed mass of online data to formatted information, and along that chain, relatively raw textual data is progressively removed from its useless elements and formatted to produce a dataset usable and well-ordered”.

But, to extract information and make the analysis whether the opinion is positive, negative, or neutral, it is necessary to use Natural Language Processing (NLP) and Opinion Mining (OM) or Sentiment Analysis (SA).

Kahn, Baharudin, Khan, and Ullah(2014) define Natural Language Processing(NPL) as a set of computational techniques for parsing natural language texts that allow computers to understand human language. From this, it is only natural that NLP has been used for social media analysis because it allows computers to process data, like unstructured texts collected from social media applications. Syeda, Shiraz, Naqvi, Parkinson, and Bamford (2017) state that NLP is characterized as a challenging problem in the area of extensive social data. Finally, while language is one of the most natural things for humans to learn, language ambiguity is what makes natural language processing a difficult problem for computers to master.

Opinion Mining (OM) or Sentiment Analysis (SA) can be the core technology behind many social media monitoring systems and trend analysis applications.

Ohbe, Ozono, and Shintani (2017) describe sentiment analysis as a class of NPL, text analysis, and statistics. The purpose of this analysis is to find the feeling of the text by classifying it as positive, negative, or neutral. In social media, sentiment analysis has several uses. For example, this analysis can be applied to identify the feelings of tourists about a specialized service, or even as it was met by the marketing and customer service department, which results in finding out if consumers are satisfied or dissatisfied with a product (Povoda, Burget, Dutta & Sengar, 2017).

Khan et al. (2014) reaffirm that sentiment analysis is not a perfect science, mainly when applied to the unstructured texts that predominate in social networks. Human language is complex, so teaching a machine to detect all the variations of language and communication mentioned above is a complicated process and is, even more, to be performed automatically. It is difficult to determine how it will evolve in the future, although there is a general belief that this analysis, using machine learning, artificial intelligence, and neural networks, goes beyond the classification of texts on a positive and negative one-dimensional scale.

In recent years, the list of challenges related to the analysis of feelings has increased (subjectivity rating, a summary of opinions, recovery of opinions, etc.). Therefore, the implementation of sentiment analysis techniques to extract sources of opinion is crucial to understand the failures and assets of the tourist service. Through web platforms like TripAdvisor, tourists can openly describe their experiences and thus affect a company's viability. Consequently, the implementation of sentiment analysis techniques to extract sources of opinion is crucial to understand the failures and assets of the tourist service.

2.7- SOCIAL MEDIA MINING (SMM)

WEB 2.0 and social media make available an avalanche of data for companies and organizations. What may be there is a lack of ability to extract them, but those who have them are collecting data at a rate never found before through web sources.

Hebert, Anderson, Olinsky, and Hardin (2014) define Social Media Mining SMM as “the process of extracting useful or actionable knowledge from this user-generated data on a large scale in the field of social media”. Thus, for them SMM “is the process of representing, analyzing and extracting actionable patterns from social media data”. These authors claim that the media mining process is at the point of convergence between social media and big data.

Crooks, Croitoru, Stefanidi, and Radzikowski (2012) show that Mining in Social Media consists of three general steps: i) extraction of data from social media providers and servers through application programming interfaces (APIs); and iii) data analysis to extract information of interest.

Although many tools and sites allow the execution of SMM phases, in most cases, the demand for strategic information has a high degree of specificity that the generalization of the available standards is not enough, requiring the development of customized systems (Crooks et al., 2012). To fulfilment this demand, SMM’s techniques and tools for collecting, sharing, investigating and viewing social media data have been widely explored and developed (Tang & Yang, 2012).

Given the enormous volume and dynamic nature of content in continuously generated social media, automatically collecting and identifying emerging themes of interest amid the vibration of constant conversations and interactions between users is pointed out by several authors as one of the main challenges in the SMM process.

Paine (2011) also emphasizes content collection as the most challenging phase of the process since SMM's services and software do not guarantee data integrity.

The main characteristics of social media data are large volume, dynamic and noisy (Zeng et al., 2012). This type of data is inherently noisy and the use of emoticons, of multiple meaning, without a strong semantic connection, typos, spelling errors and abbreviations, as well as recent slang are the most common noises and often represent unwanted data items. As a result, Zeng el al.(2012) claim that this content requires continuous monitoring and refinement of information to achieve good results.

These types of noise often cause the algorithms to lose patterns in the data, and may not reflect the accuracy of the extracted data.

It is worth mentioning that obtaining relevant content and information on social media is a major research challenge (Dai et al.,2011) and requires filtering, semantic grouping of content, markup and content mining techniques (Dai et al., 2011). In the same way, Coutinho, Lang, & Mitschang (2013) highlight the challenge of monitored terms presenting multiple meanings, which makes a challenge to recover unwanted messages for analysis, and essential terms for the analysis may be lost to the large data set.

So, the task of finding relevant information on the Internet is tough and often becomes a frustrating experience due to the heterogeneity and lack of data structure of the Web and social media.

Zafarani, Abbasi, and Liu (2014) warn differently from other authors about the system that will perform the SMM - it is called the "noise removal fallacy." For a successful data-mining exercise is vital to do extensive data preprocessing and noise removal. At this point, they warn that blind noise removal does not occur as it can make the problem indicated in the big data paradox worse, because removal can also eliminate valuable information.

2.8 - DESTINATION MANAGEMENT ORGANIZATION

The World Tourism Organization (UNWTO 2016) defines DMO *as the leading organizational entity which may encompass the various authorities, stakeholders, and professionals and facilitates tourism sector partnerships towards a collective destination vision.*

Destination Management Organizations (DMO) are responsible for the planning, management, promotion, and development of the tourist activity in their respective destinations. For them, is fundamental that know about the environment, clients, and respective reviews about attractions at the destination.

A study conducted by Parra-López et al. (2011) suggests that it is of great importance for a DMO to pay close attention to their clients' needs for information in the social media so that they can respond actively and favor the perception of functional benefits.

It is also crucial for DMO the recognition tourists' images of a tourist destination identify their strengths and weaknesses (Chen & Uysal, 2002) and to position them efficiently in the market (Pike & Ryan, 2004).

Through surveys or by monitoring and researching a CRW social media TripAdvisor, as proposed in this thesis, a DMO can know what customers are thinking and pointing out tourist destinations and even monitor the main competitors of tourist destinations.

Research reveals that DMOs use social media primarily in the areas of marketing, management, communication, and product distribution (Leung, Law, van Hoof, & Buhalis, 2013; Leung et al., 2013). Indeed, several studies focusing on different tourism contexts (e.g., hotels, airlines, travel agencies) have found that social media is actively used as a marketing tool in the industry (Denizci Guillet et al., 2016; Leung et al., 2013).

However, many organizations simply transfer their existing marketing activities onto social media platforms, rather than exploiting the transformational potential of social media. Denizci Guillet et al. (2016) found that most of them utilized social media mainly for disseminating and receiving information. A similar pattern of results was found for travel agencies, airlines (Leung et al., 2013) and DMOs (Hays et al., 2013) indicating that tourism organizations, regardless of their type, are still in the experimental stage of utilizing social media as a marketing tool.

In this way, the DMOs use the information for digital marketing and not to generate information and, consequently, strategic intelligence. And in this regard the proposal of this thesis - to generate actionable information from social media data.

3. METHODOLOGY

This topic highlights the methodology on which this study was based to collect, organize, and analyze data, as well as the presentation of research results and the justification for this approach. In line with this, the chapter points out how the research approaches and methods chosen were used to help solve research questions and answer them. So, the research strategy is an information-gathering approach to accurately answer the research question.

First, we will be following the principle of that phrase - according to IBM Big Data expert Jeff Jonas: you need to let the data “talk to you” (Mayer-Schönberger & Cukier, 2017).

Data is an important part of the research to do the analysis. Primary and secondary data were of equal importance for the study. According to Social Media Research Group (2016), mining social media can provide qualitative (sentiment, opinion) data, and quantitive (frequency of words, number of reviews) data. D'Orazio (2013) goes further and highlights that social media can be considered "qualitative data on a quantitative scale".

As seen, our proposal is, based on content reviews on TripAdvisor, to generate intelligence for a DMO. Pan and Li (2011) reinforce our proposal by making two important statements - the review of customer content that is available on travel blogs can be richer in content and is more detailed than based on questionnaire surveys. For them, moreover, using interview methods may not capture emotional experiences correctly.

The capacity to discover information hidden in these huge data and act on that knowledge is increasingly important nowadays (Kantardzic, 2011). Consequently, new forms to collect and analyze data should be considered. Text mining is an example and will be used in this work.

Therefore, the study will be based on a combination of quantitative and qualitative analytical research methods, focused on textual analysis of TripAdvisor content where the UGC is displayed, using a combination of automatic coding methods that support data mining and meet the objectives of the research.

Because it is a free program and with a large number of packages available, both for mining and sentiment analysis, R will be used.

The research will be based on the use of the R language to monitor the site Tripadvisor - Lisbon attractions.

In Chapter 5 we will detail the method and the code will be available in the Appendix II.

4-SOCIAL MEDIA INTELLIGENCE AND PROPOSAL PROCESS

The term Social Media Intelligence was coined by Omand, Bartlell, and Miller at Demos, in London 2012. They created the term for the Centre of Analysis of Social Media (CASM). They defined Social Media Intelligence (SOCMINT) as “intelligence derived from social media”.

For Norton-Taylor (2012) Social Media Intelligence (SOCMINT) is the process of identification, validation, collection, and analyzing data and information from social media using intrusive and non-intrusive methods, with the aim of developing products. The purpose of SOCMINT is to be able to reduce the “unknown” that comes within any decision- making.

Zeng et al. (2010) identified several key challenges currently face social media intelligence. First, social media intelligence research calls for highly integrated multidisciplinary research. Another issue is social media intelligence research requires well-articulated and clearly defined performance measures because much of it must be conducted in application settings with an aim to support decisions. Finally, from a pure modeling and decision-making perspective, social media intelligence represents a unique class of problems with the need for efficient data-driven, dynamic decision-making, uncertainty and subjective risk analysis, and modeling and optimization over large dynamic networks.

Ross, McGowan, and Styge (2012) pointed out that managers should consider SOCMINT activities as a formal regular business process to empowers themselves to anticipate and face future challenges, enhance their capability and ability to maintain a competitive edge over their competitors.

Chen et al. (2012) and He et al. (2013) agree and state that social media competitive intelligence can help organizations to realize the strengths and weaknesses of their products and services, enhance business effectiveness, and to improve customer satisfaction.

Starting from the data-information-intelligence-knowledge, the SOCMINT framework can be created. First, data with context equals information, information with meaning can be intelligence, and intelligence with experience generates knowledge. It reflects the qualitative changes from data to knowledge. Intelligence and/or knowledge are the basis for making decisions, and they must be a useful format to meet strategic needs for enterprises.

To acquire data from TripAdvisor, web scraping, and social media mining will be carried out. After extraction via R language, opinion mining and sentiment analysis will be performed. At that moment, we will have the information.

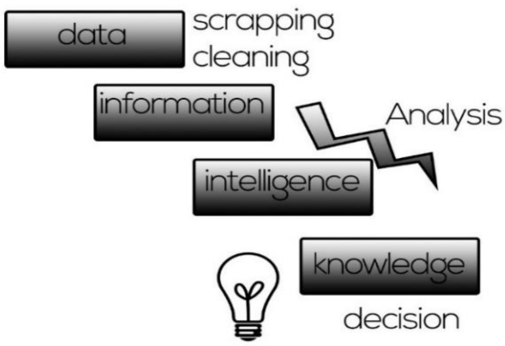

Figure 10 – Cycle Data – Information – Intelligence - Knowledge

For the transition from information to intelligence and knowledge, the presence of the intelligence analyst is still necessary. Analysis- the result of the action of the analyst – is an essential step, which includes analysis of collected data to identify patterns, relationships, or anomalies in it. It involves interpreting and translating the collected raw data into "actionable intelligence" (Miller, 2001). Several authors affirm that this is where "true" intelligence is created, that is converting information into usable intelligence on which strategic and tactical decisions may be made (Gilad, 2003; Kahaner, 1996; Prescott, Miller & Professionals,2001; Herring, 1999).

Although technology plays a vital role in SOCMINT, according to Bose (2008) non-computerized methods are still required to transform the data into actionable intelligence. Many analysis software packages can produce graphs and conceptual models, but in the end, human judgment is needed to put these analysis products in meaningful contexts and to create actionable intelligence. Wurman (2001) shows the importance of contextualizing information, by stating that a fact is only understood within the context of an idea and that due to its subjectivity, there is no way to achieve a precision absolute in terms of information. So, for now, only the human analyst can do this, despite all the advances in Artificial Intelligence.

Töllinen et al. (2012) find that traditional person-controlled research is required to create a deep understanding of social media data.

To generate the complete Social Media Intelligence Process, it is important to use some concepts from Competitive Intelligence.

As seen, to use and make sense of the information obtained from social media sites, companies need to analyze it and turn it into actionable insights, that is intelligence (Tej Adidam, Banerjee, & Shukla, 2012; Dishman & Calof, 2008). Tej Adidam et al. (2012) also observe that the analysis of information to convert it into the right type of intelligence is one of the most challenging and critical in this process. Dai et al. (2011) emphasize that to gain competitive power, to be effective, in the rapid analysis of social media content is extremely important.

To facilitate this process of identifying needs, Herring (1999) suggests the use of the KIT (Key Intelligence Topics) or Fundamental Topics of Intelligence process. This process is part of the first phase of the Intelligence cycle, Planning, being a facilitating guide in the first moment of implantation of CI in a company.

The process, through questionnaires and interviews, helps to identify and prioritize the intelligence needs of the organization's managers.

Thus, for the use of social networks as a source of Competitive Intelligence, the process can be started with the KIT.

The results of KIT interviews generate a focus on conducting the CI's operation in the organization, and allows those responsible to determine the resources needed to meet the company's fundamental intelligence needs.

Herring (1999) suggests framing KIT questionnaires into three functional categories: 1)Strategic decisions and actions; 2) Early Warning Topics; 3)Description of the main actors.

Adapting the KIT suggested by Herring (1999) “Decisions and strategic actions” we can have the following topics:- Which tourist attraction to monitor? What do they say about the attraction? What keywords represent those UGC?

The KIT functional category “Early Warning Topics” provides protection against surprises or threats, but can also generate opportunities. It also generates the possibility of making contingency plans, which can be implemented when the fears and concerns gathered in the questionnaires materialize.

The KIT functional category “Description of the Main Actors” reflects the managerial need to know the main players in the market (customers, suppliers, and mainly competitors), generating a frame of reference and profile for each of them.

As topics in the KIT “Description of the Main Actors”, supported by Herring's proposal (1999), the organization has the focus of study:

-Identify new customers, needs and interests and how competitors are serving through TripAdvisor UGC posts;

-Opinions, attitudes, and perceptions of customers and competitors about the value of the tourist attractions in Lisbon.

Cook and Cook (2000) proposed a model for the implementation of a Competitive Intelligence project, facilitated with the use of KITs in the formulation of questions. Based on this proposal, it is possible to integrate the question of information search in social media as a source of Competitive Intelligence - SOCMINT.

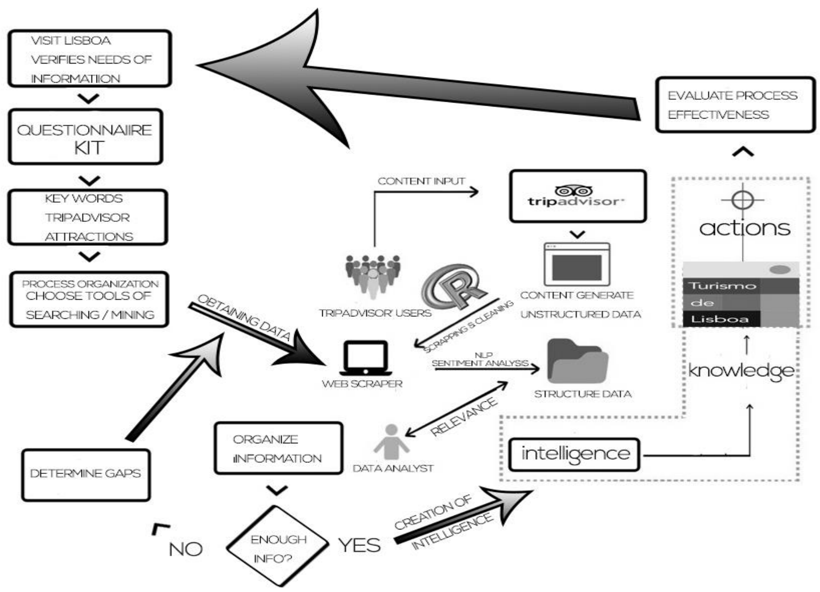

Figure 11 - SOCIAL MEDIA INTELLIGENCE CYCLE FOR VISIT LISBON-SOCMINT

5- MINING TRIPADVISOR

Using the SOCMINT framework applied to the present work, the steps proposed by him were followed.

To choose the scraping focus, we used KIT techniques. According to the results of the questionnaire and how the focus is on the DMO Visit Lisboa, we searched the content reviews on TripAdvisor related to Attractions. The choice of attractions was given by the number of ratings / TripAdvisor ranking/public and not private attraction. So, we scrapped the UGCs linked to the following tourist attractions in Lisboa: "TORRE DE BELÉM"- Belém Tower, "MOSTEIRO DOS JERÓNIMOS"-Jerónimos Monastery, "PRAÇA DO COMÉRCIO"- Commerce Square, "ELÉCTRICO 28"- Tram28, and "CASTELO DE SÃO JORGE"-S. George Castle.

It was made a comparison between Jun / Jul / Aug 18 and the entire review period on TripAdvisor – so it was chosen the form every period because of its broader scope.

This kind of semiautomated analysis, according to Költringer & Dickinger (2015), comes from the domains of data mining and natural language processing and includes four steps: (1) data preparation and preprocessing, (2) keyword analysis (meaningful words), (3) sentiment analysis, and (4) information enhancement.

For scrapping, mining, cleaning and analyzing the data from TripAdvisor we use "R Language" and its package .

All English language-based reviews collected from TripAdvisor were pre-processed using two basic procedures: tokenization and stop words removal. Tokenization is a form of lexical analysis whereby a stream of text is broken up into words, phrases, or other meaningful elements called tokens. In this study, each review was broken up into a vector of unigram-based tokens. Stop words are words that do not contribute to the meanings of the text and are usually filtered out before the processing of natural language data.

The unsupervised approach to SM analysis does not require prior training to classify the data, as only input data (X) is used. The lexicon-based method is a popular unsupervised method for determining the polarity and semantic orientation of SM statements that involves predefining lexicons of positive and negative words and phrases (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Turney,2001). The vast majority of the hospitality and tourism studies apply an unsupervised approach of sentiment identification to investigate attributes and sentiments of SM data.

Lexicon-based was chosen as a technique to be used to automatically classify texts. It is based on their sentiment approaches consist in calculating the polarity of a text by accounting the semantic polarity of words or phrases included in the document (Taboada et al., 2011). They require a dictionary of words annotated with sentiment-polarity.

As the objective of this work is to generate intelligence and negative reviews were chosen as a focus, the choice of sentiment emotion lexicon was crucial. Thus, after analyzing several lexicon sentiment applications, the 'NRC Sentiment and Emotion Lexicons' was chosen. This lexicon categorizes words in a binary fashion ("yes"/"no") into categories of emotions, like positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

According to the National Research Council Canada - NRCC (2019) the NRC Sentiment and Emotion Lexicons is a collection of seven lexicons, among which stand out:

- Word-Emotion Association (with more than 14,000 unigrams/words and 25,000-word senses),

- Hashtag Emotion (more than 16,000 unigrams /words),

- Emoticon (more than 62,000 unigrams 677,000 bigrams),

- Emoticon Affirmative Context and Emoticon Negated Context, with more than 45.000 unigrams and 240.000 bigrams in Affirmative contexts and more than 9.00 unigrams and 34.000 bigrams in Negative Contexts.

These lexicons were developed with a wide range of applications in mind and can be used in various contexts, including what we will use, which is Sentiment Analysis. Each lexicon has a list of words and their associations with certain categories of interest, such as emotions (joy, sadness, fear, etc.), feelings (positive and negative), or even colors related to the feelings.

The NRC gives the general sentiment of each sentence and the result of the analysis can score a real value between $-\infty$ (most negative) to ∞ (most positive).

6- RESULTS- REVIEWS FROM TRIPADVISOR

In this work was performed text analysis of the publicly available review data posted on Trip Advisor for five principal tourist attraction of Lisbon Portugal. For testing purposes, we carried out two UGC data mining: one taking into account the months of June / July and August and the other taking into account the first entry existing in each attraction until 31st August 2018. For the purposes of results, we present the second option .

ATTRACTIONS	REVIEWS SCRAPED
CASTELO DE S. JORGE	8747
MOSTEIRO DOS JERÓNIMOS	7075
PRAÇA DO COMÉRCIO	3155
TORRE DE BELÉM	3052
ELÉCTRICO 28	6192

Table 2 – Attractions and Reviews Scrapped

It was scraped a total of **28221** partial entries of reviews from TripAdvisor's website in english using R and analyzed them.

Such volume of unstructured textual data is tough to acquire in a tidy format. In this project, we scraped the reviews using the Selector Gadget and the Rvest package. The initial starting points are the following 5 URLs :

Castelo de São Jorge 'https://www.tripadvisor.in/Attraction_Review-g189158-d195107-Reviews-Castelo_de_S_Jorge-Lisbon_Lisbon_District_Central_Portugal.html' ; Mosteiro dos Jerónimos 'https://www.tripadvisor.in/Attraction_Review-g189158-d195318-Reviews-Mosteiro_dos_Jeronimos-Lisbon_Lisbon_District_Central_Portugal.html'; Praça do Comércio https://www.tripadvisor.in/Attraction_Review-g189158-d199878-Reviews-Praca_do_Comercio_Terreiro_do_Paco-Lisbon_Lisbon_District_Central_Portugal.html'; Torre de Belém 'https://www.tripadvisor.in/Attraction_Review-g189158-d524074-Reviews-Torre_de_Belem-Lisbon_Lisbon_District_Central_Portugal.html' and Eléctrico 28 'https://www.tripadvisor.in/Attraction_Review-g189158-d262792-Reviews-Tram_28-Lisbon_Lisbon_District_Central_Portugal.html.

To carry out the work, we did the scraping and data mining of each of the attractions separately. We applied the following script to everyone.

Using Selector Gadget extension on chrome, we obtained the CSS selector for the partial entries of the reviews. Using the Rvest package in R, we scraped the textual information contained within the corresponding HTML tags in this URL, which only contained 5 reviews.

URL <- 'https://www.tripadvisor.in/Attraction_Review-g189158-d195107-Reviews-Castelo_de_S_Jorge-Lisbon_Lisbon_District_Central_Portugal.html'
html <- **read_html**(url)

```
review_count<-html%>%
 html_nodes('.pagination-details')%>%
 html_text()
review_count <- strsplit(review_count,'of')[[1]][2]
review_count <- gsub('reviews','',review_count)
review_count <- as.numeric(gsub(',','',review_count))
reviews<-html%>%
 html_nodes('.partial_entry')%>%
 html_text()
num <- seq(5,review_count,5)
```

The next challenge was to scrape all reviews in a programmable and reproducible manner without opening each URL explicitly. After some investigation of the various URLs of the review pages, we identified a pattern. The number following the characters "OR" in the URL increments by 5 each time we move from page n to page n+1. Using this information – it was able to write code that loops over all pages and scrape all available reviews.

```
for (i in 1:length(num)){
 revurl <- paste0('https://www.tripadvisor.in/Attraction_Review-g189158-d195107-Reviews-or',num[i],'-Castelo_de_S_Jorge-Lisbon_Lisbon_District_Central_Portugal.html')
 html<- read_html(revurl)
 reviewed <- html%>%
  html_nodes('.partial_entry')%>%
  html_text()
 reviews <- c(reviews,reviewed) }
write.csv(data.frame(reviews=reviews),'reviews2.csv',row.names = F)
```

Once the content reviews of each attraction were extracted, the packages removed stop words and other characters without evaluated significance.

From that moment, through the TidyText and Syuzhet packages, we used the lexcom NRC to perform the Sentiment Analysis.

With the extraction and treatment of the data, we obtained the content reviews ratings and made the manual verification of the results. From each attraction, we remove 20 ratings and check if the result of the sentiment analysis was correct. Here are some examples:

Examples of reviews and classification

Review	sentiment
"We went on a tramway with two young children but very disappointed Too many people and we couldnt see anything so waste of time in my opinion We will try different one tomorrow."	-1.462
"I would have probably liked the tram ride very much but unfortunately our tram broke down on the second stop Which meant we either had to walk back and wait in line again or to wait for another tram at this tram stop and to..."	-1.191
"A great place with great views but 8.50 is a bit too much. Take the time to explore the small Moorish streets around the castle"	-0.8164

"Wasnt worth the money. Nice view but thats it I wish I would have spent this time doing something worthwhile. It was a barren castle that you couldnt get out of Nothing to see."	-0.8095
"Great views of Lisbon Really beautiful surroundings all very well maintained Well worth a visit highly recommended"	2.023
"Stunning monastery queues but well worth the wait stunning from the outside and even more inside well worth a visit.."	1.984
"Brilliant visit to castle getting there on hop on hop off bus and tram back to the city centre. Very interesting excellent well informed guide made visit really interesting including walking the castle walls Need a dry day or very slippery underfoot cafes and shops inside castle and several...	1.595
"Lovely place lots of variety different shopsgreat atmosphere we really enjoyed exploring great architecture really enjoyed ourselves lots to see..."	1.561

Continuing with the work of analysis, we researched the frequency of words and also the bigrams - which could indicate points of attention.

To check the programming result, viewing results such as word cloud or frequency of negative/positive words can show if there are any discrepant results. In this way, a visualization is a powerful tool for investigation, even when working with unstructured data. Helps to detect anomalies in the analysis. If a positive word appears in the negative word cloud, you can revisit the code and correct it. And it is also a quick attention tool that facilitates the performance of the intelligence/data analyst.

Commonly occurring words in Negative Reviews

Figure 12 – Word of Clouds – Words in Negative Reviews- Castelo de São Jorge

With the creation of a cloud of common negative words that occur in Castelo de S. Jorge's criticisms, some inferences can be made immediately. What can be seen is that many people found the attraction crowded, expensive, disappointing, irregular, steep and with difficult access.

What are the most common negative bigrams?

Figure 13 – Word of Cloud – Most Common Negative Bigrams- Castelo de São Jorge

According to the proposed methodology, actionable information would be created from content negative reviews, that is, the results presented were focused on problems and complaints.

From this extracted information, an intelligence analyst could supply the management of suggestions for actions - aiming to improve the quality of service and the attractiveness of the tourist spot.

In order to facilitate the analyst's work and also to visually show the predominance of feelings present in content reviews, we created two graphics – to be used in a possible dashboard:

- A- the first shows the predominant score of feelings - On the X- axis we have the emotions and on the Y- axis the weighted average of the strength of each emotion (from 0 to 8 – according to the values returned by the NRC);

- B- the second shows us the density of feeling present in the set of content reviews, which can be positive or negative.

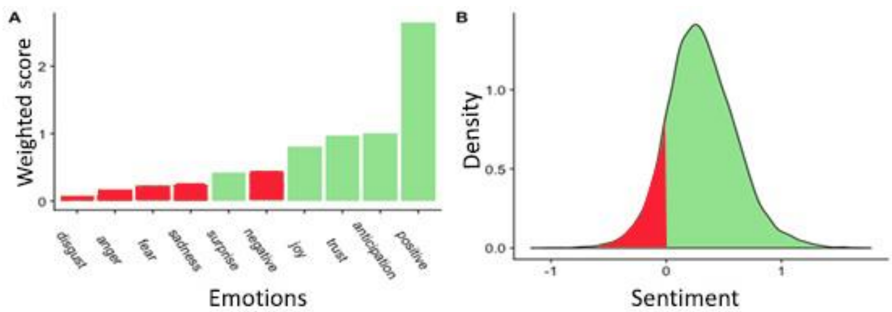

Figure 14 - A- Sentiment Score B- Sentiment X Density - Castelo de S. Jorge

A trained analyst will be able to understand the indication of sentiments (positive, negative or neutral) and the predominance and strength of the associated emotions.

To facilitate the work of the analyst, we first suggest that all the so-called “negative emotions” as well as the negative sentiment have the graphic representation in red - providing a visual sign of the situation.

We suggest the following workflow for the analyst:

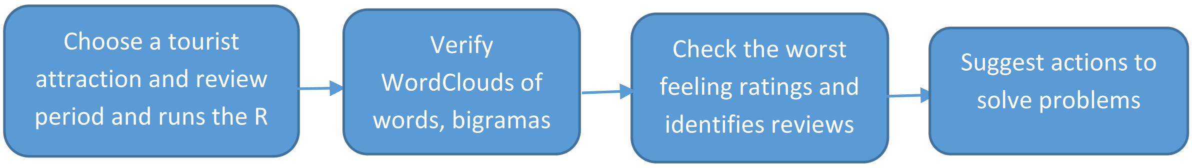

Figure 15 – Workflow for the intelligence analyst.

In the following topics we will have a simulation of the analysis of the work to be done from the data collected.

For the figure 14 -as seen above the overall sentiment of this attraction is skewed towards the positive side with the mean sentiment occuring at around 0.2. This means that the sentiment of a majority of reviews has been on a positive side.

6.1- **Analysis of the data obtained and generation of intelligence - information in action**

6.1.1-“ Castelo de S. Jorge”

Part of the data regarding the Castelo de São Jorge was presented in the previous section, but served, together with the reviews with more negative notes and more frequently, to support the creation of actionable information.

Examples of reviews classified as negative

Review	sentiment
“Beautiful castle but too many tourists Take your time for the walls its getting really jammed during the day.”	-1.032
“A great place with great views but 8.50 is a bit too much. Take the time to explore the small Moorish streets around the castle”	-0.8164
“Wasnt worth the money. Nice view but thats it I wish I would have spent this time doing something worthwhile. It was a barren castle that you couldnt get out of Nothing to see.”	-0.8095
“Its worth going there to see the views of Lisbon and it has a good atmosphere. However I wish residents in Portugal could go free as we stop taking our visitors there as we have to pay every time.Surely enough money would be made more.”	-0.7467

Possible Actiontable Information

While being busy and crowded means that lots of people are visiting the castle and that it is very popular, it also means that sometimes it could be too crowded for anyone to enjoy it properly. So based on this text analysis, the most significant insight that we can take away is that the management needs to implement some crowd control strategies such as setting a limit on the maximum number of people who can enter the monastery at any given point of time. Similarly, tickets could be sold in advance to get an estimate of the footfalls on any given day.

Another key insight that we gained was that people found the visit to this castle expensive. This could be rectified by offering discounts on ticket sales from time to time. Similarly, a flat discount can be given to those books their tickets in advance and lastly, maybe the ticket to the monastery can be included in some sort of combinatory day pass that will allow tourists to visit other places as well.

Other critical bigrams include uneven steps, people getting lost, and a steep climb. All of these can be addressed by taking the appropriate action, such as implementing tour guides and signposts while also renovating the pathway to the castle.

6.1.2- "Mosteiro dos Jerónimos"

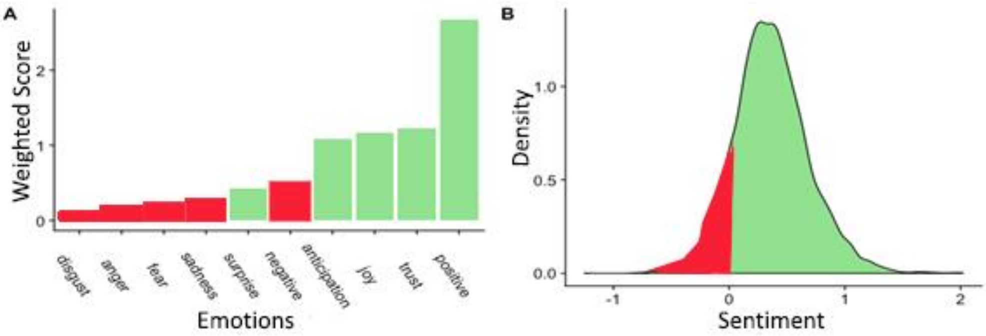

Figure 16 - A- Sentiment Score B - Sentiment X Density

As seen above, the overall sentiment of this attraction is skewed towards the positive side, with the mean sentiment occurring at around 0.3. This means that the sentiment of a majority of reviews has been on a positive side.

Examples of reviews classified as negative

Review	sentiment
"Very interesting visit but would be difficult for those with mobility issues due to the uneven stone pavements.."	-1.251
"Lovely building but hopeless crowding Pity but go somewhere else there are plenty of old churches in Lisbon…"	-0.7953
"A must do if in Lisbon.Must add to your list of gotta see The size alone of the Monastery amazed us Could have spent more time looking but crowded…"	-0.7929
"Architecturally this is stunning but didnt find it a particularly comfy ride too many overtones of too many people who suffered guess Wouldn't rush back here on a return visit to Lisbon there is plenty more to see...	-0.7139

Commonly occurring words and bigrams in Negative Reviews

Figure17   A- Word of Clouds – Words   B - Negative Bigrams

Here it can be seen a word cloud of the common negative words occurring in the reviews for Mosteiro dos Jeronimos. Immediately what can be seen is that plenty of people found the attraction crowded, expensive, disappointing, confusing, and imposing.

What descriptive monograms are the most commonly used in all reviews?

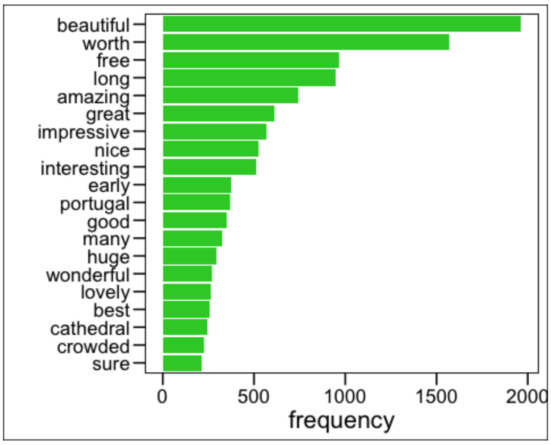

Figure 18 - Word's frequency

Possible Actiontable Information

Based on this text analysis, the most significant insight that we can take away is that the management needs to implement some crowd control strategies such as setting a limit on the maximum number of people who can enter the monastery at any given point of time. Similarly, tickets could be sold in advance to get an estimate of the footfalls on any given day.

Another key insight that we gained was that people found the visit to this monastery expensive. This could be rectified by offering discounts on ticket sales from time to time. Similarly, a flat discount can be given to those book their tickets in advance and lastly, maybe the ticket to the monastery can be included in some sort of combinatory day pass that will allow tourists to visit other places as well.

6.1.3- "Praça do Comércio"

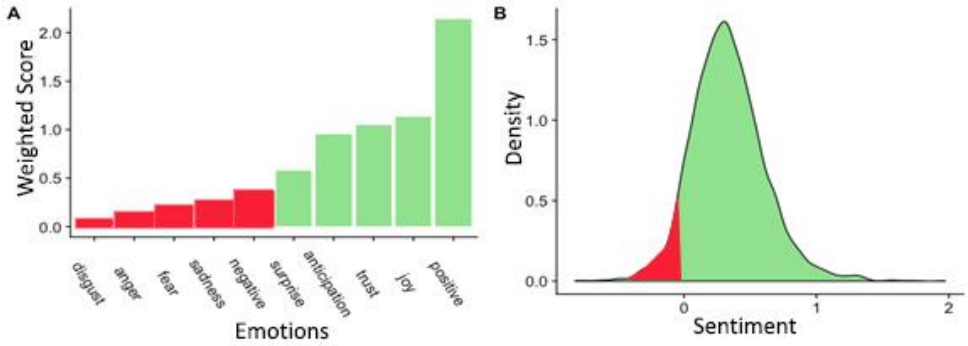

Figure 19 - A- Sentiment Score B- Sentiment X Density

As seen above the overall sentiment of this attraction is skewed towards the positive side with the mean sentiment occurring at around 0.15. This means that the sentiment of a majority of reviews has been on a positive side.

Examples of reviews classified as negative

Review	sentiment
"Too many hawkers of useless hype goodies for kids too many pushy waiters almost shoving you on a chair in their restaurant That being said the praca offers you a great walk and view on the Tagus estuary."	-0.8225
"But spoiled but cheap eateries and down scale shops a real pity that it has not been maintained in the tradition We passed by and then avoided the area."	-0.7428
"This area is loaded with tourists bakeries and restaurants Lots of touts trying to get you in their establishments We never eat on this street as per the reviews most places are below average and are tourist traps."	-0.6327
"This area of the city is beautiful but attracts too many tourists So youd probably want to get out of there ASAP Especially annoying are those drug dealers that try to offer their stuff to you on every corner The best part of it all..."	-0.6081

Commonly occurring words and bigrams Negative Reviews

Figure 20   A- Word of Clouds – Words   B - Negative Bigrams

The figures give us quickly the following information: what can be seen is that some people described their experience as negative, expensive, imposing, difficult.

What descriptive monograms are the most commonly used in all reviews?

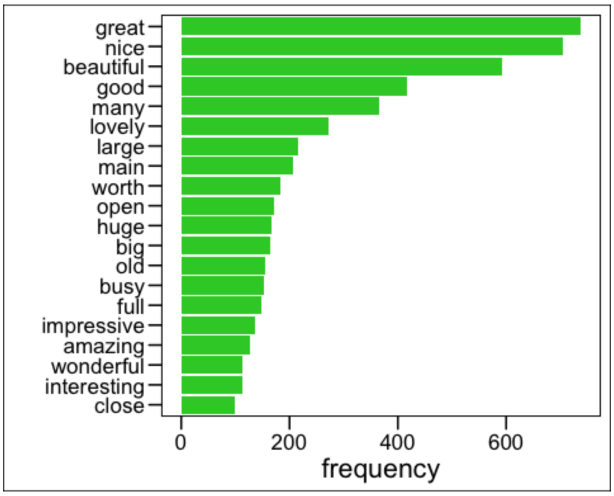

Figure 21   - Word's frequency

Possible Actiontable Information

The biggest insight that we can take away is that the management needs to implement some crowd control strategies. Another key insight that we gained was that people found the visit to this square expensive. Similarly, plenty of tourists described the attraction as a tourist trap. These kinds of things can be addressed by taking appropriate measures. The cafes and restaurants have been described as expensive as well.

6.1.4- "Torre de Belém"

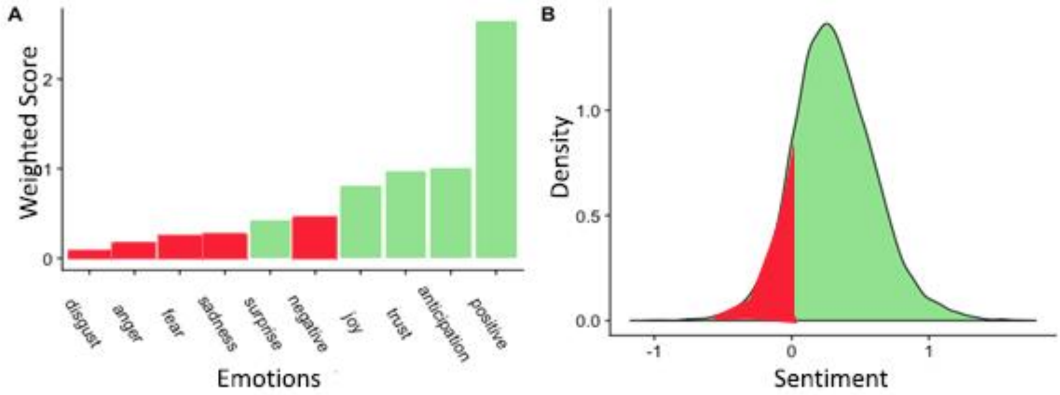

Figure 22 - A- Sentiment Score B- Sentiment X Density

As seen above the overall sentiment of this attraction is skewed towards the positive side with the mean sentiment occuring at around 0.2. This means that the sentiment of a majority of reviews has been on a positive side.

Examples of reviews classified as negative

Review	sentiment
"It is a beholding site but the vendors were very annoying and quite aggressive in trying to sell their products with you Hope the municipality will do something about these aggressive type of selling as it is quite alarming."	-1.175
"Its very beautiful but unfortunately there were way too many people on the line as you can see on my picture so we didnt get to see what inside the tower if you plan to go in to the tower try to arrive earlier."	-1.048
"It has a long line and the stairs too narrow and disordered to go up and down they have something like traffic lights but not properly working need better sync There are too many places better in Lisboa."	-0.9916
" much crowded... Oh My God!!!! Also there was a lot of crowd to catch the bus We waited 1 hour to get into our bus.."	-0.947

Commonly occurring words and bigrams in Negative Reviews

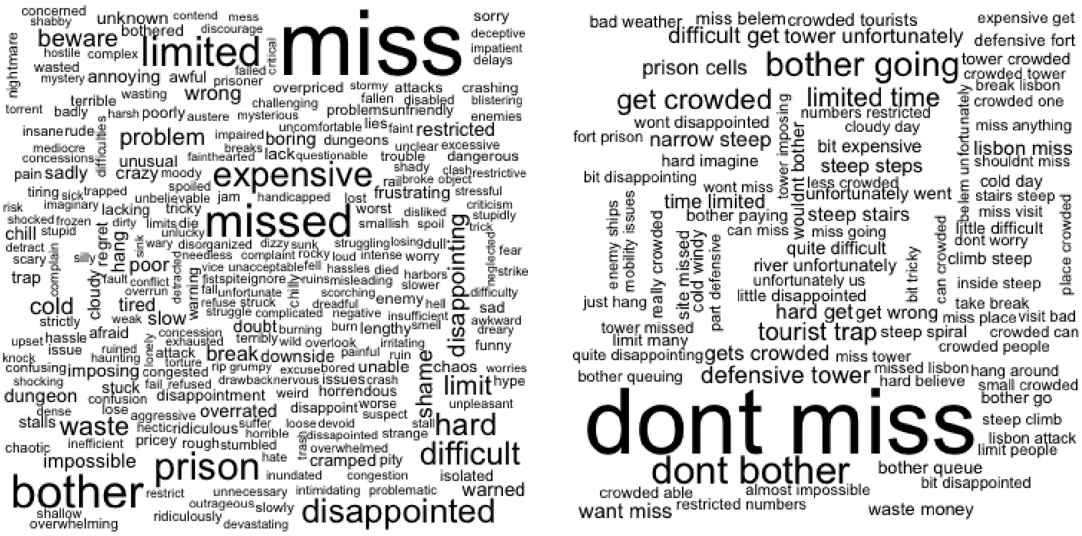

Figure 23 - A- Word of Clouds – Words B - Negative Bigrams

The figure 23 A and B give us quickly the following information: what can be seen is that plenty of people found the attraction crowded, expensive, cramped, steep, and difficult.

What descriptive monograms are the most commonly used in all reviews?

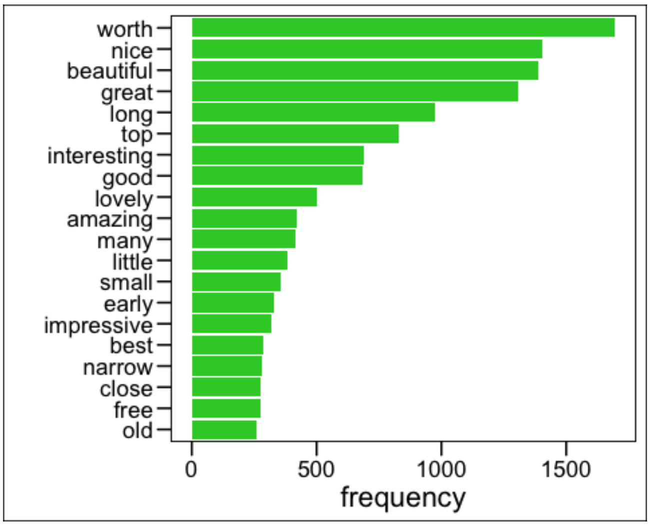

Figure 24 - Word's frequency

Possible Actiontable Information

Plenty of reviews suggest that the tourist shouldn't be bothered about going inside this attraction and that it may be a waste of time. Similarly, crowd control also seems like a significant problem. Another key insight that we gained was that people found the visit to this attraction is too expensive. This could be rectified by offering discounts on ticket sales from time to time. Similarly, a flat discount can be given to those book their tickets in advance and lastly maybe the ticket to the tower can be included in some sort of combinatory day pass that will allow tourists to visit other places as well. Another detailed point was the “limited time” . Because it is a small traction, it could in summer have the time extended to 2 hours, because it would allow a beautiful view of the Tagus at sunset.

6.1.5 - Tram 28 /Eléctrico 28

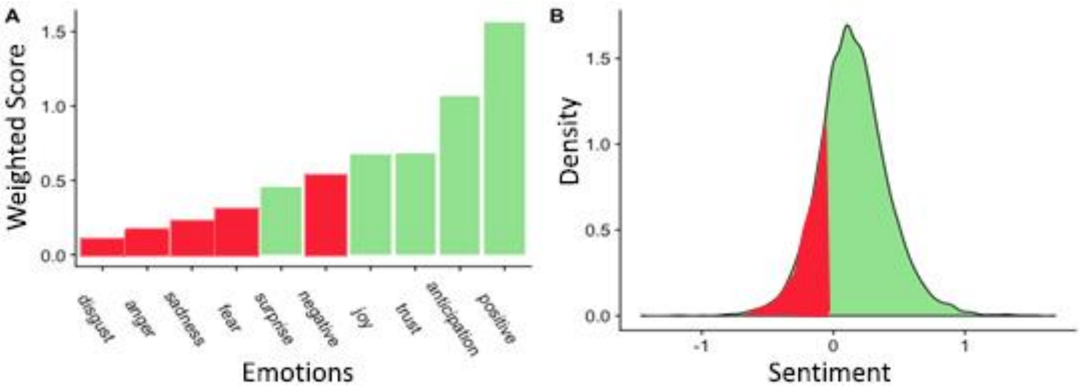

Figure 25 - A- Sentiment Score B- Sentiment X Density

As seen above, the overall sentiment of this attraction is skewed towards the positive side with the mean sentiment occurring at around 0.1. This means that the sentiment of a majority of reviews has been on a positive side.

Examples of reviews classified as negative

Review	Sentiment
“We went on a tramway with two young children but very disappointed Too many people and we couldnt see anything so waste of time in my opinion We will try different one tomorrow.”	-1.462
“I would have probably liked the tram ride very much but unfortunately our tram broke down on the second stop Which meant we either had to walk back and wait in line again or to wait for another tram at this tram stop and to...”	-1.191
“Iconic and charming but simply too crowded Something should change I wish I could have come at dawn but impossible on this stopover,”	-1.152
“We returned from castle to city centre extremely busy, but one of experiences you should have while in Lisbon but word of caution jam packed seats full standing room only for a very bumpy ride warned by some Lisbon residents on tram 28 to watch...”	-0.9912

Commonly occurring words and bigrams in Negative Reviews

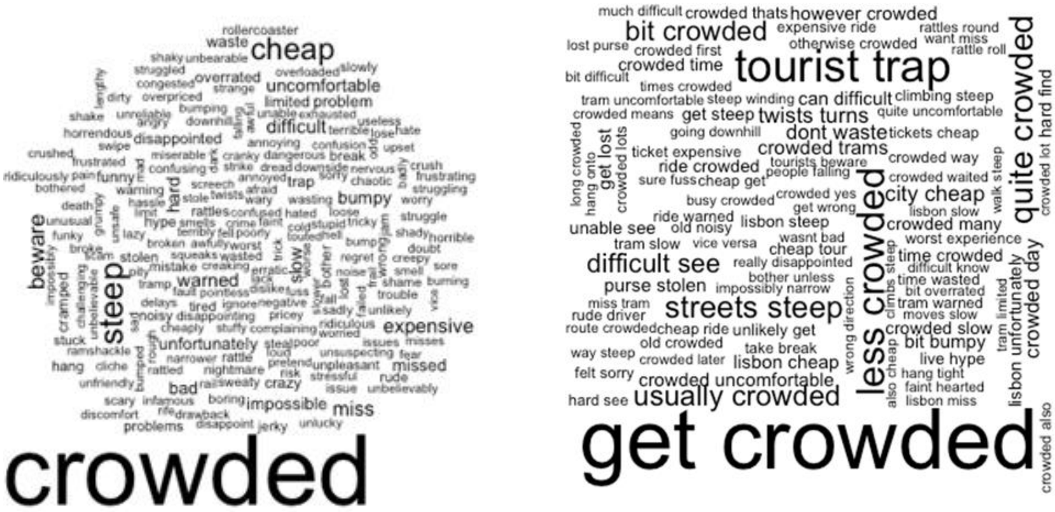

Figure 26 -   A- Word of Clouds – Words   B - Negative Bigrams

Here we see a word cloud of the common negative words occurring in the reviews for Tram 28, Lisbon Portugal. Immediately what can be seen is that some people described their experience as crowded, bumpy, uncomfortable, steep, slow and even a "tourist trap".

What descriptive monograms are the most commonly used in all reviews?

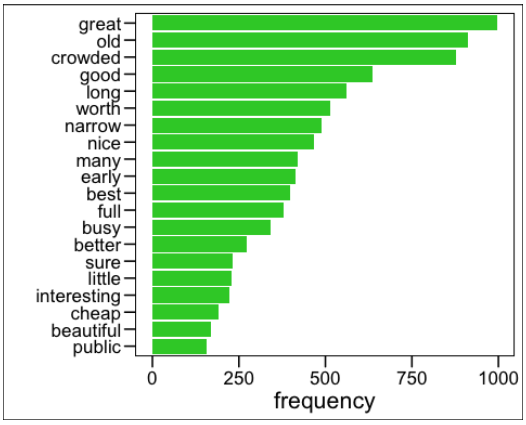

Figure 27 - Word's frequency

Possible Actiontable Information

Two terms define Tram 28 - crowded and with little security . Being crowded means that many people visit the streetcar and that it is very popular, it also means that sometimes it can be too crowded for anyone to enjoy it properly. So, based on this text analysis, the biggest perception we can get is that management needs to implement some crowd control strategies. Another key perception we obtained was that people considered visiting this streetcar unsafe, as many analyses contained words such as "stolen" and "careful". Similarly, many tourists described the attraction as a tourist trap. This kind of thing can be approached by taking appropriate measures such as increasing the frequency of trams, as well as the presence of security guards

7- CONCLUSION

Social media and customer review websites have changed the way the tourism and sector is managed. According several authors (Agichtein et al. 2008; Patin at al. 2012) social media has become a new source of information. An information that is "available" but at the same time noisy and of great volume. There reside some difficulties - that to be transposed involve technique, processing and analysis.

Some authors cite the difficulties of obtaining information from this "mess" of data (Agarwal & Ylliasi, 2010; Töllinen et al. 2012) , which can be overcome in part with the correct technique. Others highlight that social media analytics is a vital tool that must be used to innovate decision support (Senecal & Nantel, 2004) and Young (2016) goes so far as to claim that "customer feedback data is simply too valuable and too powerful to be ignored".

Given the difficulties and focused on the possibilities, we continue on our work.

One of our concerns was to see if the theme : **USING THE TRIPADVISOR TO MONITOR AND GENERATE INTELLIGENCE FOR A DMO - focus Negative Reviews** had already been explored. As seen in the Literature Review and after extensive searches in databases such as Web os Science , Ebscohost and even Google Scholar , we verified that it - in the way it was proposed has no similar . We also used the PRISMA method and the VosViewer software to support the results obtained.

So, we found that the choice of the topic was unprecedented: the use of negative UGC made from existing attractions on TripAdvisor as a source of information for management and decision support.

After this first phase, we looked for the answers to the research questions: how can organizations turn it into actionable information and knowledge?

I - How can DMO - VisitLisboa conduct media monitoring to capture and analyze media data?

II - How can DMO - VisitLisboa identify the strengths and weaknesses of tourist attractions using media information? Would it be possible to identify the strengths and weaknesses of a DMO's attractions and position them effectively in the market using media information.

Before answering them, we would like to comment on some factors that have helped us.

Two choices that helped us have more accurate results - due to the focus generated.First, it is the attractions that make the destination visited, as verified in the literature review. The tourist attraction is one of the most important elements of the tourist destination and infers a critical factor for the success rate of tourism management. Second, the choice of negative evaluations, in addition to showing us faster data, showed us many insights to improve the quality processes in serving tourists and even in the attractiveness of tourist objects.

Third, we were able to suggest a framework for Social Media Intelligence and we were able to apply it in the process of extracting data from TripAdvisor and generating actionable information – intelligence .

The use of the KIT Questionnaire, proposed in the SOCMINT framework (figure 15), facilitated the choice of the subject to be addressed and helped in the implementation of the research.

Regarding the choice of the R language as a way of performing data mining (scraping-tokenization-sentiment analysis), it proved to be effective. Mainly because it is a free tool and with a great offer of packages, mainly in the areas of Natural Language Processing and Sentiment

Analysis. The possibilities of generating word clouds and graphics also facilitated the process of analyzing the results and quickly viewing the problems.

Corroborating the views of Bose (2008) and Wurman (2001), we also believe (nowaday) that the action of a person as an intelligence analyst can still do the job more accurately, although slower, than with the use of Artificial Intelligence.

Now we can answer the questions. In order to transform data (even in Big Data and social media) into information, we were able to prove that the use of the code and its packages enabled the extraction, cleaning and selection of data and facilitated the analyst's work to generate actionable information.

The use of R and the extraction of more than 28 thousand e-Wom in the chosen period answer question I - we were able to evaluate (through the program) all e-Wom and thus monitor all the chosen attractions. Full details are available in Chapter 6.

As for question II, we were able to assess the associated word clouds and bigrams to verify the weaknesses of each attraction. In this way, the supposed analyst could generate suggestions for improvement and correction. More details are also available in Chapter 6.

We see from the results obtained that Litvin et al. (2008) were right in stating that the e-WOM can be a substantial source of strategic information to be used for the development of a series of business strategies. We believe that once the suggested actionable information is implemented, a DMO can understand the experience of the visitors and achieve increased visitor satisfaction through improved visitor experience, problem solving, competitive strategy analysis, as well as monitoring the image and reputation of a tourist attraction, and consequently, of a city.

8 –LIMITATIONS, MANAGERIAL IMPLICATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Limitations

Both theoretical and practical aspects can be addressed in the limitations and recommendations for future work.

We can highlight at least five limitations:

- the use of TripAdivisor as a source of information. Although it is the most important CRW linked to tourism, there are other representatives that could be used as reference;

- limitations of sentiment analysis still exist and an intelligence analyst can check the consistency of the results from the application of the code;

- the use of the R code/program - Python or any other program could be used, as well as dedicated software as an import and treat the data. There are several tools, like RapidMiner, WebSites Scraper, Import.io, Octoparse and WebHarvy. We tried to use these last two for comparison purposes, but the result was far short of the extraction possibility we achieved using the "R”.

- the use of the Lexicon NRC present in the packages available in the "R", such as "Affin" and "Bing".

- the language - The English language can be considered “universal”, , but in the case of Lisbon, other languages such as Spanish , French and Portuguese are important participants in the existing e-Wom in TripAdvisor.

Despite these limitations, we were able to answer the research questions and carry out a work with a different scope than has been usual in research involving the tourist industry.

Managerial Implications and Suggestions

First, we suggest the use of the social media intelligence process (figure 11), which can be adapted to any industry that has its products and services subject to social media analysis

Secondly, the whole process of extraction and analysis can be done by a computer program, but for now, as we understand it, the existence of an intelligence analyst can transform this extraction into a high-value, actionable information product.

Third, we suggest that the extracted information can be divided into future suggestions for short, medium and long term actions, according to its content.

A fourth suggestion would be to use the proposed framework with the use of the KIT (Herring 1999) to benchmark competitors, which can be, in the case of tourism, the main attractions of another city and how consumers are evaluating them. Or you can compare the results obtained from the analysis of your products/services against the competitor present in another CRW.

And finally, a fifth suggestion - using Social Media Intelligence - by looking for qualitative and quantitative data directly from the consumer, with good software or program and an experienced analyst, a company can listen to the market with much faster and less expense than a traditional Market Research.

Having data available, not extracting it and generating intelligence can lead to serious problems. In the case of this work , the Torre de Belém is a typical example of not understanding its clients - it went from 685 thousand visitors in 2016 to only 427 thousand visitors in 2019 (source - Direcção-Geral do Património Cultural- Portugal), while the other attractions remained with small variations . This may have represented a drop in revenue of 1 million euros .

As a final learning and management suggestion after this work, we would like to address that we should not act as if the data available to us were sufficient. With Big Data, we should always look for more, because one of the biggest mistakes is a biased sample. Another point to highlight is never observe the summarized data and deduce what is happening separately. The ideal is to make a more detailed analysis to get a correct view. And have a lot of experience in intelligence generation, because an analyst without the right skills may discover that he has "discovered" something that is not really in the data.

Recomendations for Future Works

Based on the limitations suggested above, future works can be suggested:

- the comparison of results applying both a program code and a pre-existing program. For example, compare the effectiveness of using R with Python , or with a commercial program like Import.io or Rapidminer;

- the comparison of results of the use of various Lexicons , for example NRC with Affin ;

- the comparison of results obtained from the e-Wom analysis of different languages;

- the comparison of results of the use of Social Media Mining and traditional Market Research;

- The use of Social Media Intelligence with Artificial Intelligence in the possible transformation of data into actionable information.

REFERENCE

Agarwal, N., & Ylliasi, Y. (2010). Information quality challenges in social media. *The 15th International Conference on Information Quality*, ICIQ, Little Rock, Arkansas, USA. Retrieved 22 mar. 2019 from https://pdfs.semanticscholar.org/ece8/48ac7435350acdb54a5a661f611f9325cbc1.pdf?_ga=2.173752547.1503290287.1598872606-1660303863.1596923915

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. Proceedings of the International Conference on Web Search and Web Data Mining - WSDM 08. https://doi.org/:10.1145/1341531.1341557

Ahmad, S. (2013). Web Mining Pedagogy: The Theoretical Support. *International Journal of Computing. Intelligent and Communication Technologies*, 2, 17–22. : <http://ijcict.com/doc/VOL2_ISSUE2_MAY13/5.pdf>

Amadio, W. J., & Procaccino, J. D. (2016). Competitive analysis of online reviews using exploratory text mining. *Tourism and Hospitality Management*, 22(2), 193–210. https://doi.org/ 10.20867/thm.22.2.3

CRAN Packages By Name. (n.d.). Retrieved from https://cran.r-project.org/web/packages/available_packages_by_name.html#available-packages-P

Barbier, G. & Liu, H.(2011) *Data mining in social media.* In: Social Network Data Analytics. Springer US, 2011. p. 327-352.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2015). Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. *Journal of Hospitality Marketing & Management*, 25(1), 1–24. https://doi.org/10.1080/19368623.2015.983631

Bindra, G. S., Kandawal, K. K., Singh, P. K., & Khana, S. (2012). Tracing Information Flow and Analyzing the Effects of Incomplete Data in Social Media. 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks, Phuket, 2012, pp. 235-240. https://doi.org/ 10.1109/CICSyN.2012.51.

Blackshaw, P., & Nazzaro, M. (2006) *Consumer-Generated Media (CGM) 101: Word-of-Mouth in the Age of the Web-Fortified Consumer*. New York: Nielsen.

Bronner, F., & de Hoog, R. (2013). Economizing on vacations: the role of information searching. *International Journal of Culture, Tourism and Hospitality Research*, 7(1), 28–41. https://doi.org/10.1108/17506181311301336

Bose, R. (2008). Competitive intelligence process and tools for intelligence analysis. Industrial Management & Data Systems, 108(4), 510–528. https://doi.org/10.1108/02635570810868362

Buhalis, D., Costa, C., & Ford, F. (2006). *Tourism Business Frontiers*. Elsevier Gezondheidszorg. 1-273. https://doi.org/10.4324/9780080455914.

Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the internet: The state of eTourism research. *Tourism Management*, 29(4), 609–623. https://doi.org/10.1016/j.tourman.2008.01.005

Buhalis, D., & Foerste, M. (2015). SoCoMo marketing for travel and tourism: Empowering co-creation of value. *Journal of Destination Marketing & Management*, 4(3), 151–161. https://doi.org/10.1016/j.jdmm.2015.04.001

Carson, D., & Sharma, P. (2001). Trends in the use of Internet technologies. World Hospitality and Tourism Trends, 2(3), 116–128. Retrieved 12 sept. 2019 from https://www.researchgate.net/publication/261878811_Trends_in_the_Use_of_Internet_Technologies

Chang, Y.-C., Ku, C.-H., & Chen, C.-H. (2019). Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *International Journal of Information Management*, 48, 263–279. https://doi.org/ 10.1016/j.ijinfomgt.2017.11.001

Chen, J. S., & Uysal, M. (2002). Market positioning analysis: A hybrid approach. *Annals of tourism research*, 29(4), 987-1003. http://dx.doi.org/10.1016/S0160-7383(02)00003-8

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems*, 36(4), 1165-1188. https://doi.org/10.2307/41703503

Cheung, M. Y., Luo, C., Sia, C. L., & Chen, H. (2009). The credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce*, 13(4), 9–38 . https://doi.org/10.2753/jec1086-4415130402

Chung J.Y.,& Buhalis D. (2008) Web 2.0: A study of online travel community. In: O'Connor P., Höpken W., Gretzel U. (eds) *Information and Communication Technologies in Tourism 2008*. Springer, Vienna. https://doi.org/10.1007/978-3-211-77280-5_7

Cook, M.; & Cook, C.. *Competitive Intelligence: create an intelligent organization and compete to win*. Glasgow: Kogan Page, 2000

Coutinho, F., Lang, A. & Mitschang, B., (2013). Making social media analysis more efficient through taxonomy supported concept suggestion. In: Markl, V., Saake, G., Sattler, K.-U., Hackenbroich, G., Mitschang, B., Härder, T. & Köppen, V. (Hrsg.), Datenbanksysteme für Business, Technologie und Web (BTW) 2040. Bonn: Gesellschaft für Informatik e.V.. (S. 457-476) https://www.semanticscholar.org/paper/Making-Social-Media-Analysis-more-efficient-through-Coutinho-Lang/aa87842a90a83dc4ab12ac60059e488beeb889a5

Coutinho, M. Marketing e comunidades digitais: do discurso ao diálogo. Revista da ESPM, São Paulo, v.14, n.2, p.28-39. 2007. : <http://www.ideiacom.com.br/gerenciador/arquivos/documentos/artigo_marcelo_coutinho.pdf>. retrieved : 22 ago. 2018.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2012). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, 17(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

Crotts, J. (1999). "Consumer Decision Making and Prepurchase Information Search." In *Consumer Behavior in Travel and Tourism*, edited by Yoel Mansfield and Abraham Pizam. Binghamton, N.Y.: Haworth Press, pp. 149-168.

Dai, Y., Kakkonen, T., & Sutinen, E. (2010). MinEDec: A decision support model that combines text mining with competitive intelligence. 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM). https://doi.org/10.1109/cisim.2010.5643661

Dai, Y., Kakkonen, T., & Sutinen, E. (2011). SoMEST: a model for detecting competitive intelligence from social media. In Proceedings of The 15th International Academic MindTrek Conference: Envisioning Future Media Environments, p. 241-248, Tampere, Finland. https://doi.org/10.1145/2181037.2181078

Dellarocas, C. N. (2003). The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback Mechanisms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.393042

Denizci G., Kucukusta, D. & Liu, L.. (2016). An Examination of Social Media Marketing in China: How do the Top 133 Hotel Brands Perform on the top four Chinese Social Media Sites? *Journal of Travel & Tourism Marketing*. https://doi.org/10.1080/10548408.2015.1064337.

Devika, K., & Surendran, S. (2013). An overview of web data extraction techniques. *International journal of scientific engineering and technology*, 2(4), 278-287. ISSN: 2277-1581

Dey, L., Haque, S. M., Khurdiya, A., & Shroff, G. (2011). Acquiring competitive intelligence from social media. *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data* - MOCR_AND '11, 0. https://doi.org/10.1145/2034617.2034621

Dickinger, A., & Mazanec, J. A. (2015). Significant word items in hotel guest reviews: A feature extraction approach. *Tourism Recreation Research*, 40(3), 353–363. https://doi.org/10.1080/02508281.2015.1079964

Dishman, P. L., & Calof, J. L. (2008). Competitive intelligence: a multiphasic precedent to marketing strategy. *European Journal of Marketing*, 42(7/8), 766–785. https://doi.org/10.1108/03090560810877141

D'orazio, F. (2013). The future of social media research: or how to re-invent social media listening in 10 steps. Retrieved January 28, 2019, from PulsarPlatform. https://www.pulsarplatform.com/blog/author/francesco-dorazio/page/3/

Faed A., & Forbes, D. (2010). *Impact of Customer Management System in Improving Customer* Retention: Optimization of Negative Customer Feedback (Version 11734). http://doi.org/10.5281/zenodo.1077839

Fotis, J., Buhalis, D., & Rossides, N. (2012). Social Media Use and Impact during the Holiday Travel Planning Process. *Information and Communication Technologies in Tourism* 2012, 13–24. https://doi.org/10.1007/978-3-7091-1142-0_2

Franzoni, S., & Bonera, M. (2019). How DMO Can Measure the Experiences of a Large Territory. *Sustainability*, 11(2), 492. MDPI AG. Retrieved from http://dx.doi.org/10.3390/su11020492

Gates, W. H., & Hemingway, C. (1999). *Business @ the speed of thought: succeeding in the digital econom*y. London: Penguin.

Gémar, G., & Jiménez-Quintero, J.A. (2014). Text mining social media for competitive analysis. *Tourism & Management Studies,* 11(1) , 84-90 . Retrieved https://www.semanticscholar.org/paper/Text-mining-social-media-for-competitive-analysis-G%C3%A9mar-Jim%C3%A9nez-Quintero/ad5bc0b68485dbdfa95f9df4c433d02d33b6ab7c

Gilad, B. (2003). *Early Warning: Using Competitive Intelligence to Anticipate Market Shifts, Control Risk, and Create Powerful Strategies* (Illustrated ed.). Amacom.

Gin, J. (2016, January 26). 7 Ways to Use Negative Customer Feedback to Beat the Competition. Retrieved 18 jan 2019 from https://www.entrepreneur.com/article/254553

Goldenberg, J., Libai, B. & Muller, E..(2001) Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters* 12, 211–223 https://doi.org/10.1023/A:1011122126881

Gretzel, U. (2006). Consumer-generated content – trends, and implications for branding. *e-Review of Tourism Research* 2006 Vol.4 No.3 pp.9-11 Retrieved 21 feb 2019 from https://www.researchgate.net/publication/242556754_Consumer_Generated_Content_-_Trends_and_Implications_for_Branding

Gretzel, U., Fesenmaier, D. R., Formica, S., & O'Leary, J. T. (2006). Searching for the Future: Challenges Faced by Destination Marketing Organizations. *Journal of Travel Research*, 45(2), 116–126. https://doi.org/10.1177/0047287506291598

Gretzel, U., Yoo, K. H., & M. Purifoy (2007). Online Travel Reviews Study. College Station, TX: Laboratory for Intelligent Systems in Tourism

Gretzel, U.; Yoo, K. H.(2008) Use and impact of online travel reviews. In: O'Connor P., Höpken W., Gretzel U. (eds) *Information and Communication Technologies in Tourism* 2008. Springer, Vienna. https://doi.org/10.1007/978-3-211-77280-5_4

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (The Morgan Kaufmann Series in Data Management Systems) (3rd ed.). Morgan Kaufmann.

Hays, S., Page, S. J., & Buhalis, D. (2013). Social media as a destination marketing tool: its use by national tourism organisations. *Current Issues in Tourism*, 16(3), 211–239. https://doi.org/10.1080/13683500.2012.662215

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, v. 33, n. 3, p. 464-472, <http://www.sciencedirect.com/science/article/pii/S0268401213000030>. Retrieved in: 02/12/2018.

Hebert, D., Anderson, B., Olinsky, A., & Hardin, J. M. (2014). Time Series Data Mining: A Retail Application. *International Journal of Business Analytics* (IJBAN), 1(4), 51-68. https://doi.org/:10.4018/ijban.2014100104

Herring, J. P. (1999). *Key intelligence topics: A process to identify and define intelligence needs. Competitive Intelligence* Review, 10(2), 4–14. https://doi.org/:10.1002/(sici)1520-6386(199932)10:2<4::aid-cir3>3.0.co;2-c

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, 72, 417–426. https://doi.org/10.1016/j.tourman.2019.01.002

Huang, X., Croft, W. B. (2009). A unified relevance model for opinion retrieval. *In Proceedings of 18th ACM conference on Information and knowledge management*, p. 947-956, Hong Kong, China. https://doi.org/10.1145/1645953.1646075

Kahaner, L. (1996). *Competitive Intelligence: How to Gather Analyze and Use Information to Move Your Business to the Top (*1st ed.). Simon & Schuster.

Kantardzic, M. (2011*). Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). Wiley-IEEE Press.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Khan, K., Baharudin, B., Khan, A., & Ullah, A. (2014). Mining opinion components from unstructured reviews: A review. Journal of King Saud University - Computer and Information Sciences, 26(3), 258–275. https://doi.org/10.1016/j.jksuci.2014.03.009

Khan, F., Bashir, S. & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2013.09.004.

Kim, D.-Y., Lehto, X. Y., & Morrison, A. M. (2007). Gender differences in online travel information search: Implications for marketing communications on the internet*. Tourism Management*, 28, 423-433. https://doi.org/10.1016/j.tourman.2006.04.001

Kim, H., Xiang, Z., & Fesenmaier, D. (2015). Use of The Internet for Trip Planning: A Generational Analysis. *Journal Of Travel & Tourism Marketing*, 32(3), 276-289. https://doi.org/10.1080/10548408.2014.896765.

Knowles, A. (2019, September 5). Why negative reviews are essential for companies and consumers. Bazaarvoice. https://www.bazaarvoice.com/blog/negative-reviews-crfa/

Költringer, C., & Dickinger, A. (2015). Analyzing destination branding and image from online sources: A web content mining approach. *Journal of Business Research*, 68(9), 1836–1843. https://doi.org/ 10.1016/j.jbusres.2015.01.011

Krause, M. (2018). How to turn negative online reviews into marketing wins. Retrieved from https://marketingland.com/how-to-turn-negative-online-reviews-into-marketing-wins-248230

Lee, Y. J., & Gretzel, U. (2014). Cross-Cultural Differences in Social Identity Formation through Travel Blogging. *Journal of Travel & Tourism Marketing*, 31(1), 37–54. https://doi.org/10.1080/10548408.2014.861701

Lehto, X., Park, J. K., Park, O., & Lehto, M. R. (2007). Text analysis of consumer reviews: The case of virtual travel firms. In Human Interface and the management of information. Methods, techniques, and tools in information design (pp. 490-499). Springer Berlin Heidelberg.

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social Media in Tourism and Hospitality: A Literature Review. Journal of Travel & Tourism Marketing, 30(1–2), 3–22. https://doi.org/10.1080/10548408.2013.750919

Leung, R. , Schuckert, M., & Yeung, E. (2013*). Attracting User Social Media Engagement: A Study of Three Budget Airlines Facebook Pages*. https://doi.org/10.1007/978-3-642-36309-2_17.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, 29(3), 458–468. https://doi.org/10.1016/j.tourman.2007.05.011

Malik, S. K., & Rizvi, S. (2011). Information Extraction Using Web Usage Mining, Web Scrapping, and Semantic Annotation. 2011 International Conference on Computational Intelligence and Communication Networks. doi: 10.1109/cicn.2011.97

Marine-Roig, E., & Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing & Management*, 4(3), 162–172. doi: 10.1016/j.jdmm.2015.06.004

Market Intelligence and Competitiveness. (n.d.). Retrieved from http://marketintelligence.unwto.org/content/conceptual-framework- UNTWO 2016

Marres, N., & Weltevrede, E. (2013). SCRAPING THE SOCIAL? *Journal of Cultural Economy*, 6(3), 313–335. https://doi.org/10.1080/17530350.2013.772070

Mayer-Schönberger Viktor, & Cukier, K. (2017). *Big data: a revolution that will transform how we live, work, and think*. London: John Murray.

McIntosh, R. W., Goeldner, C. R., & Ritchie, B. J. R. (1995). Tourism: Principles, Practices, Philosophies (7th ed.). Wiley.

Mendes, J.,Matos, N., & Valle, P. O. D. (2012). A model development of relationships between tourism experiences and destination image . In 2nd *Advances in Hospitality & Tourism Marketing and Management Conference* , Corfu, June. ISBN 978-960-287-1393.

Miller, S. H. (2001). “Competitive Intelligence an overview”. *Competitive Intelligence Magazine*, 1(11)

Prescott, J. E., Miller, S. H., & Professionals, T. S. O. C. I. (2001). *Proven Strategies in Competitive Intelligence: Lessons from the Trenches* (1st ed.). Wiley.

Mitchell, R. (2015), *Web scraping with Python: collecting data from the modern web*, O'Reilly Media, Inc.

Moe, W. W., & Schweidel, D. A. (2014*). Social Media Intelligence* (1st ed.). Cambridge University Press.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 6(7): e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers feedback on hotels through data mining: The case of Las Vegas Strip. *Tourism Management Perspectives*, 23, 41–52. https://doi.org /10.1016/j.tmp.2017.04.003

Mudambi,S. ,& Schuff , D. (2010). Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, 34(1), 185. https://doi.org/ 10.2307/20721420

Munar, A. M. (2012). Social Media Strategies and Destination Management. *Scandinavian Journal of Hospitality and Tourism*, 12(2), 101–120. https://doi.org/10.1080/15022250.2012.679047

Munar, A. M., Gyimóthy, S., & Cai, L. (Eds.). (2013). *Tourism social media: Transformations in identity, community, and culture*. Emerald Group Publishing.

Munar, A. M., & Jacobsen, J. K. S. (2014). Motivations for sharing tourism experiences through social media. *Tourism Management*, 43, 46–54. https://doi.org/10.1016/j.tourman.2014.01.012

National Research Council Canada. (2019, June 3). Sentiment and emotion lexicons. Retrieved from https://nrc.canada.ca/en/research-development/products-services/technical-advisory-services/sentiment-emotion-lexicons

Newton, P. W. (2008). *Transitions: Pathways Towards Sustainable Urban Development in Australia* (2008th ed.). Springer.

Norton-Taylor, R. (2012), "Former spy chief calls for laws on online snooping", The Guardian, 24th of April, available at http://www.theguardian.com/technology/2012/apr/24/former-spy-chief-laws-snooping. (accessed on 30.05.2019).

O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management*, 19(7), 754-772. DOI: 10.1080/19368623.2010.508007

Ohbe, T., Ozono, T., & Shintani, T. (2017). A sentiment polarity classifier for regional event reputation analysis. *Proceedings of the International Conference on Web Intelligence* - WI 17. doi: 10.1145/3106426.3109416

Omand, D., Bartlet, J. and Miller, C. (2012), “Introducing Social Media Intelligence (SOCMINT)” in *Intelligence and National Security*, Vol.27. No. 6, December at https://www.researchgate.net/publication/262869934_Introducing_social_media_intelligence_SOCMINT (accessed 21 .05.19)

Paine, D. K. (2011). *Measure What Matters: Online Tools For Understanding Customers, Social Media, Engagement, and Key Relationships* (1st ed.). Wiley.

Pan, B., & Li, X. R. (2011). The long tail of destination image and online marketing. Annals of Tourism Research, 38(1), 132–152. https://doi.org/10.1016/j.annals.2010.06.004

Pan, B., MacLaurin, T., & Crotts, J. C. (2007). Travel Blogs and the Implications for Destination Marketing. *Journal of Travel Research*, 46(1), 35–45. https://doi.org/10.1177/0047287507302378

Pantelidis, I. S. (2010). Electronic Meal Experience: A Content Analysis of Online Restaurant Comments. *Cornell Hospitality Quarterly*, 51(4), 483–491. https://doi.org/10.1177/1938965510378574

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

Parra-López, E., Bulchand-Gidumal, J., Gutiérrez-Taño, D., & Díaz-Armas, R. (2011). Intentions to use social media in organizing and taking vacation trips. *Computers in Human Behavior*, 27(2), 640–654. https://doi.org/10.1016/j.chb.2010.05.022

Patel, N. (2018) Your Business Needs More Negative Reviews. Here's Why. Retrieved from https://neilpatel.com/blog/your-business-needs-negative-reviews/

Patino, A., Pitta, D. A., & Quinones, R. (2012). Social media's emerging importance in market research. *Journal of Consumer Mark*eting, 29(3), 233–237. https://doi.org/10.1108/07363761211221800

Pike, S., & Ryan, C. (2004). Destination positioning analysis through a comparison of cognitive, affective, and conative perceptions. *Journal of Travel Research*. 42(4): 333-342 Destination positioning analysis through a comparison of cognitive https://eprints.qut.edu.au/16910/1/16910.pdf

Povoda, L., Burget, R., Dutta, M. K., & Sengar, N. (2017). Genetic optimization of big data sentiment analysis. 2017 *4th International Conference on Signal Processing and Integrated Networks* (SPIN). doi: 10.1109/spin.2017.8049932

Qu, H., & Lee, H. (2011). Travelers' social identification and membership behaviors in online travel community. *Tourism Management*, 32(6), 1262–1270. https://doi.org/10.1016/j.tourman.2010.12.002

Rezabakhsh, B., Bornemann, D., Hansen, U., & Schrader, U. (2006). Consumer Power: A Comparison of the Old Economy and the Internet Economy. Journal of Consumer Policy, 29(1), 3–36. https://doi.org/10.1007/s10603-005-3307-7

Rezende, S. O. (organizadora). *Sistemas Inteligentes: Fundamentos e Aplicações*. Editora Manole Ltda, 2003.

Rollins, M., Bellenger, D. N., & Johnston, W. J. (2012). Customer information utilization in business-to-business markets: Muddling through process? *Journal of Business Research*, 65(6), 758–764. https://doi.org/10.1016/j.jbusres.2010.12.013

Roque, V., Fernandes, G., & Raposo, R. (2012). Identificação dos Media Sociais utilizados pelas organizações de gestão de destinos: o caso de estudo do destino turístico Serra da Estrela. *Revista Turismo & Desenvolvimento*, ISSN 1645-9261, Nº. 17-18, 1, p. 311-320.

Ross, P., McGowan, C. ,& Styger, L. (2012). A comparison of theory and practice in market intelligence gathering for Australian micro-businesses and SMEs. *19th International Business Research Conference*: Research for Re-thinking, Social Science Research Network. https://doi.org/10.2139/ssrn.2253691

Santos, R. (2009).Conceitos de Mineração de Dados na Web. XV Simpósio Brasileiro de Sistemas Multimídia e Web, *VI Simpósio Brasileiro de Sistemas Colaborativos–Anais*, MM Teixeira, CAC Teixeira, FAM Trinta, e P. PM Farias, Eds, p. 81-124,2009. Retrieved from:<http://www.lac.inpe.br/~rafael.santos/Docs/WebMedia/2009/webmedia2009.pdf

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and Tourism Online Reviews: Recent Trends and Future Directions. *Journal of Travel & Tourism Marketing*, 32(5), 608–621. https://doi.org/10.1080/10548408.2014.933154

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. Journal of Retailing, 80(2), 159–169. https://doi.org/10.1016/j.jretai.2004.04.001

Sheehan, L., Vargas-Sánchez, A., Presenza, A., & Abbate, T. (2016). The Use of Intelligence in Tourism Destination Management: An Emerging Role for DMOs. *International Journal of Tourism Research*, 18(6), 549–557. https://doi.org/10.1002/jtr.2072

Sigala, M. (2014). Customer Involvement in Sustainable Supply Chain Management: A Research Framework and Implications in Tourism. *Cornell Hospitality Quarterly*, 55(1), 76–88. https://doi.org/10.1177/1938965513504030

Silva, T.(2012). Monitoramento de Mídias Sociais. In: SILVA, Tarcízio (Org.). *Para Entender o Monitoramento em Mídias Sociais*. Florianópolis: Bookess, 2012. Disponível em: < http://tarciziosilva.com.br/blog/entenda-o-monitoramento-de-midias- sociais-com-e-book-brasileiro/>. Retrieved 12/10/2018

Sirakaya, E., & Woodside, A. G. (2005). Building and testing theories of decision making by travellers. *Tourism Management*, 26(6), 815–832. https://doi.org/10.1016/j.tourman.2004.05.004

Social Media Research Group. (2016). *Using social media for social research: An introduction*. Digital Education Resource Archive (DERA). http://dera.ioe.ac.uk/id/eprint/266002016

Statista https://www.statista.com/search/?q=Tripadvisor Retrieved 15/09/18.

Statistics, B., Statistics, D., Statistics, V., Collections, G., Gadgets, G., & Gadgets, H. et al. (2018). 35 Amazing TripAdvisor Statistics. Retrieved from https://expandedramblings.com/index.php/tripadvisor-statistics/

Syeda, K. N., Shirazi, S. N., Naqvi, S. A. A., Parkinson, H. J., & Bamford, G. (n.d.).(2017). Big Data and Natural Language Processing for Analysing Railway Safety. *Innovative Applications of Big Data in the Railway Industry Advances in Civil and Industrial Engineering*, 240–267. doi: 10.4018/978-1-5225-3176-0.ch011

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), 267–307. doi: 10.1162/coli_a_00049

Taecharungroj, V., & Mathayomchan, B. (2019). Analyzing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, 75, 550–568. doi: 10.1016/j.tourman.2019.06.02

Thao, N. T. T., & Shurong, T. (2020). Is It Possible for "Electronic Word-of-Mouth" and "User-Generated Content" to be Used Interchangeably? *Journal of Marketing and Consumer Research*. https://doi.org/10.7176/jmcr/65-04

Tang, X.,& Yang, C.(2012). Social network integration and analysis using a generalization and probabilistic approach for privacy preservation. *Security Informatics*, v. 1, n. 1, p. 1-14, 2012. https://doi.org/10.1186/2190-8532-1-7

Tej Adidam, P., Banerjee, M., & Shukla, P. (2012). Competitive intelligence and firm's performance in emerging markets: an exploratory study in India. *Journal of Business & Industrial Marketing*, 27(3), 242–254. https://doi.org/10.1108/08858621211207252

Thomaz, G. M., Biz, A. A., Bettoni, E. M., & Pavan, C. S. (2015). Modelo de monitoreo de las redes sociales para orientar en la toma de decisiones de las destination management organizations. *Revista Brasileira De Pesquisa Em Turismo*, 9(2), 196. doi: 10.7784/rbtur.v9i2.835

Thomaz, G. M., Biz, A. A., Bettoni, E. M., Mendes-Filho, L., & Buhalis, D. (2017). Content mining framework in social media: A FIFA world cup 2014 case analysis. *Information & Management*, 54(6), 786–801. doi: 10.1016/j.im.2016.11.005

Töllinen, A., Järvinen, J., & Karjaluoto, H. (2012). Opportunities and Challenges of Social Media Monitoring in the Business to Business Sector. *The 4th International Business and Social Science Research Conference*, p. 1-14, Dubai, UAE Retrieved https://www.researchgate.net/profile/Heikki_Karjaluoto/publication/265040611_Opportunities_and_Challenges_of_Social_Media_Monitoring_in_the_Business_to_Business_Sector/links/55096e370cf26ff55f857903/Opportunities-and-Challenges-of-Social-Media-Monitoring-in-the-Business-to-Business-Sector.pdf?origin=publication_detail

Tsao, H.-Y., Chen, M.-Y., Lin, H.-C. K., & Ma, Y.-C. (2019). The asymmetric effect of review valence on numerical rating. *Online Information Review*, 43(2), 283–300. doi: 10.1108/oir-11-2017-0307

Turney, P. D. (2001). Thumbs up or thumbs down? *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 0. https://doi.org/10.3115/1073083.1073153

Yoo, K., Lee, K. & Gretzel, U. (2007). The role of Source Characteristics in eWOM: What Makes Online Travel Reviewers Credible and Likeable? In Sigala, M., Mich, L., & Murphy, J. (2007). *Information* *and Communication Technologies in Tourism 2007*: Proceedings of the International Conference in Ljubljana, Slovenia, 2007 (Springer Computer Science) (1st ed.). Springer.

Yoo, K., & Gretzel, U. (2008 a).The Influence of Perceived Credibility on Preferences for Recommender Systems as Sources of Advice. *Information Technology & Tourism*, 10(2), 133–146. https://doi.org/10.3727/109830508784913059

Yoo, K. H., & Gretzel, U. (2008 b). What Motivates Consumers to Write Online Travel Reviews? *Information Technology & Tourism*, 10(4), 283–295. https://doi.org/10.3727/109830508788403114

Young, J.A. (2016, October 16). Using Customer Feedback as a Source of Marketing Research Intelligence. SurveyGizmo. Retrieved from https://www.surveygizmo.com/resources/blog/market-research-intelligence/

Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering- based approach to web advertising. *Artificial Intelligence Research*, 2(1). doi: 10.5430/air.v2n1p44

Vecchio, P. D., Mele, G., Ndou, V., & Secundo, G. (2018). Creating value from Social Big Data: Implications for Smart Tourism Destinations*. Information Processing & Management*, 54(5), 847–860. doi: 10.1016/j.ipm.2017.10.006

Verhagen, T., Nauta, A., & Feldberg, F. (2013). Negative online word-of-mouth: Behavioral indicator or emotional release? *Computers in Human Behavior*, 29(4), 1430–1440. https://doi.org/10.1016/j.chb.2013.01.043

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, 30(1), 123–127. https://doi.org/10.1016/j.tourman.2008.04.008

Xiang, Z., & Gretzel, U. (2010). "Role of social media in online travel information search." *Tourism Management*, Vol.31 No.2, pp.179-188. doi: 10.1016/j.tourman.2009.02.016

Werthner, H. and Klein, S. (1999) *Information Technology and Tourism: A Challenging Relationship*. Springer-Verlag, Wien.

Wurman, R. S. (2001). *Information Anxiety 2* (Hayden/Que) (2nd ed.). Que.

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social Media Mining: An Introduction* (1st ed.). Cambridge University Press.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, 25(6), 13-16. [5678581]. https://doi.org/10.1109/MIS.2010.151

Zeng, L., Li, L., & Duan, L. (2012). Business intelligence in enterprise computing environment. *Information Technology and Management*, 13(4), 297–310. https://doi.org/10.1007/s10799-012-0123-z

Zhang, Z. J. (2011). Customer knowledge management and the strategies of social software. *Business Process Management Journal*, 17(1), 82–106. https://doi.org/10.1108/14637151111105599

Zhu, F., & Zhang, X. (2006). The influence of online consumer reviews on the demand for experience goods: The Case of Video Games. In *Proceedings of the twenty-seventh international conference on information systems* (ICIS). (pp. 367-382)Milwaukee, USA

WTTC 2018 https://www.wttc.org/-/media/files/reports/economic-impact-research/countries-2018/portugal2018.pdf

APPENDIX

I – PRISMA – Elegible Works – KeyWords and Finding

Authors	Title	Keywords	Findings
Sheehan et al. (2016)	The Use of Intelligence in Tourism Destination Management: An Emerging Role for DMOs	Destination Management Organization; knowledge agent; intelligence; stakeholder; collaboration	DMO has evolved from a meaning-centered on marketing (i.e. D. Marketing O.) to a meaning-centered on management (i.e. D Management O). The successful DMO of the future will be an intelligent agent of the destination that is able to identify, engage, and learn from disparate stakeholders both within and outside the destination. It must acquire, filter, analyze, and prioritize data and information from various sources to create knowledge that can be used to fulfill its role in destination management. DMO must gain knowledge about the competitive environment, opportunities and threats, and trends that will change the future competitive landscape.
Dey et al. (2011)	Acquiring Competitive Intelligence from Social Media	Web, Competitive Intelligence, Social Media, Decision Making.	This paper discusses methodologies to obtain competitive intelligence from different types of web resources including social media using a wide array of text mining techniques. It provides some results from case-studies to show how the gathered information can be integrated with structured data and used to explain business facts and thereby be adopted for future decision making.
Buhalis & Foerste (2015)	SoCoMo marketing for travel and tourism: Empowering co-creation of value	Personalization Mobile context-awareness Social media marketing Mobile technologies Co-creation	This paper connects the different concepts of context-based marketing, social media, and personalization, as well as mobile devices. Contextual information is increasingly relevant, as big data collected by a wide range of sensors in a smart destination provide real-time information that can influence the tourist experience The proposed Social Context Mobile (SoCoMo) conceptual model explores the emerging opportunities and challenges for all stakeholders.

Thomaz et al. (2017)	Content mining framework in social media: A FIFA world cup 2014 case analysis	Social media, Content mining, Twitter, Tourist services, Brazil FIFA world cup 2014	This paper proposes a social media content mining framework that consists of seven phases. The framework is effective to collect relevant content and identify popular topics in social media toward strategic and operational tourism management.
Dai, Sutinen & Kakkonen (2010)	MinEDec: a Decision-Support Model That Combines Text-Mining Technologies with Two Competitive Intelligence Analysis Methods	Decision support system, competitive intelligence, text mining, the Five Forces framework, SWOT analysis	Proposal for the integration of two Competitive Intelligence Analysis (FFA and SWOT) to monitor and analyze the competitive environment of businesses, making analysis with various text-mining technologies in a decision support model Mining Environment for Decisions (MinEDec).
Thomaz et al. (2015)	Social media monitoring model to guide decision making of Destination Management Organization	DMO, Social Media Monitoring, decision making	The objective of this study was to present a social media monitoring model to support decision making by DMO. The proposed model has been tested in the pre-period and during the 2014 FIFA World Cup.
Leung et al.(2013)	Social Media in Tourism and Hospitality: A Literature Review	Social media, Web 2.0, journal review, tourism research, hospitality research, consumers, suppliers	This study reviews and analyzes all social media-related research articles published in academic journals of Tourism and Hospitality from 2007 to 2011. This paper details the role of the consumer and social media provider. Supplier-related studies have concentrated closely on promotion, management, and research functions, but few discussed product distribution. Research findings demonstrate the strategic importance of social media for tourism competitiveness.
Gémar & Jiménez-Quintero (2014)	Text mining social media for competitive analysis	Competitive intelligence, social media, text mining, hotel industry, financial performance.	This study used a text-mining tool to analyze the primary social media sites, with a focus on a sample of hotels. The dimensions analyzed were sentiments and reach. A dependence was found between several variables obtained through text mining and financial performance. The results indicate that the analysis of

			social media using these techniques can be a method to improve financial performance.
Franzoni & Bonera (2019)	How DMO Can Measure the Experiences of a Large Territory	DMO; experiences; reviews; TripAdvisor; methodology; descriptive analysis; sentiment analysis; content analysis.	Destination Management Organization (DMO) can collect useful information to make decisions and take action to protect and/or increase the competitiveness of the destination. They collected opinions from TripAdvisor to find corrective measures to be taken to preserve or enhance the interest of a tourist destination. They used empirical observation.
Moro et al. (2017)	Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip	Customer feedback; customer reviews; online reviews; knowledge extraction; data mining; modeling; sensitivity analysis	This paper presents a data mining approach for modeling TripAdvisor scores using reviews published in 2015 for the 21 hotels located on the Strip, Las Vegas. Nineteen quantitative characteristics were used to characterize the evaluations, hotels, and customers were prepared and used to feed a support vector machine to model the score. Sensitivity was applied on the model to extract useful knowledge translated into the relevance of the resources to the score. It was found that seasonality, day of the week, and user profile on TripAdvisor influenced the scores.
Vecchio et al. (2018)	Creating value from Social Big Data: Implications for SmartTourism Destinations	Big Data Business analytics Decision making Smart tourism creation Social media measurement	The article explores a set of regional tourist experiences related to a region of Italy, and through the KeyHole tool, it extracts patterns and generates opportunities for value creation generated by Big Data in tourism.The findings present and discuss evidence in terms of improving decision making, creating marketing strategies with more personalized offers, transparency, and confidence in dialogue with customers and stakeholders and the emergence of new business models.Finally, implications are presented for researchers and professionals interested in managing. Big Data exploration in the context of information-intensive industries such as Tourism.
Schuckert et al. (2015)	HOSPITALITY AND TOURISM ONLINE REVIEWS: RECENT TRENDS AND FUTURE DIRECTIONS	Data Mining Tourism and Hospitality Literature Review	This study analyzed articles related to online evaluations of tourism and hospitality published in academic journals between 2004 and 2013. Based on keyword-oriented research and content analysis, 50 articles were identified as relevant. As a result, it was found that more than half of the analyzed articles focus on hotels and apply empirical methods based on secondary data. Another finding is that Data Mining can make up the quantitative part of the search and

			may find product defects or service failures among numerous online reviews.
Taecharungroj & Mathayomchan (2019)	Analyzing TripAdvisor reviews of tourist attractions in Phuket	Online reviews TripAdvisor Destination marketing Destination management Latent Dirichlet allocation Naïve Bayes	This search drew the online reviews present on TripAdvisor for tourist attractions, Phuket, Thailand. Online assessments were analyzed using two machine learning techniques: latent Dirichlet allocation (LDA), which helps researchers determine the dimensions of each type of attraction, and naive Bayes modeling used in the Machine Learning area (Machine Learning) to categorize texts based on the frequency of the words used. This research also resulted in two practical tools - dimensional salience-valence analysis (DSVA) and lexical salience-valence analysis (LSVA) - and used them to suggest actions for the Thai DMO.
Chang et al. (2019)	Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor	Sentiment analysis Hospitality Natural language processing Social media analytics Visual analytics Google trends TripAdvisor	The paper proposes an integrated framework that includes data tracker, data pre-processing, construction of sentient sensitive trees, convolution tree kernel classification, aspect extraction, and category detection and visual analysis to gain insights into ratings and reviews of hotels. The results achieved show that the approach surpasses known methods of classifying feelings. The analysis revealed that business travelers tend to give lower ratings, more often use negative keywords, such as "rude", "terrible", "horrible", "broken" and "dirty", to express their dissatisfied emotions with the hotel stay. Couples tend to give higher ratings.
Marine-Roig & Clavé **(2015)**	Research paper Tourism analytics with massive user-generated content: A case study of Barcelona	User-generated content Smart city Smart tourism destination Big data Business intelligence	Through analysis social media - review over 100,000 online travel reviews (OTRs) and relevant travel blogs written in English by tourists who have visited the city in the past 10 years, the authors propose a methodology that facilitates the collection, cleaning and analysis bulk UGC related to tourism from the most appropriate sources. The results obtained help to define the transmitted image of the city. It is also used to extract business intelligence (BI) from OTRs on the main attraction of Barcelona - La Sagrada Família.
Amadio & Procaccino (2016)	COMPETITIVE ANALYSIS OF ONLINE REVIEWS USING EXPLORATORY TEXT MINING	Text mining, Online reviews, Competitive analysis, Visual analytics, ReviewMap, SWOT	The study used the text mining/visualization tool, ReviewMap intending to extract information in the comments written by customers of three competing hotels. An application was also a SWOT analysis. The approach was exploratory, whose objective was to determine Usable competitive intelligence can be found in a typical collection of online analytics from a set competing hotels. The SWOT analysis provided by the data extraction revealed the strengths, weaknesses, opportunities and threats revealed several promising competitive actions for the hotels in the study.
Dickinger & Mazanec (2015)	Significant word items in hotel guest reviews: A feature	Text mining, Hotel reviews, Social media,	This document shows text mining and provides information on reviewing reviews published on TripAdvisor. A supervised classification SVM was used, especially for high dimension data. This machine

	extraction approach	Quality management, Support vector machine	identifies keywords that represent the most positive and negative reviews. These terms can be used as an early warning system by managers to efficiently monitor customers' online dialogue with the hotel.
Hu et al. (2019)	What do hotel customers complain about? Text analysis using structural topic model	Online hotel reviews Customer dissatisfaction Structural topic model Text mining Tripadvisor	This study adopts a new text analysis method of a structural topic model to analyze 27,864 hotel reviews in New York City. The ability to understand the causes of customer complaints is fundamental for hotels to improve the quality of service, customer satisfaction. It was found that customer complaints vary between different categories of hotels. The results indicate that complaints from high-end hotel customers are mainly related to service problems, while low-cost hotel customers are often troubled by problems related to facilities.
Berezina et al. (2015)	Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews	Hotel reviews, Text mining, User-generated content, Customer satisfaction, Dissatisfaction	The authors made a text mining approach followed and online analyzes of satisfied and dissatisfied customers were compared. The survey revealed that satisfied customers who want to recommend a hotel to others refer to intangible aspects of their business more often than dissatisfied customers. The study has managerial implications related to the understanding of satisfied and dissatisfied customers through the use of text mining and hotel ratings on review sites.
Tsao et al. **(2019)**	The asymmetric effect of review valence on numerical rating a viewpoint from sentiment analysis of users of TripAdvisor	On line review Text Mining Asymmetric effect Brand Strength	The authors performed a sentiment analysis using text mining, extracting a set of data from the TripAdvisor website. This study found that there is an asymmetric relationship between the valence of the (verbal) review and the numerical classification. For a stronger brand, the content of negative reviews will have a greater impact on numerical ratings than the content of positive reviews, while for a weaker brand, the content of positive reviews will have a greater impact on numerical ratings than the content of reviews negative.

II- CODE R for Scrapping and Analysing Attractions in TripAdvisor

"**Text analysis of Mosteiro dos Jeronimos Reviews from Trip Advisor**"

author:"MMAF"

output :word_document: default   pdf_document: default

html_document: default   --- ```{r setup, include=FALSE}

knitr::opts_chunk$set(echo = F,warning = F,message = F,cache = F,fig.align = 'center') options(scipen = 999)

library(rvest) – Easily harvest (Scrape) Web Pages

library(cowplot) - Streamlined Plot Theme and Plot Annotations for 'ggplot2'

library(dplyr) - Essential shortcuts for subsetting, summarizing, rearranging, and joining together data sets

library(syuzhet) - Extracts Sentiment and Sentiment-Derived Plot Arcs from Text

library(sentimentr) - Calculate Text Polarity Sentiment

library(data.table) - Extension of 'data.frame'

library(tidytext) - Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools

library(tm)- Text Mining Package

library(wordcloud)- Create Word Clouds

library(tidyverse) - Simple, Consistent Wrappers for Common String Operations

library(stringr) - Simple, Consistent Wrappers for Common String Operations

library(udpipe)- Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit

library(knitr)- A General-Purpose Package for Dynamic Report Generation in R

library(kableExtra) - Construct Complex Table with 'kable' and Pipe Syntax

library(webshot) - Take Screenshots of Web Pages

library(htmlwidgets)- HTML Widgets for R

library(gridExtra)- Miscellaneous Functions for "Grid" Graphics

library(ggthemes) - Extra Themes, Scales and Geoms for 'ggplot2'

library(quanteda) - Quantitative Analysis of Textual Data

library(topicmodels) - Provides an interface to the C code for Latent Dirichlet Allocation (LDA) models and Correlated Topics Models (CTM) by David M. Blei

library(flextable) - Functions for Tabular Reporting

library(pander)- An R 'Pandoc' Writer

'https://www.tripadvisor.in/Attraction_Review-g189158-d195318-Reviews-Mosteiro_dos_Jeronimos-Lisbon_Lisbon_District_Central_Portugal.html'

~~~
```{R initial setup, echo = T,eval = F}

URL <- 'https://www.tripadvisor.in/Attraction_Review-g189158-d195318-Reviews-

Mosteiro_dos_Jeronimos-Lisbon_Lisbon_District_Central_Portugal.html'   html

<- read_html(url)

review_count<-html%>%

html_nodes('.pagination-details')%>%

html_text()

review_count <- strsplit(review_count,'of')[[1]][2]

review_count <- gsub('reviews','',review_count)

review_count  <- as.numeric(gsub(',','',review_count))

reviews<-html%>%

html_nodes('.partial_entry')%>%

html_text()

num  <- seq(5,review_count,5)
~~~

the URL increments by 5 each time we move from page n to page n+1. Using this information I was able to write code that loops over all pages and scrape all available reviews.

~~~
```{R review scraping, eval=F}

for (i in 1:length(num)){ revurl

<- paste0('https://www.tripadvisor.in/Attraction_Review-g189158-d195318-Reviews-

or',num[i],'-Mosteiro_dos_Jeronimos-Lisbon_Lisbon_District_Central_ Portugal.html')

html <- read_html(revurl)

reviewed  <- html%>% html_nodes('.partial_entry')%>%  html_text()

reviews  <- c(reviews,reviewed) }

write.csv(data.frame(reviews=reviews),'reviews1.csv',row.names = F)

``` ```{R review filtering} reviews  <- read.csv('reviews1.csv')

reviews <- gsub("[[:punct:]]", "", reviews$reviews)

reviews <- reviews[!duplicated(reviews)]
~~~

``` ```{R} pander(head(as.data.frame(reviews)),caption = 'Example of scraped reviews',split.cell = 120, split.table = Inf)

``` ```{R,fig.width = 10}

Sentimentr

sentiment <- sentiment(as.

sentiment_nrc   <- get_nrc_sentiment(as.character(reviews))

sentiment_nrc <- data.frame(score = colMeans(sentiment_nrc))

sentiment_nrc$sentiment <- row.names(sentiment_nrc)

sentiment <- sentiment%>% group_by(element_id)%>% summarise(sentiment=mean(sentiment))

reviews_sentiment

<- data.frame(review = reviews,sentiment)

p1 <- ggplot(data = sentiment_nrc,aes(x = reorder(sentiment,score),y = score)) + geom_bar(stat = 'identity',fill = 'limegreen',alpha = 0.6) + xlab('Sentiment') + theme(axis.text.x = element_text(angle = 300))

P2 <- ggplot(data = reviews_sentiment,aes(x = sentiment)) + geom_density(fill = 'limegreen',alpha = 0.6) +xlab('Sentiment') plot_grid(p1,p2,labels = 'AUTO')

``` **### Examples of reviews classified as negative**

```{R}

pander(head(reviews_sentiment %>%

    arrange(sentiment) %>   select(review,sentiment),4),split.cell = 120, split.table = Inf)

```'

Examples of reviews classified as positive   ```{R}

pander(head(reviews_sentiment%>%arrange(desc(sentiment))%>
select(review,sentiment),4),split.cell = 120, split.table = Inf)

``` To properly explore the textual data, some preprocessing steps must be taken. These are as follows   ```{R}   reviews

<- tolower(as.character(reviews)) tokenized <- tokens(as.character(reviews), what = 'word', remove_number = T,remove_punct = T, remove_symbols = T,remove_hyphens = T)

tokenized

<- tokens_select(tokenized, stopwords(), selection = 'remove') tokenized

<- tokens_select(tokenized,stop_words,selection = 'remove')

token_list <- tokens_ngrams(tokenized,n = 1:2,concatenator = ' ')

token_list <- dfm(token_list)

token_freq <- textstat_frequency(token_list)

bigram_freq <- token_freq[grepl(' ',token_freq$feature),]

monogram_freq <- token_freq[!grepl(' ',token_freq$feature),]

a <- get_sentiment(monogram_freq$feature,method = 'bing')

b <- get_sentiment(bigram_freq$feature,method = 'bing')

monogram_freq <- data.frame(monogram_freq,a)

bigram_freq <- data.frame(bigram_freq,b)

``` ### Commonly occurring words in Negative Review

```{R} Eng

<-udpipe_load_model('english-ewt-ud-2.3-181115.udpipe')

annotations

<- as.data.frame(udpipe_annotate(eng,unlist(monogram_freq[monogram_freq$frequency >=5,1]))) frequent_monograms <- monogram_freq[monogram_freq$frequency >=5 & a < 0,]

annotations <- annotations[annotations$token_id == 1,c(4,8)]

frequent_monograms <- merge(frequent_monograms,annotations,by.x = 'feature',by.y = 'sentence')

wordcloud(frequent_monograms[,'feature'],frequent_monograms[,'frequency'],

rotateRatio = 0.6)

``` ### Commonly occurring words in Positive Reviews

```{R} annotations

<-as.data.frame(udpipe_annotate(eng,unlist(monogram_freq[monogram_freq$frequency >=50,1])))

frequent_monograms <- monogram_freq[monogram_freq$frequency >=50 & a > 0,]

Annotations

<- annotations[annotations$token_id == 1,c(4,8)]

frequent_monograms <- merge(frequent_monograms,annotations,by.x = 'feature',by.y ='sentence')wordcloud(frequent_monograms[,'feature'], frequent_monograms[,'frequency'], shape = 'circle', rotate Ratio = 0.6)

``` ### What descriptive monograms are the most commonly used in all reviews?

```{R} annotated_monograms <- as.data.frame(udpipe_annotate(eng,unlist(monogram_freq$feature)))

annotated_monograms <- group_by(annotated_monograms,sentence) %>%slice(1) %>%

select(sentence,upos)

monograms <- merge(monogram_freq,annotated_monograms,by.x = 'feature',by.y = 'sentence')

monograms %>% filter(upos == 'ADJ') %>% arrange(desc(frequency)) %>%top_n(20,frequency) %>%ggplot(aes(x = reorder(feature,frequency),y = frequency)) +

geom_bar(stat = 'identity',fill = 'limegreen') +coord_flip() + theme_base() + xlab('')

``` ```{R} Bigrams <- bigram_freq bigrams_positive <- bigrams[bigrams$b > 0,]

bigrams_negative <- bigrams[bigrams$b < 0,]

``` ## What are the most common negative bigrams?

```{R} wordcloud(bigrams_negative[bigrams_negative$frequency >= 2,'feature'],

bigrams_negative[bigrams_negative$frequency >= 2,'frequency'])

```## What are the most common positive bigrams?

```{R}wordcloud(bigrams_positive[bigrams_positive$frequency >= 5,'feature'],bigrams_positive[bigrams_positive$frequency >= 5,'frequency'])
```