Università
Ca'Foscari
Venezia

DEPARTMENT OF ECONOMICS

Master's Degree in
Global Development and Entrepreneurship

Final Thesis

**The central role of customers in web 2.0 applied to tourism:
an analysis of Italian tourist destinations based on TripAdvisor reviews.**

Supervisor:
Ch. Prof. Dario Bertocchi

Graduand:
Tania Rossi
Matriculation Number 863797

Academic Year:
2020-2021

Qualsiasi mezzo si utilizzi, qualunque sia la meta, metterci in gioco, ripartire ogni volta, dare fiducia all'inspiegabile curiosità che ci assilla è l'unica vera strada che ci resta da percorrere.

Christian Carosi, *Viaggiatori intelligenti*

Contents

Introduction

Tourism and communication are two inseparable concepts that complement each other and progress hand in hand, particularly thanks to the progress of new technologies. Both sectors have evolved qualitatively and quantitatively, renewing themselves completely at a structural and operational level, as well as at an IT and technological level. In the second half of the 1990s, the world of travel and tourism was hit by a rapid technological acceleration, in which new communication tools and innovative development processes spread across all sectors. The advent of Internet technology has therefore represented an entirely new phenomenon, potentially capable of disrupting and revolutionising the entire tourism sector.

The explosion of the web has enabled supply and demand to create and globalise a flow of information that can run freely through dialogue platforms with global reach. By transferring social interactions into the virtual world, various communication channels have developed that have allowed millions of people to interact with ease, thus eliminating the geographical distance between them. The highly participatory and collaborative nature of Web 2.0 means that the final consumer has become a protagonist with a central role in the construction of the Internet: the context has thus become democratically accessible to all, and the focus of analysis has shifted to interpersonal relationships and communities.

This social component is essential, and this context turns out to be an "architecture of participation", in which interaction between subjects is the main asset from which collective intelligence is created. With reference to the tourism industry, a strategy was developed in which understanding latent needs and enhancing relations with the tourist was more appropriate to the sector. One of the most innovative aspects in this new set-up derives from the web's ability to rebalance the information asymmetry between supply and demand: the overload of information is now accessible to all, and the consumer is free to collect qualitative and quantitative data from different

sources, compare them, and then build an opinion and make the most appropriate choice for his or her needs. In order to improve the usability of the Internet and make the whole tourist offer more transparent, a number of sites have sprung up spontaneously where consumers publish ratings and scores on goods and services based on their own personal experience.

Known as rating sites, these tourism portals act as neutral information sources, where peer-to-peer relationships form the basis of the data exchange: TripAdvisor represents the icon of the traveller community in Web 2.0. Born as a digital and interactive variant of paper guides, it is constantly growing thanks to UGC, the contents generated by the users registered in the portal, who, by reporting personal experiences and impressions on the web, have the power to influence the choice of a potential customer who is looking for suggestions online. It should be stressed that the purpose of those who review a tourist facility on a travel portal is not to promote or denigrate the online reputation of an establishment in order to make money from it, but an attempt to push the reader in the direction they consider most appropriate. In most cases, reviewers share their consumer experience so that it can be helpful to other travellers, simply by highlighting the positive aspects and discovering the negative ones.

"Information is the oil of the 21st century, and Analytics is the combustion engine" [45]. In the world in which we now live, we are subjected to an uninterrupted flood of data, which accompanies us on a daily basis and represents an invaluable treasure, so much so that it has been called the new oil, capable of profoundly changing individual and collective behaviour, with important economic and social consequences. economic and social consequences.

There are several reasons for this explosion of information. Certainly, the increase in the world population has been accompanied by an increase in the number of people owning electronic devices, with capabilities that improve over time (especially smartphones and computers), through which they interact with others.
Technological developments have greatly driven this explosion with increasingly efficient software and storage systems that have greatly reduced the cost of transmitting and capturing shared data. Finally, this information is being applied in different domains, more frequently and widely than in the past: people use it to regulate certain aspects of their lives; companies aggregate it to optimise numerous activities and transform their sectors; the State analyses it to guide policy choices and better understand current and

future macroeconomic trends. Big Data is also widely used in, and has revolutionised, the tourism sector, especially for data from UGC sites.
It is important that this data is extracted and analysed in order to become a resource for tourism operators.

The intention of this thesis is to provide an overview of the evolution of the tourism industry with the advent of the Internet and, consequently, the change in the tourist's decision-making process. In fact, tourists are increasingly influenced by the web and user-generated content (UGC) when inspiring, planning and booking their trip. The maximum expression of this phenomenon are online reviews. Today, in fact, a DMO in its destination marketing strategies is increasingly making use of web reputation analysis as a means of verifying how the market perceives and evaluates its destination and, as a result, the image it has of it.

In the first part of the paper, the evolution of the tourism market to what we know today is explained, where UGC sites, where people can leave their opinions on what they have visited, play an important central role. Hence the importance of rating sites, in particular TripAdvisor, the world's most famous and widely used platform for this purpose. In this first chapter, the site and the reviews found on it are analysed, with a focus on fake reviews.

The second part of the thesis introduces Big Data and describes its context, the change in our society that led to the emergence of the phenomenon. It analyses the characteristics, the "4Vs" (volume, variety, speed and value), highlighting the advantages that the use of data can bring, especially to the tourism sector. The methodology used in this paper to extract the data is Web Scraping, this system allows the huge amount of data present online to be made usable, so that it can be processed. However, this technology is quite new and therefore not very well regulated, which is why we look at the legal problems associated with it in the following chapter.

The last chapter puts into practice what has been presented in the literature of the previous two chapters. an analysis of the reviews left on TripAdvisor for tourist attractions in all Italian locations was carried out. In this chapter the objective of the research is described, how the web scraping technique was used to create the database, the analysis method used to obtain the necessary data to obtain the desired results.
This project is concluded by examining and interpreting the results of the analysis and explaining how this analysis and these results can be used and the policy implication.

Chapter 1

The world of tourism in the era of digitalization

In this chapter we are outlining the world of tourism, from its birth to its evolution to how it is today. With a closer look at the Italian market.

1.1 Tourism market

Tourism is an important added value to the economy of a city, a region and therefore a country. Consequently, the development of this sector can be decisive in improving the conditions of a place. Indeed, tourism has taken on a new importance on the international scene in recent years because of both its increased size and the irreplaceable contribution it can make to balanced economic, cultural and environmental development. There is no country in the world that does not have tourist potential, which can be cultural, gastronomic, sporting, entertainment, natural, conference-related and so on.
Tourism can bring many benefits if it is properly developed, it can bring valuable currency, employment, income, investment and allows diversification of the economy. In addition to economic aspects, it also has social and cultural impacts, which can be both positive and negative.

1.1.1 Tourism phenomenon

The term tourism was born at the turn of the seventeenth and eighteenth centuries, during which it was used to define the social habits of the wealthier classes: the young European aristocrats of the time were used to travel through the countries of Europe. These journeys of exploration were un-

dertaken for educational, cultural or leisure purposes and understandably served as an element of social distinction. Over the years and the advent of the first industrial revolution, the practice of tourism spread to the less wealthy classes and soon became a personal need, practicable by all social classes. The concept thus lost its character of elite and gradually turned into a mass activity: tourism is since then an effective response to the need to escape from the place of residence and work.

The interest in tourism on the part of scholars from different disciplines (economics, sociology, social psychology, statistics, etc..) and its continuous evolution, has led to the succession of various definitions and adjustments, sometimes conflicting with each other.

A definition that constitutes the starting point for all social sciences that scientifically analyse the tourism phenomenon is that proposed by the World Tourism Organisation, according to which:
" Tourism is a social, cultural and economic phenomenon which entails the movement of people to countries or places outside their usual environment for personal or business/professional purposes. These people are called visitors (which may be either tourists or excursionists; residents or non-residents) and tourism has to do with their activities, some of which involve tourism expenditure." [1].

While on the one hand it is individuals, with their movements, who fuel the existence of tourism in economic terms, on the other hand the phenomenon manages to influence the entire community: it facilitates the exchange of information and knowledge between individuals and supports the creation of an interpersonal network [43]. Although the social approach is important in the study of tourism, the economic sphere prevails, emphasising how tourism has succeeded in bringing fruitful economic and financial benefits to communities that have invested resources and space to support the phenomenon itself.

The sector is nowadays a key sector for the economy of a country and is undergoing year after year strong processes of change and restructuring. New tourism professions are developing thanks to technological innovations and new trends to be analyzed and studied. Tourism is, in fact, a versatile industry that uses services and infrastructures of every single city and involves a large percentage of the world's population.

[1] `https://www.unwto.org/glossary-tourism-terms`

The tourism sector plays a central role at world level because it contributes strongly to the global economic growth process. In fact, it is one of the top five economic sectors in the world. It is important to look at the development of tourism globally over the last twenty years as this period has seen a huge expansion, thanks to income growth in emerging economies, reduced transport costs and the emergence of new destinations.

These are the elements that have mainly contributed to the increase in the number of tourists and have led to estimates of further significant growth in tourism expenditure over the next two decades. In recent years, global tourism has accounted for 10% of GDP, 7% of exports and 1 in 11 jobs. Comparing the number of tourists over time shows that in 1950 there were 25 million tourists, in 2015 there were 1186 million and in 2030 there are expected to be 1.8 billion[14].
Europe is the most popular destination, with around 620 million visitors per year, followed by Asia, America, Africa and Oceania. Europe, however, has experienced limited growth in recent years while other continents have exhibited higher rates of higher growth rates.

Tourism requires free time, income and freedom of movement. It should be noted, that since March 2020, following the rapid spread of Covid-19, international tourism has essentially stopped. The economic crisis has reduced the availability of income and increased uncertainty about the future, both of which have a negative impact on susceptibility to travel[6]. This has been compounded by government restrictions and public fear.

1.1.2 Tourism in Italy

Italy is one of the most touristic countries in the world, especially thanks to the enormous artistic and cultural heritage it possesses. It is the country with the highest number of places recognized as World Heritage Sites. There are, in fact, 54 UNESCO sites out of the 1092 UNESCO sites recognized worldwide, 5000 cultural sites including monuments and archaeological areas, 3,609 museums, 46,025 listed architectural properties and 34,000 places of entertainment.[41]

Some fundamental stages in the development of tourism in Italy can be identified. In the 1980s, when tourism was limited to a few international destinations, Italy was in second place in the ranking of world tourism spend-

ing. Before, in fact, there was only the United States. Later on, however, Italy's position, like that of other mature destinations, deteriorated due to the great expansion of tourism on a global level. Italy, therefore, went from 8% of global tourism spending in the 1980s to 3.4% in 2010. The deterioration, moreover, has been more significant for our country than for the main European competitors.

Since 2010, the first signs of recovery have emerged, as foreign tourist spending has once again begun to grow at a high rate, averaging 4.3% per year. This recovery occurred thanks to the development of geopolitical issues that discouraged travel to countries with a high risk of terrorism, improved price competitiveness and the growth of interest in cultural travel in Italy by foreign tourists. Since 2015, moreover, the first signs of recovery in domestic tourism have emerged, which in the years of the crisis had deteriorated more than international tourism. Recent years have been characterized by a strengthening of tourism especially thanks to tourists from non-European countries, which have increased from 37% in 2010 to 41.5% in 2017. In particular, foreign tourists from the United States, Canada, Australia, Japan and China have increased. To be considered, moreover, that the expenditure incurred by these tourists is above average and increasing over time. Considering only Europeans, the majority of tourists come from Germany and France. It should be taken into consideration that tourists coming from European Union countries are favoured by the presence of the common currency [4], and the possibility of moving freely between States.

Tourism in Italy generally has two main characteristics:

- Short trips to many locations;
- Cultural trips. In fact, 60% of foreign tourists' trips to Italy are cultural trips. This aspect is linked to the fact that a peculiarity of Italy is represented by the presence of a unique artistic and cultural heritage, which is appreciated especially by those who come from abroad and who are visiting Italy for the first time.

Italy is characterised by a large number of art cities, museums and archaeological sites throughout the country. However, tourism is characterised by a strong concentration in the most attractive areas. This is demonstrated by the fact that the first 20 museums of the 5,000 Italian museums comprise 30% of the annual visits. Museums represent a major problem as they have a low number of visitors and relatively low revenues compared to the main

European countries. The cause of this problem can be found in the huge offer and in the gaps concerning the management of the facilities.

In the field of management, it has to be considered that the capacity of exploitation is very limited and that many small museums are not known by potential investors. In order to be able to valorise museums, public transfers are important, which allow them to contribute to the promotion and preservation of their heritage, as well as resources obtained from the participation of private individuals. In 2014 an important reform was introduced, which provided for an accreditation system for museums, increased their autonomy and improved coordination between structures. In the same years, regional clusters were created, which allowed for the development of economies of scale for the management of facilities.

The accommodation facilities in Italy are numerous and for this reason our country is at the first place in the ranking of the number of accommodation facilities present in different countries. However, Italy is second in terms of number of beds because in first place there is France. The number of beds, however, would be much higher if those offered by private companies were also considered. These beds are not usually taken into consideration because very precise data are not available.

Comparing Italian accommodation facilities with those in other European countries, similarities and differences emerge. The similarities include the size of the structures and the characteristic of seasonality. Differences, on the other hand, include the homogeneous distribution of facilities in our country. This depends on the low concentration of the Italian population and on the great diffusion of tourist areas in the Italian territory. Moreover, structures are characterised by a level of bed utilisation and operating profitability that is slightly lower than that of structures in other European countries, and by a slightly higher added value per employee, which increases as the size of the enterprise increases.

Two elements have emerged in the last period:

- An increase in the number of beds in accommodation facilities other than hotels. In particular, there has been an increase in the number of beds in agritourisms and bed and breakfasts;

- An increase in the quality of hotels. In fact, the number of one- and two-star hotels has decreased, the number of three-star hotels has remained

unchanged and the number of four- and five-star hotels has increased. The total number of hotels has remained constant.

1.2 Digital innovation and new needs

The era of Web 2.0 began in 2004 when for the first time Tim O'Rielly [2], during a series of conferences on the new applications of the Internet, used the term to indicate the turning point that the network had reached. From this moment on, the term has entered everyday language to mean a new way of using the network and technologies, as opposed to the previous Web 1.0. If in the nineties the web was static and based on an asymmetrical model in which there was a clear distinction between producers and consumers of services, starting from 2004 a new dynamic web model began to emerge, based on the sharing of resources and the free creation of content by users. From consumer of information and services, the user has become both producer and consumer so prosumer. The term is given by the fusion of two words, product and consumer to define the consumer who participates in the construction of the product. A new figure is thus delineated that replaces that of the traditional consumer, easily influenced by the marketing strategies of the company. The consumer ceases to be a passive figure and becomes an active subject, which is not limited to the use of content, but becomes the player who personally generates the content and shares the information.

Web 2.0 is therefore based on equal models, such as peer-to-peer (P2P) architectures in which the prosumer no longer has to limit himself to reading static content produced outside the web (offline) but can publish his own ideas, share them freely, collaborate with other users and use dynamic content produced by others in the network. It is a much simpler and easier to use web, composed of a series of interactive applications in which there is an active participation of the user aimed at producing dynamic content. Web 2.0 can be considered a second generation of web-based services, based on online collaboration and sharing among users. Consumers no longer depend on information provided by companies but increasingly rely on unfiltered content produced by users [38].

The main principles of Web 2.0 are collaboration, sharing, and user participation in generating content. For this to summarize has been defined a:

[2] Tim O'Reilly is the founder, CEO, and Chairman of O'Reilly Media

- social web: because it is formed not only by a network of pages but also by a network of people who interact with each other (e.g. social network);

- participatory web: because there are user-generated contents, that through new tools, can easily publish and spread contents (e.g. blogs);

- collaborative web: in which contents are not necessarily generated by a single individual but are the result of the work of several users who collaborate and contribute to the production. This is possible through the tools made available in Web 2.0 in which users can work together (e.g. Wikipedia).

Among the main advantages offered by Web 2.0 are the uniformity of the interface on any device, the greater reliability and ubiquity of data that can be accessed through any device. On the other hand, it is possible to identify some disadvantages due to the constant need for connection, the environmental impact they produce and the necessary security of the data stored [32].

In this scenario of changes and innovations lies Travel 2.0. This term refers to a new way of planning a trip inspired by the principles of Web 2.0, that is based on a participatory and collaborative approach of the user that affects the tourism industry in a new way of traveling and organizing the trip. As already mentioned, even in tourism, the consumer is evolving and changing methodologies to reach the product, becoming prosumer: the 2.0 tourist is able not only to read but also to generate content.

According to the 2019 research of Digital Innovation in Tourism Observatory of the Politecnico of Milan, 97% of Italian digital travellers inquire online before buying and 85% conclude online at least one purchase (accommodation, transport, activity or experiential service). Supply chain companies can no longer ignore the use of digital tools in their relationship with these customers: in 2019, the digital tourism sector grew by 9% and reached €15.5 billion (driven by the use of mobile, with +32%). As far as mobile is concerned, smartphones are used by 90% of tourists in the various stages of the tourist journey, in particular for searching for information (71%), purchasing services before travelling (33%) and sharing on social networks and through reviews (33%). The most used Apps are those for searching and booking restaurants (41%) and guides to the location and territory (35%).

In particular, 2020 and 2021 are years of great change, including for tourism. The research carried out by the Digital Innovation in Tourism Observatory of the Politecnico of Milan as shown that the digital sector of tourism dropped by more than 9 million euros in 2020, 60% less than the previous year. This negative downturn due to the heavy restrictive measures can be seen as an opportunity for innovation, also in the way of thinking about holidays and tourism.

Innovations on the digital side are the trends that have been most evident over the last two years.
One of these is the digital journey. In 2019 about 8% of Italian facilities offered payment solutions usable by mobile and online check-in, now the percentage of accommodations has risen to 30%, as well as the integration of chatbot in the sites of facilities has risen from 2% to 14%, and the increase of virtual room tours has risen to 13% of the total. These increases indicate how the digital approach in the world of tourism is actually becoming more and more relevant: offering customers more information and tools to make payments and bureaucracy easier and faster is the way to increase turnover, because it is what the market is asking for.

Another innovation in the tourism sector is the neverending tourism. This is a neologism that sums up a mode of tourism that starts before the trip and continues after the return home, through digital content.
The experience can start even before the trip, 22% of Italian tourists used immersive tools, such as virtual tours, to view accommodation before booking, augmented or virtual reality for the on-site experience. Many more used Google Maps and Street View to visit the destination.
But the experience is also determined by what happens post-trip: the percentage of tourists who, once back home, buy online a product related to the location visited (to maintain the experiential link with the destination) has increased from 4% in 2015 to 12% in 2019 (mostly food products - in 70% of cases - followed by clothing and catalogues or books). Proposing goods (food, clothing, but also merchandising or books on the places visited) can therefore represent an added value of engagement for operators in the sectors.

Also interesting in the post-trip sector is the percentage (30%) of tourists who review their experience online. Social media are also of particular importance: while they are essential for inspiration, they are also widely used to share their experience during and after the trip.

The year 2020 was the year when most of the world's population stayed

indoors and a lot of places and facilities did their best to offer online content of cities, museums, beaches, forests, mountains and experiences in general. Creating content of this kind has allowed people to be in touch with their dream destination even without leaving home. The strategy of creating a relational bond with the customer before and after the trip is a way of guaranteeing new sources of income, in fact, the more the tourist feels attached to the place and the accommodation, the greater the likelihood of his return.

In recent years, the way people do tourism has changed a lot. New forms of tourism have developed to meet the needs of an ever-changing society. The elements characterising the new tourism trends are the low possibility of reaching distant locations and the concern for safety.
Proximity tourism sees tourists reducing their trips in favour of nearby locations that can be reached in compliance with regulations, using their own means, also to avoid contact with strangers.
Undertourism, on the other hand, involves choosing less frequented destinations to avoid particularly crowded places, an indication of how the Covid-19 pandemic has changed people's approach to travel. Closely related to undertourism is sustainable tourism, in which tourists choose places in contact with nature, with many outdoor activities and small local excellences. Sustainable tourism aims to promote knowledge and appreciation of local cultures and traditions, while respecting the environment and the way of life of the host countries, territories and populations.

Since the 1990s, attention to the phenomenon of ecotourism[3] has grown more and more until it exploded into a real ecological boom. For this reason, a number of alternative definitions have begun to circulate, ranging from "green tourism" to tourism with a low environmental impact. This has certainly increased the awareness of and attention to environmental issues. But it is not just a trend, sustainable tourism is becoming a way of life, and with the Covid-19 pandemic, people increasingly need to travel with awareness. According to the latest June 2021 research by Booking.com, the global pandemic would lead more people to travel more sustainably. In fact, 92% of Italian travellers think that sustainable travel is extremely important and 57% are interested in doing so in the future. While 3 out of 4 accommodation establishments say they have adopted at least some sustainable practices, only 1 out of 3 tells potential guests clearly on various channels.

According to the World Tourism Organisation (UNWTO), sustainable

[3]Tourism focused on environmental and social commitment

tourism is a form of travel that meets needs but at the same time benefits the host country in order to enrich opportunities for its future development. On the other hand, responsible tourism, as defined by the Italian Association of Responsible Tourism (AITR), means a type of tourism that is carried out on the basis of social equality criteria and with respect for human rights so that they can be guaranteed for everyone. It therefore assesses the ethical impact of tourism on the local population and its economic and social development.
But today the two concepts are becoming increasingly close. In particular, sustainable and responsible tourism is a form of tourism that recognises the importance of the host community and its right to participate actively in the development of its territory.

Digital innovation, therefore, is at the heart of the new tourism, whatever form this takes, and the tools available to facilities are innumerable and can guarantee considerable revenue growth in a strongly changed world where travel is still complicated.

For Fabio Galetto, Google's Travel Director, *"to capture the new traveller, the tourism sector must find new ways of investing in technology and human capital. Mass digitisation, cloud and artificial intelligence, training oriented to new professional figures such as data scientist, data architect, chief data officer are some of the fundamental chapters to be written"*[4].

All these innovations are moving ever closer to the new horizon, which is Web 3.0. The new generation of Internet technology, which relies heavily on the use of machine learning and artificial intelligence (AI). Its goal is to create more open, connected and intelligent websites and applications that focus on automated data understanding. Through the use of AI and advanced machine learning techniques, Web 3.0 aims to provide more personalised and relevant information at a faster rate. This is possible through the use of intelligent search algorithms and developments in Big Data analytics. Web 3.0 also aims to make the Internet more open and decentralised.

Antonio Preiti, director of Sociometric and professor at the University of Florence, at a conference entitled "Big Data and Tourism 2.0" said: *"The great prospects of Italian tourism are linked to the 4.0 digital transformation that has the distinctive feature of creating solutions for tourists through direct dialogue between machines, through 'machine learning'. This matching*

[4]`https://www.visitcomo.eu/it/vivere/eventi/Big-Data-e-Turismo-2.0/`

must be profiled on the consumer and available in real time, also allowing the direct connection between the individual tourist and the individual player in the hospitality industry. Tourism is the most important 'instant economy' [5] *known to man. Without incorporating technology, and without storing consumer data in Italy, you can't do business intelligence"* [6].

1.3 User Generated Content

The digital age, characterised by the introduction of new technologies and digital channels, has further changed the habits of modern life: from shopping to job hunting, from the way people communicate with others to the way they search for information. In this historic moment of 'digital disruption', many businesses have been forced to abandon the modus operandi adopted for many years or to reinvent it. The need to meet new market requirements in order to maintain an adequate level of competitiveness has triggered a process of renewal of the economic system as a whole.

It has been seen that social media enable consumers to actively gather information and share opinions; this makes them no longer passive receivers but active generators and distributors of information in the form of video, text, audio and so on. Consumers are thus encouraged to share their vision and exert a collective influence on others and on the brands themselves [47]. The main mode of interaction is the use of User Generated Content (UGC) [10].

Studying the literature, it can be seen that various researchers have tried to define UGC, but what emerges is essentially a heterogeneity of definitions that agree on some points and contradict on others; among these, some have achieved greater consensus. Kaplan and Haenlein [25] define UGC as the sum of all the ways people use social media; Daugherty [10], on the other hand, proved that UGC refers to media content created by members of a general audience, and includes any form of online content that has been created, circulated, and consumed by users. Krishnamurthy and Dou [28], after stating that UGC are now ubiquitous in the vast online world and that they are also often identified as Consumer Generated Media (CGM), specified that they include opinions, experiences, advice and comments on products, services,

[5] Using a computer or smartphone to get all the information and make decisions in a potentially better, more timely and more rational way

[6] `https://www.visitcomo.eu/it/vivere/eventi/Big-Data-e-Turismo-2.0/`

brands, and companies; however, all of this normally has to be substantiated by existing personal experiences, which are then reported through Internet posts.

Finally, it is specified that the content can take different forms such as texts, messages, photos, videos, podcasts and so on. However, the definition that has been most successful, and that is consequently adopted in the reference study environment is the one dictated by the Organisation for Economic Co-operation and Development (OECD) which states that a UGC, which is usually used to describe various forms of media content created and made publicly available by users, must comply with three mandatory requirements:

- The content must be published exclusively on a site that is publicly accessible, or in social networks that are accessible to selected groups of people; thus excluding all content that is exchanged via email or instant messaging applications in private.

- Content must also demonstrate a certain level of creativity without being shamelessly copied from something already present on the web; this restriction therefore excludes all replicas of content that has already been made public.

- Finally, content must be created outside of exclusively professional contexts and practices, thus excluding content created for commercial and advertising purposes.

UGC is defined here as any type of content or interaction created and shared by a user and which is accessible to one or more users on the online web.

Blackshaw[8], in order to underline the importance of UGC, defined it as "a disruptive force that emerges spontaneously from the creativity of users". The essential peculiarity of UGC lies in the fact that they can be created by anyone and are consequently shared and catalogued through a popular spontaneous indexing system based on tagging systems: folksonomy.

Folksonomy refers to the social nature of content classification on the web: in this case it is individual users who classify web pages by associating keywords (tags) to the content. This method of classification reflects the

mental schemes of individuals who create and consume content on the Internet, it is an intuitive indexing system because it is close to the natural model used to classify information in the physical world, and finally it creates a system of correlation and cross-referencing that favours the discovery of new content and serendipity, which consists of discovering something unsearched and unexpected while looking for something else. This allows companies to analyse the tags users attach to online conversations, and enables them to interpret the way people define the characteristics of one or more products offered on the market by directly observing the point of view of consumers.

A distinction must be made between UGC published voluntarily, organic content [7], and those incentivised [8] by a company or other 'web authority' [40]. User-generated content is generally organic, i.e. its creation is motivated by an intrinsic intention and is therefore not driven by a brand, company or sponsor. This type of content is definitely valuable as it comes from an emotionally engaged consumer. Several reasons have been identified as to why users are driven to create content.

According to Krishnamurthy and Dou [28] the motivations behind the creation of UGC can be either rational or emotional. The first group includes the need to share knowledge and the public support of one's principles (influencing); the second group - the emotional one - includes building social connections with friends, neighbours or other users and self-expression.
Other researchers went on to specify how people find themselves sharing and creating content because they find it interesting and entertaining, and think it might trigger the same reaction in their network of contacts. In this case, people are then driven to create content by the pleasure they get from entertaining others, the idea of educating their network or justifying their desire to feel important and useful [52].

Finally, it must be understood that although most UGC is organic, sometimes the content is 'driven' by 'external forces'. A distinction of this type of UGC can be [40]:

- Consumer-solicited-content: solicited but not rewarded creation of content.

[7] A person feels emotionally and intrinsically motivated to produce this content.
[8] This content is encouraged by the offer of an incentive.

- Sponsored conversations: they refer to content created by the consumer paid by the company.[9]

- False conversations: increasingly less used today, these are those that occur when a company posts content purporting to be original material posted by the real consumer.

One of the most commonly used methods is the creation of participatory UGC campaigns: the company asks the user to generate content according to certain guidelines. All content, thus produced by the users, is displayed and then shared again, either by the company itself or by other users. Another case is when the site itself or other 'authorities' incentivise creation, with compensation that is less and less economic and more and more reputational [52] [10].

The evolution of the web has changed the power relations between consumers and companies. Today, everything is more user-centred, and the user has in fact gained and obtained power as a consumer over companies [7]. Consumers are no longer mere passive receivers of information; their role has now been elevated from observers to publishers and creators of online content [8]. Everyone can replicate a content proposed by other users with a comment or a re-share, but even more importantly every user can create content; this makes every user a potential danger or opportunity for the subject matter of the content itself [42].

Forrester Research [11] introduced a method of social etnographics based on some research conducted on the social and digital lives of consumers; it then identified seven types of users based on the way they use and interact with social media. The system introduced and profiled people based on the user's habits over the last month, and proposed a scale indicative of the degree of engagement.

The following categories are then identified:

[9] Often in these cases the company actively seeks out the user, who increasingly identifies with an influencer.

[10] This is the case of the 'degree of trustworthiness' attributed to the user by some review sites.

[11] `https://www.forrester.com/blogs/10-06-25-the_data_digest_the_social_technographics_profile_of_facebook_and_myspace_users_us/`

- Creators: they add value to the social web and their social communities by creating content that can be shared with other users; these personalities are the basis of the social web.
- Conversationalists: users who frequently talk through social media.
- Critics: users who react to content; often these users do not create content but interact with content created by others with comments, ratings, reviews and edits (they are considered good contributors by the community).
- Collectors: users who efficiently organise the accumulation of other people's content.
- Sociables: users who simply visit social networks on a regular basis;
- Spectators: users who often consume content in a 'secret' way, keeping their true identity hidden;
- Inactives: users who are online but do not really participate in social life.

The last categories of users are those who limit themselves more to content consumption than content creation [52].

There are several 'sharing activities' implemented by users: activity-stream, gift application, continuous sharing, upload functionality, embed code, experience sharing and so on [52]). It has been seen that each type of user 'participates' in a different way; equally different is the degree of participation adopted by users depending on the platform on which they act or the environment of reference. The highest degree of participation is found in the sharing and creation of projects, where people with complementary skills cooperate on common projects.

The degree and modalities of co-creation give rise to four different types of participation:

- Co-creation: this takes place between experts and is suited to specific contexts, as well as challenges requiring high-level expertise and disruptive ideas. Contributors generally have to meet certain criteria and then participate in a selection process.

- Coalition and collaboration: this takes place between different teams that can share ideas and investments; a common competitive advantage is necessary for collaboration to be successful.

- Crowdsourcing: is the form of co-creation in which the design, implementation and scope of the project is entrusted to an undefined set of people. This process only has a reason to exist thanks to the tools made available by the web.

- Community: This mode is the most relevant if the aim is to develop something for the benefit of many people.

The web in this sense plays a role in the dissemination of a collective intelligence that exists thanks to a distribution of knowledge and skills in the network, between individuals placed on the same level [42].

In fact, even in the tourism sector, if until recently the most used method to exchange travel information between travellers was word of mouth, now online review sites have been added to this traditional method [18]. Among these user-generated content (UGC) sites, the most famous is TripAdvisor. These sites share images and text on a wide variety of topics and a variety of places of interest and tourist attractions, and in this way user experiences and opinions are made public. This type of sharing is very useful for travellers who want to embark on a new trip, in order to find information on places, attractions, tourist facilities and much more in relation to the place they are visiting. It is possible, therefore, to affirm that UGCs are playing an increasingly important role as sources of information in the tourism sector[36] and this is because they provide news that is highly visible, free and easily accessible from online search engines. Indeed, when conducting an online search in tourism, online search engines tend to direct the user to the GCOs and this is a factor that increases the ease of access to the relevant information [56].

UGC influence people in tourism in different ways and at different stages. Firstly, they are important during the decision-making and planning phase of a trip: thanks to the information contained in them, it is possible to choose one's own destination and to understand which of the many destinations is more inspiring and reflects one's own needs [58]. Secondly, these tools play a motivational role regarding the chosen destination, as the experiences shared

by other users create certain expectations in the traveller, as well as providing ideas regarding activities to do there. In addition, UCGs are the most important external sources of information for both domestic and international travellers. This is also due to the fact that online users do not speak the same language as those who posted their experiences, but can easily and immediately find the information they are looking for [18].

Moreover, if a user has been strongly helped by the information he or she has found in a UGC in order to choose his or her travel destination and experiences, he or she is, for reasons of sociability and emotional support, even more motivated to leave a review on the site [36]. Since the UGCs are a very important source of information in the field of tourism, and their main content consists of reviews, the analysis of the reviews themselves appears to be a possible method to use in order to assess the level of tourist attraction of a given location.

1.4 Rating websites

People use sites that collect user-generated reviews and ratings to find out about a product, brand or service they are interested in. Rating sites are third-party websites that collect evaluations and rankings about different product alternatives on the basis of certain criteria, provide consumers with a neutral source of information and managers with a source of feedback, as well as a form of communication. Thus, the fundamental objective of rating sites is to provide a set of comments, judgements and rankings concerning the offer of a given sector [9].

There are rating sites for various types of products and the main rating sites in the tourism sector used in Italy are TripAdvisor, Booking.com, Expedia and Trivago. These collect and make available to a wide audience the experiences told by consumers so that other people can have a wealth of information. It is possible to search for a hotel or tourist destination and read the impressions of those who have had a previous experience, with ratings on various aspects of the service offered including cleanliness, comfort, friendliness of staff, value for money, attractions. Consumers consult these websites because they believe that by doing so they will make better decisions easily, learn from other consumers, and save time and effort [9]. Moreover, thanks to rating sites they can easily compare alternatives and make quick assessments.

Online review sites of any kind usually contain similar dynamics, functions, interfaces; those dedicated to tourism usually contain a part dedicated to the sectors reviewed (such as 'hotels', 'flights', 'activities', 'restaurants'), a part to offers and a part to the community. A rating site usually contains the following elements: numerical evaluation (rating), the reviews, identification characteristics of the reviewer (at least username), date of writing the text, usefulness of the review [27].

These web platforms contain numerous reviews on different products, but also on the same good, and thus each comment is preceded by reviews that have already rated a product [46]. A reviewer can be considered a member of an online community of reviewers (social group) in which commenters who act as opinion leaders for future reviewers are in turn influenced by other opinion leaders. As a result, a consensus is formed about the evaluation of a product, a conformity to the group [46]. Aral [2] also states that ratings are influenced to a greater extent by positive rather than negative reviews. When people read that others liked a tourist destination and recommend it, this generates the former the same positive feeling about the establishment [2]. In addition, there are also other consumer evaluations that are different because they want to stand out from others (need for uniqueness) [21]. There is thus a social influence on rating sites and this is usually adaptive and bidirectional.

The social influence of other consumers' ratings can both strengthen and weaken another consumer's rating [46]. In these sites, trust is an essential element [21] and, since the authors of the ratings are often unknown, the quality and quantity of the reviews and ratings are important. The rating in itself, not justified by an explanation, does not generate credibility and trust in the reader, whereas the content of the review (the quality), and thus the length and objectivity of the commentary are important to overlook the lack of information about the source. The review with high quality is more logical and persuasive, and supports the evaluation with reasons based on objective factors of the product.

It must also provide a sufficient amount of information. The number of reviews and ratings (the quantity) may represent the popularity of the product since it is assumed that the number of ratings is related to the number of consumers who have made the purchase. Uninvolved consumers are influenced more by the quantity of reviews rather than the quality, while involved consumers are influenced by both, especially by quantity when the quality of information is high [39]. However, rating sites seem to lend themselves well to false ratings due to their free and open nature. While OTAs (Online

Travel Agencies) can guarantee that only consumers who have purchased are allowed to write reviews, on other sites that do not allow direct booking, such as TripAdvisor, this is not guaranteed. The limited information about the author and the possibility of creating false identities online allow for cheating and strategic manipulation of content and ratings [13].

Many tour operators may use false identities and share untrue information to promote their own image or to harm that of their competitors. Three types of evaluations can be distinguished: true evaluations written by consumers who have used the product; defamatory false evaluations usually written by competitors who seek to damage a competitor; and positive false evaluations, usually written by the establishment itself or by users paid by it who praise the hotel's characteristics. There are various measures to assess the veracity of a review: TripAdvisor for this reason has added the 'Star badge' system that allows users to give recognition, giving visitors the possibility to check which reviewers are more experienced, therefore reliable.

In addition, many scholars have identified criteria to identify which reviews are fake: Keates [26] talked about the reviewer's solo visit [19], Yoo, Lee, Gretzel and Fesenmaier [57] stated that fake reviews differ from real ones in terms of lexical complexity, use of first person pronouns, inclusion of the brand name and personal sentiments, admitting that it might be difficult to distinguish reviews based on structural properties. Vásquez [55], stated that the inclusion of positive comments next to negative ones causes the reviewer to be seen as more reasonable, able to grasp what is good and what is missing or substandard. Other content analysis research has revealed similar factors and concluded that most user-generated ratings are authentic [37].
There are many rating sites, not only in the tourism sector but in every industry, and sometimes they contain contradictory information, so consumers have to choose which one to consult and which information to use for their choices. Consumers usually read more than one of these sites and on average look at 2.3 [9]. This is because they want to avoid mistakes by comparing ratings, analysing whether they coincide, and including all possible alternatives with the desired characteristics [9]. Research conducted by Dabholkar [9] shows that the factors influencing the choice of a rating site are the credibility of the platform, the wide availability of information on many alternatives and the possibility of customising the information. A rating site is sufficient if it is perceived as credible and if it provides a large number of reviews. If consumers find a rating site that meets their needs and satisfies them, they will continue to use it for future decisions and become loyal to it [9].

1.5 TripAdvisor

One of the best-known review platforms in the world is TripAdvisor; indeed, user-generated content is at the heart of its mission. In 2020, TripAdvisor received around 59 million reviews and opinions from its members worldwide. This includes different forms of user-generated content (UGC), reviews (26 million), management responses (which are submitted by company representatives in response to reviews) and forum posts.

Travellers submitted reviews from all continents: 54.1 per cent of the reviews concerned customer experiences that took place in Europe, 23.5 per cent in North America, 13.7 per cent in Asia and the South Pacific, 4.7 per cent in Central and South America and 3.9 per cent in Africa, Antarctica and the Middle East. Reviews can be made on this platform for every type of place visited during one's stay. In 2020, travellers made over 8 million reviews for hotels, over 12 million reviews for restaurants and over 4 million reviews for experiences, attractions and activities.

1.5.1 The world's largest travel community

Founded in 2000 by Stephen Kaufer and Langley Steinert, TripAdvisor is a travel site that provides reviews and other UGC information to users, with the aim of helping them plan and book their holidays.

CEO and co-founder Kaufer, conceived TripAdvisor with his wife Caroline from his office in Newton, Massachusetts:*"It was 2000, my wife Caroline and I were planning a holiday to Mexico, so I started looking online for information on a hotel that looked particularly appealing. I had found thousands of sites all showing the same great picture and idyllic description.' How can I get an idea of what I will actually find?' I wondered. Then, using my computer skills, I honed in on the search and finally found a first-hand description of a couple who had stayed in the same hotel we liked: their photos showed rusty deckchairs and a very different beach from how it had been described to us. We had dodged a bullet. Caroline then suggested I create a site to help travellers in similar situations: 'Just make sure it's easy to use and gives correct information'"*. The world's largest online travel site was born.

The insight of Kaufer, a Harvard engineering graduate, came from a problem he had personally experienced: getting unbiased travel information on the Internet: *"I wasn't interested in looking at rich brochures or attractive*

hotel websites that allow you to make an instant booking, but don't provide in-depth information. I wanted to know what others had to say about their stay".

The aim was to create a website where users could give their opinions and personal reviews about their travel experiences, a platform that would be a source of useful information for other travellers, because it was objective and unfiltered like the information on accommodation websites.
Since its creation in February 2000, TripAdvisor, in less than twenty years, has become not only a site that collects feedback from travellers on tourist facilities but also a valuable tool to support the B2B world, able to offer suggestions for the management of online presence to the owners of the structures reviewed.

The business model initially designed was to monetise TripAdvisor by inserting banners and content from other sites, but later, Kaufer and the other co-founders came up with the idea of offering online travel agencies (OTAs) the possibility to place text ads on the portal and travellers to give free reviews of their travel experiences (initially they could only be directed at hotels).

The first results were surprising, the number of clicks through rate was 8 percentage points higher than the industry average (taking as a reference the sites of major competitors including Expedia, Yahoo Travel and Expedia, Yahoo Travel and Hotels.com)[49].
In November 2002, the 'popularity index' was introduced, which, based on feedback from travellers, assigns a ranking to each property reviewed, which is then included in an overall ranking. The success of the early years led in 2004 to the acquisition of TripAdvisor by Interactive Corporation (IAC), now Expedia Inc., a holding company that controls Expedia, Hotels.com, and Hotwire.com, all companies operating in the online travel business, the same year, restaurants were also included among the facilities reviewed.

In 2006, the portal expanded overseas, reaching Ireland, the UK, Italy and Spain. In 2009, the new company structure allowed it to expand its offerings to include services beyond accommodation and tour searches, including a tool for searching for airline tickets. May 2010 saw the launch of 'TripAdvisor for Business', a section dedicated to businesses that provides owners with tools and advice to interact optimally with travellers.

In 2011, TripAdvisor became independent and started to attract travel-related mobile applications, including a system that allows tourists to map

each place they visit by entering photo albums, and an application that allows travellers to get complete information about their flight and airport conditions.

In the Italian market, the latest innovation introduced, when, given the important role played by reviews in the traveller's purchasing process, it announced that it would complete the same within its own site, offering the additional service of "TripConnect instant booking". In fact, according to research conducted by PhoCusWright, an American tourism research and consultancy company, for consumers reviews are a tool that gives them greater confidence in their choice of trip, so much so that most of them do not make a booking until they have consulted opinions and reviews issued by other travellers about a given structure.

The introduction of the new service has transformed the tourism portal into a full-fledged Online Travel Agency (OTA), now forced to compete with industry giants such as Expedia and Booking.com. The new "Instant Booking" service is commission-based for hoteliers, and allows visitors to book directly on TripAdvisor the property whose opinions they have read from other users, without being diverted to the latter's official web page to complete the transaction. TripAdvisor works with the booking engine provider of an accommodation to show its rates and availability on the site. In this way, the tourist looking for information on TripAdvisor, by means of a simple click on the "Book on TripAdvisor" button, makes the booking without having to leave the portal. The owner of the accommodation facility that uses this additional service pays commissions to TripAdvisor only on successful bookings.
It's a quick and practical service that is optimised for desktops, tablets and smartphones.

TripAdvisor continually makes significant investments in technology, brand building and relationship management with partners and advertisers, in order to improve the platform globally, attracting more users worldwide. The TripAdvisor business model is based on the management of three main digital marketing activities, which are important sources of income for the brand.

1.5.2 TripAdvisor tools and functionalities

TripAdvisor was founded as a user-generated content (UGC) site operating in the international travel industry, whose activities focus exclusively on on-

line travel management and online advertising. Its main function is to act as an intermediary between users wishing to plan and book a holiday and travel accommodation providers around the world. The founder and his team, convinced of the central role played by the Internet as an integral part of the travel planning process, have firmly believed in the potential of the travel portal, creating around authentic and quality reviews, a large community of travellers, representing today, one of the elements of strength of the brand.

Registration to the portal is free of charge and is open to all accommodation providers and users. In particular, for users, registration is only required to issue a review or make a booking and is not required if its use is limited to reading travellers' opinions or finding travel information. On the other hand, for tourist accommodation facilities subject to review, the creation of the profile can be done not only by the owner himself, but also in the following cases:

- Subsequent to a traveller reporting or reviewing an establishment;
- After a report on a given property has been sent to TripAdvisor by its business partners, such as Expedia or hotels.com;
- Following a recommendation from the editors of the site that they have found a reference to a structure in an article or guide.

Moreover, once the profile has been created (directly by the owner or indirectly by users or TripAdvisor itself), its removal can only take place if the property has been permanently closed, precisely because of its user-generated nature.

1.5.3 Reviews

The presence on the portal represents an opportunity to increase visibility and enhance brand reputation, although the fear of receiving negative reviews about one's own activity can be discouraging for any tourist facility owner.
Despite the risk of receiving unfavourable reviews, the portal represents a complete source of information for travellers and therefore, having a TripAdvisor profile constitutes a great opportunity for facility managers. In this regard, the portal, within the "TripAdvisor Insights" section, provides structures with useful tools and guidelines to strengthen marketing initiatives and manage customer feedback (especially if negative) in order to consolidate

their online presence.

The reviews issued by users both within the TripAdvisor portal and on other UGC platforms on the Internet are determining factors in the development of the online reputation and image of each facility: the more positive opinions recorded, the higher the value perceived by users around that brand.

Specifically, the term "review" is used to indicate an opinion given by a guest regarding a service used at a reception facility; for the user, these represent the information that he or she would like to acquire on the network regarding a specific product/service, not generated by the seller/offerer but by those who have already purchased it previously. From the point of view of the offer, however, they are nothing more than a measure of customer satisfaction, as they make it possible to understand how the experience offered by a given facility was perceived by guests.[50]

The important role played by reviews lies in the reliability of the judgement expressed by a person who, although unknown, has no economic purpose in sharing his or her experience: it is a spontaneous sharing, born with the sole aim of offering further useful decision-making tools to all users surfing the net.

The reviews provided by travellers, by virtue of the fact that they express an opinion on a service that is difficult to assess a priori, because they are developed around subjective relationships established between the seller and the customer, represent the main source of information for the user, capable of filling, at least partially, his initial lack of information.
The possibility "granted" by the Internet to each user to generate and create multimedia contents (e.g. reviews) and the ease with which it is possible to acquire information on a product/service on sale, has revolutionised, in recent years, the traditional seller-buyer relationship on which the concept of information asymmetry introduced by Akerlof [1].

Indeed, in the current scenario in which the potential buyer, thanks to the tools provided by the Internet, is able to acquire more information than the offerer, one is led to think of an information asymmetry on the contrary [3], in which a third party has entered the sales relationship: the reviewer. During the phase preceding the booking, the potential client tends to consult a multitude of information on the web in order to have an overall idea of the services offered by the facilities in the area of the tourist destination and, after analysing and comparing them, he chooses the one that could meet his

needs. The reviewer is defined as a "narrating guest with an unexpected ability to influence", by virtue of his or her ability to condition the decisions of the purchaser, explaining to him or her not only what is offered by a given tourist facility, but also how it is offered.

The important role attributed to reviews is also reflected in the exponential growth they have recorded within the world's largest travel portal, representing for potential buyers an indicator of the high level of reliability and truthfulness of the information. Their wide use is due to the fact that the information reported is not limited to the description of tangible elements of the tourist offer, but often includes a strictly personal description of the emotions experienced, whether positive or negative, which, by touching the traveller's emotional sphere, are able to influence him more in his final choice. The image and reputation of an area or structure is the result of a dense network of opinions and subsequent subjective interpretations that vary from user to user.

The management's objective will be to guarantee a level of service quality that is in line with or higher than that expected by guests, so that their satisfaction can lead to positive reviews and opinions. A satisfied customer is a valuable digital marketing tool that can influence potential buyers more than the brand can directly. For this reason, monitoring and managing what users say about your facility is one of the most important aspects to consider. The definition of the customer's actual experience is essentially based on the difference between the expectations accrued during the booking phase and the reality of the overall service offered by the facility.

An important aspect of UGC sites to be taken into account are the negative reviews. For those who use the service to make a decision before trying out an experience, it is possible to come across false reviews that can negatively, if unfairly, affect the status of hotels, restaurants and tourist destinations in general. This is what emerges from the latest report published by TripAdvisor, the annual Report on Review Transparency[12], in which it examined the reviews issued throughout the year 2020 on its platform. This report highlighted the fact that TripAdvisor rejected and removed over two million reviews that did not meet the platform's community standards. In total, 3.6% of all reviews posted last year were identified as fake, and most were declined before being published on TripAdvisor's site.

[12] `https://www.tripadvisor.com/TransparencyReport2021`

TripAdvisor uses certain methods to combat the phenomenon of fake reviews. One of these is the expulsion of members who do not meet the standards set by the platform. In the last year, it expelled and banned 20,299 users from the service. It took action against 34,605 organisations for 'fraudulent behaviour'.
But the method that seems to work best is a special algorithm created ad hoc, which 'hunts' for fake reviews. In 2020, thanks to this system, it managed to stop 67.1% of fake reviews before they could be read by the platform's millions of users. There is also another problem for Tripadvisor, that of paid reviews, all those positive evaluations that are made by fictitious users bought by the owners of the various structures present on the platform. In this regard, the company has announced that it has removed this type of review in 131 different countries and has identified 65 new fraudulent sites and blocked submissions from 372 external platforms.

The report states that "While the number of reviews has decreased overall in proportion to the decline in travel, fake reviews, which are obviously not based on actual customer experiences, have not followed the same trend." So TripAdvisor is taking important steps to protect all of its users during this pandemic period, with the introduction of new community standards designed to prevent the spread of misinformation about COVID-19[13] and protect establishments that have kept customers safe. As a result of these measures, TripAdvisor removed nearly 50,000 reviews that did not adhere to the platform's COVID-19 posting guidelines.

"Knowing that you can rely on trusted guidance from travellers who have been there before has never been more important. As we continue the work to earn the trust travellers place in our business, we take the enforcement of our community standards incredibly seriously as we use the best in technology and human moderation practices to fight fraud. Today's report demonstrates how effective our team, tactics and technology are at maintaining those standards," said Becky Foley, head of Trust and Safety at TripAdvisor.

Finally, the platform itself provided some useful advice on how to avoid being fooled by the fake reviews circulating on its site. The review should be 'recent', written 'first-hand', touch on relevant topics, but above all respectful and impartial.

[13]Infectious disease caused by the SARS-CoV-2 virus.

Chapter 2

Retrieval of information from the web

Since the evolution of the World Wide Web (www), the scenario of the internet user and data exchange is changing rapidly. As ordinary people join the internet and start using it, many new techniques are promoted to enhance the network. At the same time, new technologies are introduced to improve computers and network facilities. The daily use of the internet means that a huge amount of data is available on the internet. Companies, researchers, academics, all share information on the Internet so that they can be connected and reach people quickly and easily. As a result of exchanging, sharing and storing data on the Internet, a new problem arises: how to manage the data overload and how the user can get or access the best information with the least effort. To solve this problem, the researcher identified a new technique called Web Scraping.

2.1 Big Data

Today, the world is flooded with information and this information, especially in the last twenty years, has been growing in number ever more rapidly. Trying to quantify the data and information that surrounds us is very complicated.

A satisfactory result has been achieved by Martin Hilbert, professor at the "School for Communication and Journalism, Southern California, achieved an interesting result by trying to quantify all the information and data produced in the world from 1986 to 2007 by including data from analogue and

digital sources in his research.

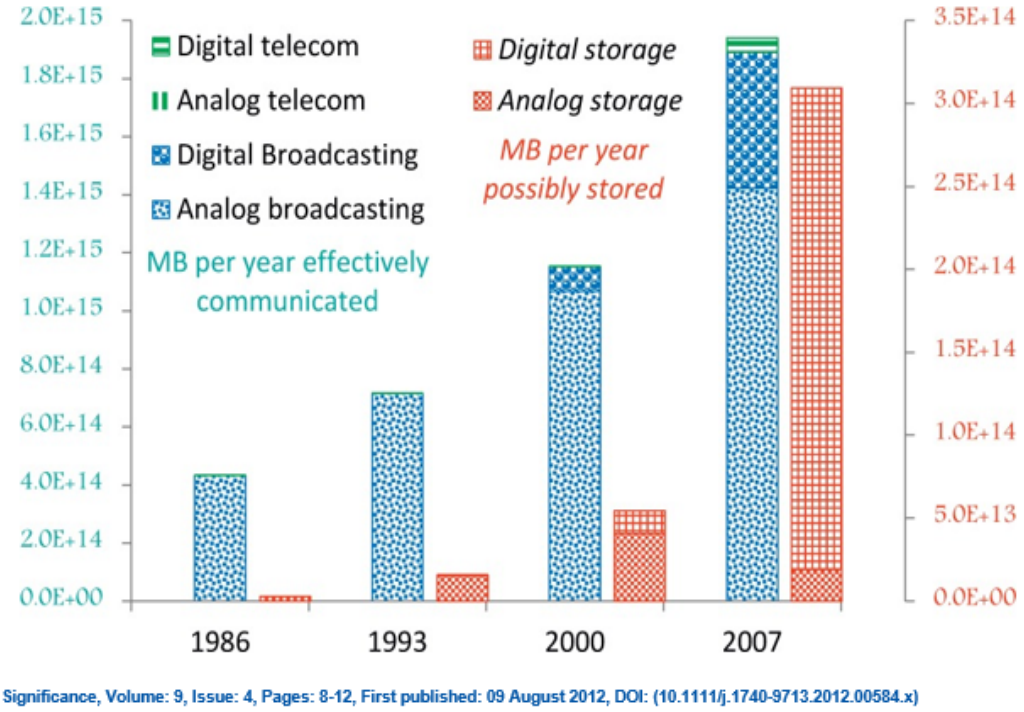

Figure 2.1: How much informations is there in the "information society"?

As the graph shows, around 300 exabytes[14] of data were stored in 2007. While in 2000 only 25% of the data stored worldwide was in digital form, in 2007 the situation was reversed, resulting in only 7% of the data being in analogue form and the remainder in digital form.
Hilbert's analysis concludes that the data is doubling every three years. As proof of this, for 2013 the amount of information stored was estimated at 1200 exabytes[23].
As reported by Data Science Central[15], the amount of data created, copied and consumed globally reached 59 zettabytes[16] in 2020. In the last year alone, due to the pandemic, this has grown by 44% from 41 to 59 zettabytes and is expected to grow by 152% to 149 zettabytes in 2024.

[14]One exabyte is equivalent to one trillion bytes and is the sixth largest unit of measurement for data volumes.
[15]`https://www.datasciencecentral.com/`
[16]One zettabyte corresponds to one trilliard bytes

According to Peter Norvig, Google's artificial intelligence expert, quantitative change will produce qualitative change.

2.1.1 Main features of Big Data

Big Data is the new tool that makes society 'measurable'. They point to a new data science capable of measuring and, in perspective, predicting economic crises, the spread of opinions, the distribution of economic or energy resources, and mobility needs.

Big Data has been very successful in the IT field since 2012, even though large companies such as Google have been using data processing technologies for a long time and have invested a lot of resources in them. This great success of Big Data in the last years in the IT field is due to the availability of open source technologies that use cheap hardware and to the availability of cloud platforms, both factors that contribute decisively to the reduction of costs.

Although Big Data is a recent phenomenon, for which there is not yet an exact and exclusive definition, some researchers have attempted to explain the phenomenon.
In 2010, Apache Hadoop (a framework supporting distributed applications with high access to data under a free licence) defined them as: *"Data sets that cannot be captured, acquired and managed by general computers within a specific domain"*.
In 2011, McKinsey and Company (a multinational management consulting firm) defined the phenomenon as the new frontier of innovation, competition and productivity. According to the company, Big Data are in fact that set of data that cannot be acquired and managed by classic databases. The definition shows that the volume of data is not the only criterion to be taken into account, two other key features are the ever increasing flows of data and the management that can no longer use traditional technological databases.

Finally, in search of a more complete definition of the phenomenon, De Mauro, discussing the existing proposals and investigating their main characteristics, tried to synthesize an expression that formally represents the essence of the phenomenon: *"Big Data are the information assets characterized by such a high volume, velocity and variety of data as to require specific technology and analytical methods technology and analytical methods for their transformation into value'*[12].

By analysing all the present definitions for Big Data, it can be said that the core of the concept includes the following aspects:

- Volume, Velocity and Variety, to describe the characteristics of information;
- Technology and Methods of Analysis, to describe the requirements needed to make proper use of this information. of this information;
- Value, to describe the transformation of information into insights that can create economic value for companies and society.

A useful way of characterising the key attributes of Big Data is usually to refer to the '5 Vs': Volume, Velocity, Variety, Veracity and Value. These are characteristics whose convergence helps both to define and distinguish Big Data from data in general. This characterization is inspired by the definition of the "3 Vs" originally introduced by Laney [29], who considered as principal dimensions of the Big Data the volume of the data, the speed with which the data is collected, used and diffused and the variety of aggregated information.
Subsequently, the additional V of veracity was included by IBM [44] to emphasise the importance of addressing and managing the uncertainty inherent in certain types of data. Finally, to further complete the picture one can include value, both as a basic dimension and as a consequence of the other four characteristics. as a consequence of the other four characteristics.

- *Volume*: the term refers to the huge amount of all kinds of data generated by different sources that cannot be managed by traditional databases, but need to be organised and analysed. The amount of data is of the order of Zettabytes, or billions of Terabytes; this requires dedicated parallel and massive computing power, running on tens, hundreds of thousands of servers.
  Over the years, two key factors have contributed to the exponential increase in the volume of data collected and exploited[16]: reducing the cost of collecting, storing, processing and analysing data and the increasing online activity of consumers, driven by increased access to high-speed internet, as well as more online and connected goods and services, including those provided by IoT devices[17].

[17]Path of technological development whereby, through the Internet, potentially every object of everyday experience acquires its own identity in the digital world

- *Variety*: this is data of a different nature, which may be structured or unstructured, collected in a variety of ways, for example, through smartphones, social networks or commercial transactions. Both because of the presence of different data sources and the great opportunities arising from the possibility of combining data of different formats, variety is one of the most distinctive features of many data-driven business models, such as that of online platforms.
  This is confirmed by the results of research conducted by NVP (New Vantage Partners) which shows that 40% of the companies surveyed - belonging to the Fortune 1000 - refer to variety as the main driver in their investment choices for Big Data management, compared to 14.5% who cite volume and 3.5% who indicate speed.

- *Velocity*: refers to both the speed of data generation and access, and the speed of data processing and analysis. Also this further dimension of Big Data, like the previous ones, has undergone such a growth that the traditional techniques of traditional techniques of data collection and processing no longer work.
  While it is true that in some contexts the value of data is not perishable over time, it is equally true that for many business opportunities related to the ability to rapidly exploit available data, real-time analysis is now a fundamental prerequisite. However, this requires sophisticated skills, technological infrastructure and software solutions.

- *Veracity*: the quality and meaningfulness of the data collected or processed, especially in the light of the current exponential increase in the three dimensions defined above and by virtue of the ever-increasing of machine learning processes.
  Data quality and integrity is a prerequisite for organisations working with 'truthful' data to produce useful and reliable analyses. In order to achieve this goal, pre-processing techniques are often required to eliminate non-significant and/or non-integral data.

- *Value*: is considered the most important aspect of big data and refers to the process of identifying high value hidden within a large, diverse and rapidly growing amount of data. It is therefore essential to assess the veracity and quality of the data in order to effectively generate new value.

> The ability to extract value from Big Data is mainly due to the proliferation of increasingly sophisticated analytical techniques that allow, through the extraction of information from data, to improve decision-making and performance, contributing to the efficiency and quality of processes and qualifying the supply of goods and/or services, in particular, in terms of innovation and customisation.

Big Data is a combination of these characteristics which confirm a very rapid and complex evolutionary dynamic of the phenomenon, but which could create a revenue opportunity for organisations and a competitive advantage in today's digitised market.

In the process of extracting value from Big Data it is possible to identify a sequence of main activities: Collection, Archiving, Analysis and Use. These four steps form the Big Data value chain.
From collection to use, in fact, the data set proceeds through several interdependent phases, each of which can be compared to the links of a chain, which gradually increase its value.
Each of these four pillars has special characteristics and sees the potential involvement of different stakeholders, including individuals, organisations, enterprises, and public institutions. Most of these stakeholders are only involved in selected parts of the value chain, while only a few are more vertically integrated, as in the case of companies that are able to collect data on consumers, then store them, aggregate them and finally process them and use them in their business model.

The first step in the value chain corresponds to data collection, which has become increasingly important due to the fact that most economic and social activities are on the Internet and all media content is available in digital format. For most business models of companies operating online, the collection of data from users' online activities is of particular importance. online activities of users. Technically, the acquisition of user-generated data presupposes the use of a number of different technologies dedicated to tracking them.

The main data collection mechanisms are:

- First-party data collection: occurs whenever a company collects information directly from its customers/users as part of their use of the goods or services it offers. services it offers.
- Third-party data collection: collection of user data that does not have a

direct relationship with the online service provider and is done through third-party tracking. Third parties may accept such tracking as part of commercial agreements to receive website analysis and ad serving services. Tracking by third parties is prevalent on both websites and applications.

- Purchase from intermediaries: companies in the data collection phase may acquire data from intermediaries, called data brokers, if they do not have their own large first-party data set or are looking to enrich existing data. These are companies whose main activity is collecting personal information about consumers from a variety of sources and aggregating, analysing and sharing that information, or information derived from it, for purposes such as marketing products, verifying a person's identity or fraud detection. These intermediaries, by gathering and combining data from a variety of online and offline sources, are able to create detailed user profiles which are then made available to companies for purposes including marketing, advertising and financial services.

After the collection phase comes the storage of data in retrievable forms so that it can be processed. In view of the increasing volume of data that is acquired, large data centres are required for this activity. These are made up of large computer clusters with large storage capacities that allow easy scalability and fast access and transfer times. Data can be stored on internal servers or on external cloud computing services.

The third step in the Big Data value chain is analysis, which plays a central role because data by itself has little value but acquires value when it is organised and processed. This key value-adding step involves organising, integrating and analytically processing raw data in order to deduce information that can be used for economic purposes. In essence, this activity is concerned with transforming information derived from large, often unstructured, data into knowledge and is conducted through the use of sophisticated analytical techniques and algorithms.

The last link in the Big Data value chain is the use of knowledge extracted from the data. There are several forms of use, the main ones being: improving products or services, exploiting new business opportunities by offering innovative products and services, developing more target-oriented business models by better targeting potential customers.

The real Big Data revolution is the emergence of new tools and operations, capable of linking information to provide a broader visual approach to data, suggesting previously unthinkable structures and models of interpretation. One might think that Big Data is a field of interest only for the IT sector. But while Information Technology is undoubtedly the first step for Big Data, it is useful everywhere. No sector that is based on marketing and in which there is data to analyse can really be said to be immune to the Big Data revolution. This revolution touches the lives of every single person without anyone being directly aware of it: all the data coming from a user's browsing enables the giants of commerce (electronic and otherwise) to identify and propose the products that best suit the customer's needs and desires, those that arouse their curiosity and drive them to buy out of momentary or permanent need or on the wave of a simple impulse.

However, all these new technologies are not without their difficulties. After all, the large amount of data from online sources is often unreliable, and their dynamism makes any attempt at systematic study or investigation for assurance and consolidation purposes complex. Moreover, errors and possible gaps may be magnified when several different datasets are used together. In the absence of a clear research objective and an equally rigorous data collection plan, the risk of obtaining insignificant or misleading results is rather high. Moreover, the difficulties and costs of accessing Big Data risk producing a new kind of digital divide between those who can afford the necessary economic effort and obtain significant benefits and those who, instead, cannot make use of the indications that can be obtained from such studies. In fact, when one has at one's disposal a potentially boundless set of data, the probability of finding a significant correlation between any two 'series', even if they are completely random, can be as high as 90%.

In summary, Big Data scholars and practitioners agree that having access to a large amount of data, covering virtually every aspect of human life, brings considerable advantages. But in order to enjoy these advantages, it is necessary to take meticulous care in the processing phase of the data obtained from the movement of any survey that uses information from an enormously populated online world. Having overcome, therefore, an understandable initial diffidence, the Big Data can best support the collection, classification, analysis and synthesis of the data of a given sector, offering precious information which goes beyond the simple raw data.

It is therefore easy to see that data is one of the main drivers of value creation in the digital age and, moreover, the dramatic increase in its importance

is basically due to two relatively recent technological trends: technological progress and the development of sophisticated analytical techniques. As economic and social activities become increasingly digitised, they are constantly being produced. The development of sophisticated analytical techniques, on the other hand, has made it possible to achieve advanced degrees of data processing thanks to which more value can be extracted from the data. The importance of these two factors can be illustrated by the fact that data are at the heart of the highly profitable business models of the world's largest companies, so much so that the expression 'data is the new oil' [15] is not uncommon, although unlike oil, which is a finite and non-reusable resource, data can be - taking into account of course ownership and access rights - infinite and reusable.

2.1.2 Market power of Big Data

Big Data has become one of the most important topics in the ongoing debate on competition and regulation in the digital economy. The collection and use of data by companies for commercial purposes is not a new phenomenon, but the current situation is notable for the scale and scope of the data collected and processed and its extreme importance for many of the business models of the most successful technology companies.
In particular, the role of data in the competitive process between firms in the digital economy and especially between online platforms is a controversial issue that has started to attract the attention of policy makers and scholars in the competition law community. In fact, although for some years Big Data has been considered one of the most promising drivers of economic development, more recently it has been a source of concern for the possible market power that can be accumulated by those who collect large amounts of data and have the capacity to process them. Adding to this concern is the evidence of the disruptive commercial success of four huge companies (Google, Facebook, Amazon, Apple) that are able to collect huge amounts of data from consumers and offer them data-driven services. These companies control the largest market share in major services provided online.

In general, there is no agreement on the role of Big Data as a basis for sustained market power and consequently no agreement on the role of competition law with respect to the Big Data business. There are two prevailing opinions, those in favour of a more proactive application of antitrust in the realm of Big Data, and those who are against such an intervention.
On the one hand, it is argued that Big Data does not constitute a basis for

market power. Tucker [51] and Lerner [30] argue that Big Data does not create a significant barrier to entry and is therefore unlikely to be a source of market power. This assertion is based on the non-exclusive and non-rivalrous nature of data (in the sense that the collection and use by one provider does not prejudice the collection and use by another provider.) and the supposed ease of data collection, without taking into account many potential barriers to entry.
On the other hand, it is argued that superior data may not only temporarily lead to a dominant market position, but that it allows the dominant firm to improve the quality of its database faster than potential competitors and may therefore lead to a permanent advantage. The European Data Protection Supervisor stated that *'the collection and control of huge amounts of personal data is a source of market power for the major players in the global market for Internet services'* [48].

Those who benefit from a greater number of platforms to collect data, who have a substantial database with which to compare new data, or who possess unique tools for its synthesis and analysis, are able to establish significant data power on which to base a competitive advantage. Due to the non-rivalrous nature of information, it has been argued that a company cannot gain power over user data from which market power could potentially arise. But despite this characteristic of data, it is also true that not everyone is free to collect, analyse and use Big Data in the same way and under the same conditions. This results in barriers to entry, and smaller companies or start-ups are more restricted.
Barriers to entry can be of various kinds, and to these are added legal barriers such as data protection and privacy laws that can affect both the activity of data collection and the activity of data storage and use. Another set of legal barriers arise from data ownership issues, which affect the ease of access to data despite its non-rivalrous nature. non-rivalrous nature. The financial burden imposed on small or start-up companies as a result of complying with these rules may discourage them from entering the market, thus reducing competitive pressure and limiting the potential for new innovative products and services.
So it can be said that entry barriers have to be analysed on a case by case basis, because they can be different and specific for each market.

For a company to gain a real data-driven competitive advantage, it must be able to effectively monetise what is processed and analysed through the data it collects. Therefore, the quantity, variety and quality of data a company possesses is not an indicator of its competitive advantage if it fails to

utilise it. The turnover generated by a company through the monetisation of data can be an indication of its competitive strength. Factors specific to data-driven markets, such as high economies of scale[18] and scope[19], network effects, feedback loops, under certain conditions, may give firms a durable competitive advantage, which may help them to persistently deflect current and future competition.

2.1.3 Big Data in tourism

The online digital landscape has enabled us to benefit from waves of innovation that have radically transformed most sectors and the way people seek information, buy products or services and communicate with each other. Whether on a computer, a tablet or any connected mobile device, it is as if everything is literally at your fingertips. In this scenario, it is not possible to perform any online activity without it leaving a data trail like a sort of digital footprint.
The revolution that the use of Big Data represents for every sector is perhaps even more significant for the tourism industry. In fact, among the companies that first took advantage of this resource are mainly airport companies and airlines.

British Airways, for example, in order to counter the competition, has decided to invest in in-depth knowledge of its customers by collecting online and offline information from loyalty programmes. This enables them to understand the most frequent needs and problems of travellers and to develop more effective proposals and solutions. Other airlines, such as Swiss Air, Air FranceKLM and Lufthansa, are using Big Data to improve their revenue management strategies[20], and several hotel chains have also started to implement interventions based on the use of Big Data.
Hilton, for example, introduced the use of a Balanced Scorecard[21] to understand what factors drive organisational performance. Thanks to this activity, it was able to identify correlations between customer satisfaction and cus-

[18] Phenomenon of cost reduction and increased efficiency linked to higher production volume.

[19] A business case in which different goods are produced at the same time with the same production factors.

[20] Set of coordinated activities aimed at optimising occupancy and maximising revenue

[21] A support tool for the strategic management of the company that makes it possible to translate the company's mission and strategy into a coherent set of performance measures, facilitating their measurability

tomer behaviour. Some hotels, on the other hand, use Stem's platforms that are able to facilitate, on the basis of Big Data analysis algorithms that continuously analyse the climate, the building and the use of electricity, an efficient management of energy itself, reducing costs by at least 10-15%. Even the large OTAs (Online Travel Agencies) are not neglecting this aspect, Expedia, for example, is making significant investments in this area, considered to be the key to the future of travel.

The tourism sector recognises a central role for the customer, and uses an approach that primarily values the customer's needs, wishes, preferences and requirements. This is to improve both consumer satisfaction and the quality and memorability of the tourist experience. Big Data plays a fundamental role in this new model, as it allows to achieve and maintain a competitive advantage [34].
Through the analysis of the infinite amount of data, both structured and unstructured, that is developed through the various internet channels, in different formats and at a speed that is only sustainable by new generation technologies, it is possible to guarantee companies in the tourism sector a series of precise and detailed information on the behaviour of their existing and potential customers.

Online platforms where tourists leave their reviews have a very high potential to be useful for Destination Management Organizations (DMOs)[22] [54] and other tourism stakeholders to gain information about how tourists perceive destinations, how they spend their time when they are there and, subsequently, on what basis they create and communicate their experiences and opinions and the related text and image content [24]. It is for this reason that such platforms are gaining importance as an element of DMOs' marketing strategy, useful for improving their branding policy and strategic positioning among the various tourism organisations at territorial level [35]: they offer a valid tool for reaching a global audience and obtaining information from it even with limited monetary resources. In addition, UGCs are important for developing new management policies in tourist places and attractions, based on the real interests of users [58].

The strategies formulated on the basis of this information can be a starting point for designing or providing more appropriate tourism products and

[22]It is an organization whose primary function is to attract visitors for the purpose of optimise the local economy through purchase of room nights, food and beverage, retail items, transportation or visitor services.

services, which leads to improving the tourist attractiveness of a location [33]. In addition, the quantity and quality of reviews contained in UGCs can be used as a measure to divide tourists into segments based on their interests [22] and also to identify the level of touristisation of destinations, based on the geographical concentration of reviews.

Although several studies have shown the very high usefulness provided by the information contained in UGC reviews, and although many individuals make extensive use of it in their daily lives in order to make tourism decisions, the same information is not used in large quantities and proactively by most DMOs [20]. This often occurs because the analysis of the quantity and amount of reviews, which are presented in the form of big data, is not immediately understandable, but, on the contrary, requires a certain type of analytical interpretation [5]. Through their interpretation, in fact, it would be possible to transform the information of big data into precise knowledge regarding the true interests of tourists: this would represent a great information resource for DMOs and a solid base on which to build their strategies [53].

Big Data for tourism does not only offer important insights into collective behaviour, but also into people's movements and the relationship between places, things and people. According to what was recently confirmed by the TDLAB (Digital Tourism Lab), people's daily behaviour is now always characterised by some form of digital intermediation which, in fact, feeds enormous flows of data. If analysed properly, this complex and diverse data makes it possible to substantially implement the decision-making processes of tourism businesses, but also to improve the offer by responding adequately to the complexity of demand.

The advantages of analysing Big Data in the tourism sector are primarily of two types.
First of all, Big Data brings strategic advantages because it allows to know the reputation of a given structure, territory, service or itinerary. This seemingly simple information brings with it an almost incalculable value: the possibility of optimising one's own policy with a view to improving reputation.
The second, fundamental advantage of Big Data for tourism is of an operational nature, because all the information collected and analysed can lead to the maximisation of tourist satisfaction, through the personalisation of their travel experience and offer.

According to Travelport research[23], more than 70% of hotel management is willing to pay for access to additional external reports, and more than 55% of them already do so in order to get an aggregated view of the industry. Clearly, in times of such rapid change, data is needed, but the ability to interpret it is even more important. Exogenous factors (epidemiological, economic, regulatory) must be combined with the effects on habits, behaviour, expectations and emotions, in a new travel scenario that is being shaped day by day.

The types of big data used in existing tourism research are mainly the following: online textual data, online photo data, GPS data. In the graph below all types of Big Data can be seen.

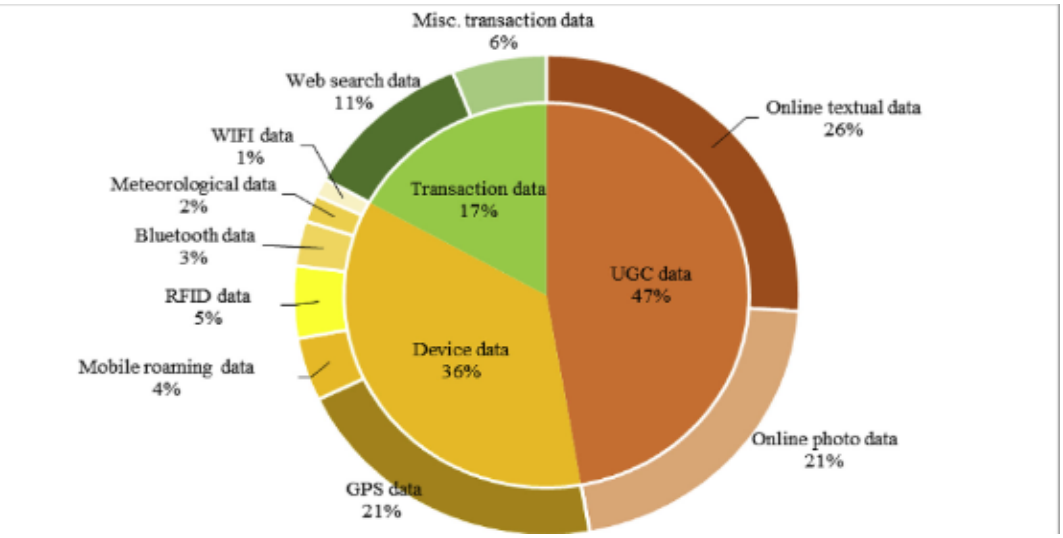

Figure 2.2: Distributions of data sources and data types.

It can be observed that the Big Data used in tourism research is mostly derived from the following data sources: users (accounting for 47%), devices (36%) and operations (17%) [31].
The successful application of data from UGC platforms to tourism research is attributed to the low cost and easy access to such online data. In contrast, the lower level of use of transaction data is mainly due to the fact that most of it is private information, which is only held by tourism organisations or governmental sectors.

[23]https://marketing.cloud.travelport.com/recovery-guide

Therefore, Big Data useful in the tourism sector comes from different sources. Web 2.0 and social media have provided great platforms for tourists to share their tourism experiences, sharing textual data online (product and service reviews, blogs) and photographic data online. In addition, the expansion of Internet of Things (IoT) has made it possible to employ various devices and technologies to track tourists' movements and movements (GPS loggers, Bluetooth sensors, e-readers and WIFI scanners).

The Internet of Things (IoT) is rapidly gaining ground in the ICT world. The term IoT was first coined by Kevin Ashton (MIT) in 1999, defining it as a network that connects anything at any time and in any place to identify, locate, manage and monitor smart objects. The idea behind the IoT is to generate automatic real-time interactions between real-world objects that connect to the Internet, thereby reducing the gap between the real world and the digital world.

In addition to the above sources, there are those from a variety of transactions (transactions, activities or events in the tourism market) such as web search, online booking and purchase etc., these data are typical cases of transaction data. For each type of source and corresponding Big Data a specific systematic analysis is needed in terms of research focus, data characteristics, analytical techniques, as well as main challenges and further directions [31].

It is increasingly evident that it is very difficult to cultivate and sustain a competitive advantage for a long time, as the current economic environment is increasingly dynamic, competition is fierce and consumers are demanding and savvy. We are living in an era of 'temporary advantage' and 'hypercompetition' [11], where organisations need constant innovation to gain a temporary advantage and stay ahead of the competition for a continuous series of periods of time.
Anticipating trends, directing flows and offering personalised services: these are the three characteristics of the contemporary economy to which the tourism sector must also adapt in order to keep up with society. The tools to do this already exist, however, and they are Big Data, digital information, which are great opportunities if the players in the tourism sector learn to manage them, collect them and select them on the basis of quality. However, data alone are not enough, and a network is needed to interpret, study and process them, especially in a sector as complex as tourism.

This is what was said at the conference organised by the Municipality of Como in collaboration with Fondazione Volta and Lions Club Interna-

tional, by the 16 speakers including experts and professionals on the theme of big data and tourism 2.0. The speakers discussed the topic of big data and tourism 2.0, how to systemise the mass of data and how local, national and international players in the sector should respect the large international platforms but have their own keys to interpretation. They are the ones who know the sector best. As far as Italy is concerned, it is essential to use big data, but it is necessary to ensure that it is processed and interpreted by experts in the sector, who are the ones who can best exploit the greatest deposit of biodiversity of historical and cultural heritage.

2.2 Web Scraping

Convenient copy-paste is not an option when the data to be retrieved is enormous: if, for example, information were to be saved manually on a spreadsheet, the process would consist of copying and pasting every single piece of data from its web page to the document to be created, requiring time that would be better saved by using a program to do all the work. Fortunately, although the data on web pages is not perfectly structured, it is possible to use programs, called scrapers, which, depending on the format and structure of the page, can adapt and scrap the data. Depending on what the scraper is used for, the amount of information needed from the data changes. Web scraping is a resource-intensive, time-consuming and complex technique, due to the fact that the software has to analyse different types of data. Obviously, the complexity increases as the number of sites to be analysed increases.

2.2.1 What is and evolution

Since the internet has become an ocean of easily searchable data sources, people have started to find it easy to extract any publicly available data they want. As mentioned above, however, information from the web is heterogeneous, coming from many different sources. Therefore, these two factors: heterogeneity and the extent of data meant that the traditional copy and paste is no longer possible, because it is inefficient and tedious. And that's when the method or word Web Scraping is born.

WebHarvy writes *"Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to*

a local filein your computer or to a database in table (spreadsheet) format. Data displayed by most websites can only be viewed using a web browser. Examples are data listings at yellowpages' directories, real estate sites, social networks, industrial inventory, online shopping sites, contact databases etc. Most websites do not offer the functionality to save a copy of the data which they display to your computer. The only option then is to manually copy and paste the data displayed by the website in your browser to a local file in your computer- a very tedious job which can take many hours or sometimes days to complete." [24]

Web Scraping is a technique used to generate structured data from the unstructured data available on the web. The structured data generated by scraping is then stored in a central database and analysed in spreadsheets. Various techniques can be used to do scraping, from copy and paste to the use of special software. Compared to all possible techniques, web scraping software is the simplest scraping technique. Now, there are many software available on the market that allow you to do this.
Scrapers are often used in conjunction with crawlers, programmes that visit all the pages that can be reached from an address provided as input, following the links. Web scraping is actually powered by bots and web crawlers that work in the same way as those used in search engines. Web scraping focuses on extracting any specific data from the website, and allows users to highlight the necessary information from a web page and structure that data into a usable excel file, or database that provides an opportunity for non-programmers to join in and easily extract data from the web.

Web crawling is the process of analysing websites in order to index all their contents. A crawler, also known as a spider, is specialised software to take all the content of a web page and follow its various links to analyse related websites or sub pages.
Most crawlers provide settings to control and limit execution, the most important of which are important are:

- Depth control: a parameter can be set to tell the crawler how many links to follow from the starting page, generally following 5 links deep is more than enough to analyse an entire website.

- Pages to be followed simultaneously: some web crawlers may make requests in parallel to web pages and allow a limit to be set on the maximum number.

[24] `https://www.webharvy.com/articles/what-is-web-scraping.html`

- Pause between requests: In order not to clog up the website or risk being blocked, you can set a pause to be respected before following a new a new link.

- Saving logs: As the execution of a crawler takes time and can be interrupted, it is recommended to set up a file that keeps track of all the operations performed and errors found, as well as the list of pages already obtained before the interruption.

The most common use case of crawlers is use by search engines, where pages found by the crawler are indexed to provide results relevant to the search query of the user using the service.

Crawlers generally run systematically on the same pages to update information and are initiated by passing the result of the previous crawling as a listener. Since crawlers follow each link found on a web page and can return the content to a scraper, it is possible that information that should remain private or at least not be indexed is saved and made available by search engines. This problem encountered when using these technologies will be analysed in the next chapter.

In order to collect large amounts of data, and not just a specific piece of data, the web scraper is often used as an extension of a web crawler: modern web scrapers can analyse entire websites by following links to retrieve various contents, and then analyse them individually by extracting the necessary data.
Web scraping, also called in different ways web data extraction, is a technique that consists in extracting data from websites. Although the extraction of data from a website can be done manually by the user, web scraping is mainly used when the extraction process is automated through the use of software that emulates a user's visit to a website and when the data obtained are entered into a database or style sheets, or more generally saved locally, and then analysed as a whole. The content of the web page is then analysed to obtain information.
Examples of the use of web scraping are comparing prices of the same product or service found on different websites, analysing data from different users of a website in order to create an organised list or, more generally, monitoring data from websites.

Using scraping software has the following advantages:

- Data extraction is automated and requires little human intervention. In fact, this is indispensable when there are huge amounts of data to be extracted from millions of web pages on a regular basis.
- Data are collected at high speed, due to the fact that the system is automated.
- The information collected is very accurate, as the automatism ensures that human error is eliminated.
- It can also be done without having to write a code, as there are numerous web scraping tools and services available online, some of which are even free.
- Web scraping tools convert unstructured and semi-structured data into structured data, and the information on web pages is reorganised into presentable formats, resulting in ordered, structured and clean data.

Due to all these advantages of using scraping software, they are widely used in all sectors.

2.2.2 Legal issued

As mentioned above, one of the questions concerning the web scraping technique is its legality. And this question is never properly answered. In fact, there are different opinions on the legal and illegal aspects of this technique and the software used to carry it out.

Web scraping is not illegal as long as the 'captured' data are freely accessible on the sites and are used for statistical or content monitoring purposes. That is why it is important to know what kind of data is used, to understand if it is legal, illegal. Scraping becomes illegal when the data extracted is used for other purposes, such as the publication of content in breach of copyright, use for profit and in breach of competition rules, or when personal data is collected for commercial purposes (e.g. for e-mail marketing with addresses extracted from sites) without the knowledge or consent of the persons concerned.

In most cases, the degree of ease with which web data can be accessed determines more or less where the data lie in the sphere of legality.
Scraping data from public websites is perfectly legal. This refers to data

and information on websites that can be obtained without the need to log in or authenticate one's identity. Examples of such websites are e-commerce platforms such as Amazon and BestBuy. Although these data sources may attempt to protect public information by placing various obstacles in the way of scrapers and crawlers, extracting data points from them is perfectly fine.
Private or personal data is identified as all data that can reveal a person's identity, such as name, address, date of birth, medical and financial details and contact information. As a general rule, it is illegal to scrape any personal information without the person's consent or without any legal reason. The EU and California currently have the strictest laws in this regard.
It is illegal to scrape any accessible data such as images, songs, articles, etc., which are the intellectual property of any company or individual, copyrighted data. Since their owners have full control over their use and reproduction, scrapers require explicit consent to extract them. As an alternative solution, you can use snippets of data or cite and credit sources to use the data.

In general, before scraping any website for its data, you should be aware of what its policies are regarding access to your data. If they explicitly contain restrictions on scraping, it is worth assuming that scraping is a violation of their website terms of service (ToS). Also, even if there are no such policies, beware that their content may still be protected by copyright.
Services such as LinkedIn require users to have an account before their data is visible. By signing up for these services, the user almost always agrees to their terms prohibiting data scraping. Since bots and crawler scrapers use account credentials to access data, the service provider can easily identify you and ban you completely from their platform. Therefore, the best thing to do in order to avoid web scraping is to find publicly available data.

The guidelines set out by the Investment Data Standards Organization Best Practices are used by all those who use web scraping to collect information. Among them is the financial world, which is much more careful about assessing the legal aspects of how the data was obtained.
The guidelines require the following points to be followed[25]:

- Follow the instructions in the robots.txt file[26].
- Finding the Terms of Use on the site and understanding what they are. If one has to click on the acceptance of these in order to access the

[25] `https://www.investmentdata.org/publications`
[26] File that contains directives that tell search engines which parts of our site they can scan and which they should not.

site (known as a clickwrap), the user of the site is obliged to read the Terms of Use and can therefore be accused in a possible court case of having read them and not having respected them. In the very frequent case where the Terms of Use are on a dedicated page or section of the site, the user can access the site without viewing them and so can the scraper. Consequently, in a possible lawsuit, these can hardly be used to the detriment of the web scraper.

- Do not overload the site's servers with requests.
- Only extract information that is factual and not covered by copyright (e.g. prices, not photos).
- Always access publicly available information.
- If there is an application programming interface (API) available, use that instead of scraping the site.
- Do not extract information to gain a competitive advantage over the target site.

Since there are currently no clear laws determining whether or not web scraping is legal, and international legislation is still unclear on this issue, cases are dealt with on a case-by-case basis. In most cases in Europe and the US, reference is made to the General Data Protection Regulation (GDPR) and the US Privacy Act. Therefore, in order to understand how to deal with this issue, it is useful to analyse cases that have occurred in reality and how they were concluded.

The GDPR came into force in May 2018 and protects the personal data of individuals within the European Economic Area (EEA). Examples of personal data include people's names, emails, phone numbers, dates of birth, IP address, credit card and bank details, medical records and media content such as photos, audio and video.
The GDPR classifies the protection of personal data as a 'fundamental right'. As such, it prohibits the processing of personal data unless it is carried out on one of six legal bases: consent, contract, public task, vital interest, legitimate interest or legal requirement. When processing is based on consent, the data subject has the right to withdraw it at any time.
In addition, data controllers must clearly disclose any data collection, state the legal basis and purpose and indicate how long the data is retained and whether it is subject to sharing with third parties or outside the EEA.

As for the US, although it does not have a federal regulation tying privacy and data protection like the EU, there are several sector-specific legal acts, such as the GLBA for finance, HIPAA for healthcare and COPPA for children's data.
In 2020, however, California passed a state law, the Californian Consumer Privacy Act (CCPA), which requires companies that collect personal data to explicitly disclose how they intend to use that data and also allows consumers to remove their information or opt out of data collection. . The same rules also apply to data scraping companies.

Comparing the two laws shows that the GDPR and the CCPA both allow consumers to access and remove their personal information and opt out at any time. However, users can change their data under the GDPR, but not the CCPA. Similarly, the CCPA only requires privacy notices on websites, whereas the GDPR requires explicit user consent.

The most frequent legal problems in the context of web scraping are the following:

- Copyright violation: any violation of copyrighted data is punishable, regardless of how you access and collect the data.

- Violation of the Computer Fraud and Abuse Act

- Transgression of movable property: A breach of assets (or site security) occurs when a website or its servers are hacked or damaged in any way. In the context of web scraping, a crawler that repeatedly sends requests can affect the performance of the target website by crashing or slowing down its server. From a legal point of view, site owners might consider frequent requests as an intentional attack on their system. Consequently, it is important and morally responsible for DaaS providers to create scrapers that do not damage the target website.

As mentioned above, there are some historical cases that have established legal precedence in web scraping lawsuits.

eBay v. Bidder's Edge (1999)
Bidder's Edge, a Web site that collected auction listings, was sending 100,000 daily requests to eBay's servers for access to its ongoing auctions, causing damage to eBay's systems. In late 1999, eBay filed an injunction against Bidder's Edge, alleging violation of the Trespass to Chattels Act.

Although both parties subsequently settled the case out of court for an undisclosed amount, it set a legal precedent for future cases.

HiQ Labs vs LinkedIn (2019)
This landmark case began when HiQ Labs, a data analytics company, sued LinkedIn for prohibiting it from deleting public profiles on LinkedIn. HiQ Labs used the data to consult employers about candidates.
In 2019, the Ninth Circuit Court of Appeals ruled that the CFAA did not apply because the data was publicly available and not protected by copyright. As a result, LinkedIn was unable to prevent HiQ Labs from accessing its public profiles. However, it restricted access to user profiles only after access.

In Italy the case of Trenitalia versus Trenit, an app that showed fares and train delays, is famous. After an initial suspension of the app, the court gave reason to the latter, based on the following criteria:

- No personal information is collected without the consent of the individuals involved (as might instead be the case with a massive scraping of data from Facebook)

- The database is partially collected: in reality, the judge probably understood as partial the display by the user of the database, who cannot see all prices, timetables and routes of all trains but must select a route to see the data of interest.

- This does not harm the owner of the data (in fact, it does not compete with Trenitalia) but, on the contrary, it offers a service to the user of the app.

Although legislation varies from country to country, this ruling follows, at least in the most general concepts, the guidelines set out by the Investment Data Standards Organization Best Practices.

Other cases in Italy occurred when the Data Protection Authority (GPDP) intervened and prohibited a company from using the personal data of twelve million users, which had been identified and collected by scraping from various web pages. The company in question had subsequently created its own website where it had published the information collected in the form of an online telephone directory that could also be consulted by other companies

for telemarketing purposes [27]. On another occasion, the Garante declared the practice illegal, prohibiting a company from sending commercial e-mails to professionals whose e-mail and PEC addresses had been taken from public domain lists, but without asking for and obtaining the necessary authorisation from the legitimate owners [28].

As the issue of web scraping is not black and white, it is necessary to carefully analyse each use case to avoid unintended consequences. It is necessary to consider existing legislation, the types of data collected, the terms and policies of the data source and also the ethical use after extraction.

[27] https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/6053915

[28] https://www.altalex.com/documents/leggi/2018/03/01/spam-vietato-sulle-pec-dei-liberi-professionisti

Chapter 3

Analysis of Italian tourist destinations based on TripAdvisor reviews

As said in the previews chapters, the tourism sector has undergone profound changes in a short period of time due to a deep innovation in the way people organise their trips and choose their destinations, using digital tools at all stages of the tourist experience. The ease with which users find and disseminate information of all kinds on the web has favoured the creation and development of User Generated Content (UGC) sites, which, in the tourism sector, allow customers to quickly disseminate comments and evaluations on their experiences, places visited and accommodation facilities, visible to billions of people around the world.
The importance attributed by consumers to reviews issued by other users on the web, as a more neutral and reliable source of information, has made these digital platforms supplementary communication tools to the traditional ones.

For this reason, an analysis was carried out on the most famous and globally used UGC portal, TripAdvisor. The aim of this platform is to offer users a space to exchange opinions and reviews on hotels, restaurants and tourist attractions. Its strength is the reviews, especially since these are written directly by those who have experienced them first-hand.

3.1 Research objectives

Given the growing popularity of UGC and the possibility of having a huge amount of data, Big Data, created by the publication of reviews on tourist attractions, accommodation facilities, restaurants by the users of these services, it was decided to focus initially on the quantitative aspect of the reviews and later on the qualitative aspect of travel 2.0. The main objective of this thesis project is to investigate this macro-topic in relation to the tourist attractions of the locations found in Italy and present on the tourist portal of reference, TripAdvisor.

The literature analysed in the previous chapters attempts to understand and study this complex and ever-changing phenomenon. Xiang, Schwartz, Gerdes and Uysal [17] argue that Big Data analytics could develop new insights to reshape the understanding of the hospitality industry and support its decision-making.
Through the UCG analysis it is possible to know how much they liked their service and/or attraction, to know the characteristics and the places that were liked more or less. It is essential to monitor all these opinions and experiences that tourists leave behind like a fingerprint, and to do this you need the right tools. There is no doubt about the importance of analysing the reputation that is created on social networks or specific tourism portals, not least because a user's opinion is conditioned by that of other people. This conditioning is even greater if these people have some social influence or act in a close in a close environment.
The exploitation of Big Data provides a better understanding of tourism demand, tourist behaviour, tourist satisfaction and many other issues associated with tourism. It is important to identify who one's customers are, in order to be able to offer an offer that is as personalised and integrated as possible.

So, as mentioned in the previous chapters, Big Data and UGC sites play an increasingly important role in the world of tourism, and this innovation has become even more pronounced in the last two years due to the global pandemic. In this paper, an analysis has been carried out, starting with these data and ending with their processing and results.
It was therefore decided to take a more in-depth look at the reviews published on the famous portal and to analyse the quantity of reviews using an index that relates them to the quantity of tourist attractions present in a given location in Italy. For you to analyse this index with the quality offered.

3.2 Database

The analysis is based on data from the world's most famous UGC tourism portal, TripAdvisor, where you can find information about any tourist attraction, hotel, restaurant and location anywhere in the world. The first thing that was done in order to start the analysis was to create the database. The data used was extracted from TripAdvisor using the web scraping technique. Web scraping is a technique for extracting information and data from the pages of websites by means of automatic procedures [60]. This technique is also known as data scraping. The main purpose of scraping is to extract information from the corpus of a text, available on the Internet. The data are extracted, processed and stored in a database. Scanned documents can be recorded in the database in an unabridged form or in a reprocessed form.

The software used for scraper analysis is WebHarvy. It allows to easily extract data from websites and does not require programming or scripting knowledge and works with all websites. It is possible to use WebHarvy to extract data from product lists/e-commerce sites, yellow pages, real estate listings, social networks, forums and so on. It is easy to use and allows users to select the data they need in just a few clicks. Collect data from different directories by following each link. WebHarvy works anonymously, protecting the privacy of its users.

The data refers to all locations in Italy where there is at least one user review on TripAdvisor. The data has been extrapolated up to April 2021, and refers to all reviews found on the site prior to that date. The data taken into account is the sum of all reviews found, without any distinction according to the year of publication, language of the review or origin of the user. When analysing this data, it must be taken into account that the tourism sector has suffered a sudden decline in all its aspects in the last two years due to the global Covid-19 pandemic. The Covid-19 pandemic has had a negative impact on domestic and even more so on international tourism flows. Therefore, if this pandemic had not occurred and the upward trend in the tourism sector had been maintained, the data on which this analysis would have been based would have been much higher. Bank Italia, in its Covid note on the subject, points out that the sector is among those that have suffered more than others from the consequences of the restrictions, which have limited flows and presences and forced several operators to support themselves with loans or to close down completely.

The restrictions on movement due to the closures that have limited recre-

ational, accommodation, cultural and entertainment activities have meant that overall attendance has fallen by more than 52%, of which 33.8% by domestic tourists and more than 70% by international tourists. This has also meant that the use of these platforms, including TripAdvisor, has decreased. This is despite the fact that the website and the industry have been working hard to spread messages to protect users and ensure they can travel safely. What has decreased are the reviews left on UGC sites by users, because as they move and travel less they have had fewer opportunities to try out facilities, restaurants and activities in locations other than their local area. In fact, in the last two years we can see that the reviews of people who live close to the location reviewed are greater than those of tourists from distant regions or even from abroad, both in Italian and in other languages, and for any type and form of tourism and therefore tourist attraction.

In any case for the analysis of this article, is considered the total number of reviews found in the platform per Italian location. From TripAdvisor the data taken in consideration to create the database are: the macros (the types of tourist attractions), the reviews, the stars; all these data divided for each Italian location asua time divided into Italian regions.

In order to carry out the data extraction, with the WebHarvly software, it was initially necessary to select in the portal, in a page taken as reference, the sections of interest. And then set this to be done in all pages related to Italian locations found on TripAdvisor. The initial step of highlighting the necessary data correctly is very important to ensure that the data is extracted correctly by the scraping software.

In each Italian location the total number of macro, that is tourist attractions, grouped according to the following categories, has been calculated[29]:

- Bar and clubs (bar, clubs, pubs..)
- Cultural sites (museums, statues, observatories/ planetariums..)
- Entertainment and events (food & drink festivals, sporting events, cultural events, seasonal fireworks, exhibitions, music festivals..)
- Landmarks and sites of interest (architectural buildings, monuments/statues, bridges, scenic drives, observation decks/towers, lighthouses, ships,

[29] `https://developer-tripadvisor.com/content-api/business-content/categories-subcategories-and-types/`

wharfs/ piers/ boardwalks, fountains, arenas/stadiums/fields, neighborhoods, universities & schools, religious sites, historic sites, ancient ruins, educational sites, castles, cemeteries, scenic/ historic walking areas, government buildings, reservations, mines, civic centers, race car tracks..)

- Natural sites (national parks, playgrounds, other nature and parks..)
- Relax and wellness (Roman baths, Arab baths, Hammams & Turkish baths, yoga & pilates, thermal spas & hot springs, health clubs, spas, onsen resorts..)
- Shopping (malls, shops, specialty shops, art galleries, shopping tours, fashion shows & tours, department stores, factory outlets, antique shops, flea/ street markets..)
- Tour and activities (day trips, all types of tours, in any destination, sea, mountain, city..)
- Transports and services (ferries/boats, airport services, bus services, rail services, mass transportation systems, trams, airport & ground transfers..)

These categories are those used by the TripAdvisor platform. Each tourist attraction can be found in more than one category, but in order to obtain the database without repetition, it was decided that each tourist attraction would be found under the category that the site reports first. Therefore each tourist attraction is considered within the macro category in which it is considered first. In recent years TripAdvisor has been doing a great job of controlling and managing the pages that are created in its portal, inserting or moving them under the right categories, so that users can more easily find them and have an immediate idea of what they are talking about and what they are seeing, and that this is correct and true.

So for each macro category is shown the number of tourist attractions in each Italian location, divided by region. For the same variables we collected the number of total reviews, without distinction of language, left by users for each Italian location and with reference to each type of macro listed above. The average value of stars, which identify the quality found by users in a tourist attraction, was collected for each type of macro in each tourist location. The value of stars represents on TripAdvisor an average rating given from 1 to 5. These numbers indicate:

- 1 - Terrible
- 2 - Poor
- 3 - Average
- 4 - Very good
- 5 - Excellet

So 1 is the most negative value instead 5 is the maximum value and the most positive value that can be achieved. This value indicates the average quality perceived by all users who reviewed that tourist attraction. The selected scores influence and are used on the position of the tourist attraction in TripAdvisor's popularity rankings and/or how the attraction is positioned on the platform when searching for the location.

Then all these values, extrapolated to TripAdvisor by the WebHarvly software, created the database on which the analysis of this paper is based. The database was then exported to Excel to proceed with the analysis. In the generated Excel, the regions, locations, macros, ratings, and stars are shown in the columns. In the rows for each location, the extrapolated values are shown. The file generated after extraction contained data on 26,515 Italian locations.

3.3 Method of analysis

Starting with the database created using the WebHarbly software and extrapolating data from TripAdvisor, the next step was to process the data. First of all, it is important to understand some characteristics of the Italian territory.

According to ISTAT surveys[30], Italy is a country characterised by the presence of small municipalities, which cover 69% of the total Italian territory. But residents here are only 16.2% of the national population. In many regions, more than 70% of the land area falls under the control of these smaller municipalities. In medium-sized municipalities, most of the population lives, 68.5% of the national total. In large municipalities, on the other hand, which occupy 1.1% of the national surface area, 15.3% of the

[30] `https://www.istat.it/it/files//2020/12/C01.pdf`

population lives. In regions with a high proportion of land occupied by small municipalities, density levels tend to be low. This is particularly the case in Valle d'Aosta, where small municipalities cover 99.3% of the regional surface area and have the lowest average density.

Based on an ISTAT report[31], carried out in September 2020, there are 1,704 non-tourist municipalities in Italy, i.e. 21.5%, areas where there are no tourist attractions and/or where tourist flows are absent. The regional distribution shows a higher concentration of non-tourist municipalities in Piedmont and Lombardy due to the high number of municipalities in these regions (2,691 out of 7,926, or 34% of the total), many of which are small.

This distribution of the Italian population means that in many localities, especially in the small municipalities, there are fewer tourist attractions available to visitors than in the larger ones. This figure does not take into account the quality offered by small localities but only the quantity of opportunities offered, both for each type of tourist attraction and in relation to the number of the same type.

Given this, locations that did not satisfy the criterion of having an attraction in at least six of the nine macros analysed were removed from the dataset. Therefore, locations where the number of zeros in the nine macros was greater than six were eliminated and not considered in the analysis carried out.
Following this first compilation of data, another one was made. In this second step, data that referred to the entire region as a location was eliminated. Therefore, the rows relating to the total count of macros in each region were eliminated, 20 variables. And for the same criterion the lines concerning the total count of tourist attractions considering the Italian provinces have been removed. This was done in order to have fairness in the data analysed and therefore to consider only Italian locations, otherwise some data would have been taken into account more than once. The elimination of the rows, which did not respect the above criteria, were done directly in the Excel sheet, applying formulas so that they could be identified quickly and correctly.
So, after eliminating all the values that reflected the criteria listed above, the sample on which the analysis was carried out is 4,449 Italian places.

Following the reworking of the database, the analysis began in two steps.

[31] https://www.istat.it/it/files/2020/09/Decreto-rilancio_Classificazione-territori_16_09_2020.pdf

The first part of the work involved a quantitative analysis of the data. The objective here was to understand the number of people who reviewed each tourist attraction in relation to the number of macros present in each location. The first assumption that was used is that the number of opinions corresponds to the number of people who visited that tourist attraction. This is because it is taken into account that in order to leave a review on TripAdvisor and generally in any UGC and social platform, a person has physically been there. And therefore that the review is truthful and consistent with what the person has experienced. The hypothesis that fake reviews exist, i.e. reviews made with the aim of deceiving other users, recommending or discouraging an attraction on the basis of untrue information, is not considered. This is because in the case of false negative reviews, the probability that they were reported by the tourist attraction operator is very high. However, in the case of false positive reviews, it is very likely that users who chose the destination for the positive reviews and then had a bad experience would not only leave a very negative review but also a very positive one, if it is considered to be false. Moreover, normally fake reviews are more likely to be found in reference to hotels and restaurants, hardly any fake reviews are found on tourist attractions. The high competitiveness of the market leads to fake reviews, often between business owners; but hardly this can happen for museums, statues or places of interest.

Furthermore, no distinction is made between negative and positive reviews, only the total number of reviews on the portal for each macro is considered.
To do this, an index calculated through the ratio between the number of macros and the number of reviews was used. This index was calculated for each location in the database and for each type of macro.

The second step is a qualitative analysis of the results of the previously calculated index. For each location, the macro that had the highest index number was taken into consideration; this factor indicates people's preference to visit that specific category of macro, type of tourist attraction, in that location. At this point, stars were used as a method of comparison to compare the quality offered by the area, giving this variable the purpose of indicating the quality perceived by users who had previously used, seen and visited the tourist attraction. This gives an overview of the perceived quality from the users' point of view. The value attributed to the stars, being a score, is issued by the user taking into account all the factors that are important to him/her to be evaluated and shared. Obviously, the value reported in the database and used is the average of all the stars left by users for each macro

at each location.

So to sum up, the tourist attraction for each of the most relevant Italian locations was taken into consideration and grouped by region. Then for each region the preferred and most visited tourist attraction by TripAdvisor users was found. For each region the favourite tourist attraction with the highest number of preferences was then analysed. Then the preferred tourist attraction for each region was taken into account and the distribution of stars was considered.

This helps to understand whether the market offer corresponds to the demand, and therefore whether the needs and preferences of tourists, especially those who have visited that tourist attraction, correspond to what is offered in the area.

3.4 Findings

Starting from the analysis described above and the results obtained, the next step was to observe and discuss the results obtained. Interpreting them in order to give them meaning and to understand how they can be useful to all actors in the tourism sector. We then analyse the situation region by region and evaluate and discuss the results to see if supply and demand match or where the area should improve to ensure that tourists find the highest quality. These observations will be made from different perspectives and from different points of view, so that the study will be more valuable and can be used and exploited by more stakeholders.

With the data obtained from the analysis described above, a table was created, in order to have a more immediate view and to be able to understand the data better, so as to be able to give them a meaning. The table was created by creating a pivot table in the Excel file, which allows the data to be grouped in a customised manner, in order to extract the results concerned.

The first observation that we can make, starts from creating a map 3.1 of the Italian regions, in which in each area is indicated the preferred macro based on the data obtained from TripAdvisor and the analysis carried out, taking into account the first part of the analysis, that is, the results from the index applied to the review data. Therefore, in this map you can have an immediate visual feedback of the macro from a quantitative point of view; the set of tourist attractions present in that location grouped under the same

category, preferred by visitors who have reviewed it on the portal.

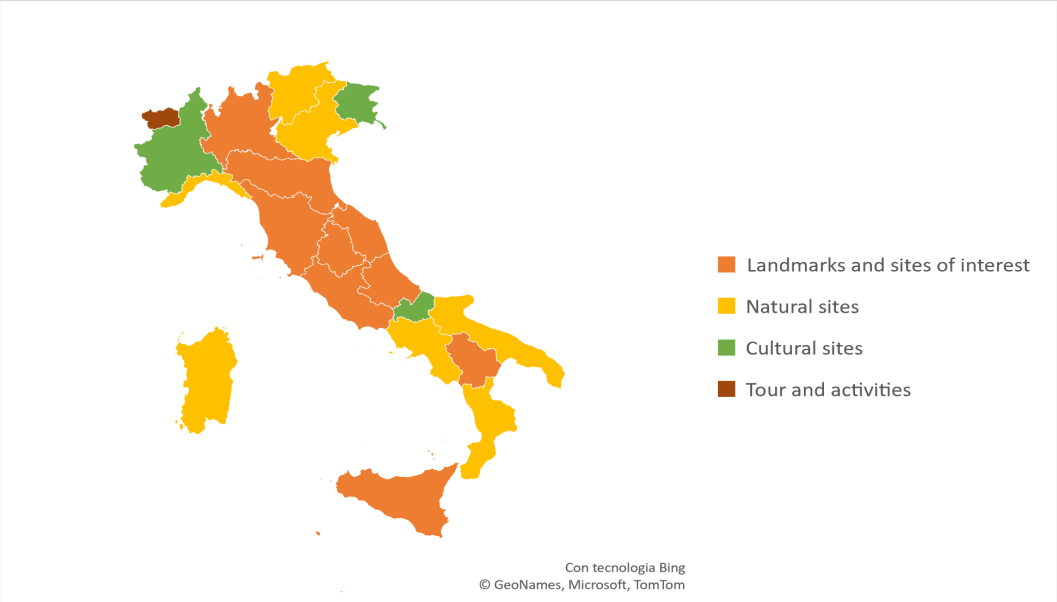

Figure 3.1: Most reviewed macros by region.

The quickest observation to make is that of the nine macro categories into which tourist attractions are divided on the TripAdvisor site, only four are identified as the favourite and best in the various Italian locations, and these are: Landmarks and sites of interest, Natural sites, Cultural sites, Tour and activities. Moreover, it can be clearly seen that these four categories are grouped into zones within the map of Italy. The most representative category of the Italian territory is "Landmarks and sites of interest", which is also the macro within which there is a greater number of tourist attractions.

At this point we begin to analyse the territories in which the macro result is the same, in order to try to understand the reason for this result and therefore what these territories have in common to make people choose them for that particular type of tourist attraction.
Let's start with the most common, namely "Landmarks and sites of interest", found mainly in the central part of the peninsula and in Sicily. Certainly, in all these regions there are major Italian cities, important from a cultural, social and artistic point of view and well known in Italy and abroad, such as Rome, Florence, Bologna, Milan, Catania, Palermo.

These large cities are the best known and the ones for which many visitors organise a visit there. They offer a wide variety of attractions in every sphere, and have some of the most well-known Italian attractions which therefore leads visitors to choose them as their destination. But these Italian cities also have much more to offer than just the most famous and touristy destinations, and this makes a visitor stay longer in the chosen place and take the opportunity to visit even smaller, or less famous and well-known attractions. All these regions of central Italy have in common an enormous wealth and cultural heritage, and this can be found not only in large cities but also in smaller towns and villages.

The small towns and villages of central Italy are becoming increasingly popular, especially in the last two years of the pandemic, as they offer tourists the opportunity to visit authentic places and to try, eat and savour typical local products. Therefore, in these regions there is a strong combination of big cities and small villages but linked by the presence of sites of interest of various kinds: artistic, cultural, socio-economic; which make tourists very attracted to these areas.

Sicily, no less, is a region which offers a very large variety of places and landscapes. Tourists who choose it as a destination can choose a more cultural tourism, visiting the important World Heritage sites in the region, or a more natural or seaside tourism. In short, in this region the offer is very wide and certainly can satisfy many requests of different types of tourists. Precisely for this reason, the preferred and most reviewed macro category in these areas is "Landmarks and sites of interest", because in this category are grouped all these places and tourist attractions that are not well defined within the other categories, because they are more transversal and can offer the tourist a more transversal experience. Besides being a more generic category.

It is interesting to note that Basilicata also has this category as a preference. When one thinks of this region, the city that immediately springs to mind is Matera, certainly the most famous and culturally rich. This city was the European Capital of Culture in 2019 and this meant that tourism in the city increased by 44% in that year thanks to this and all the events that took place. Because of this, the region is known and appreciated more and tourism in the region has increased and because of this, online reviews have also increased.
The regions in which this category is the favourite, and therefore the most important according to the analysis carried out and the data from TripAd-

visor, have a great importance from a cultural, artistic, social and economic point of view for the whole of Italy.

The second most present category is "Natural sites", found in Trentino-Alto Adige, Veneto, Liguria, Sardinia and in the South of Italy.
Starting from Northern Italy, it is easy to notice that in the north-east there is the most famous mountain range in Italy, the Dolomites. One of the most renowned naturalistic-environmental and touristic areas of the Alps, home to a national park and nine natural parks and the largest Italian ski area. Therefore, thanks to their mountains, these areas are frequented by tourists who are passionate about nature. This type of tourism and the tourist attractions that are offered are suitable for all seasons. The activities that people can do are different between winter and summer but the offer is always there. The case of Veneto is a clear example of how the perception we have does not always reflect the truth and that, in addition to traditional methods, it is important to base one's strategies on the results of Big Data analysis. When we refer to Veneto, we think of the most famous cities such as Venice, Verona, Padua and therefore we expect a tourism in this region more cultural, but the analysis made and the results obtained show that in reality the tourists in this area are very attracted by natural sites.
Just in these last two years of pandemic, the tourism sector related to the mountains and in general the whole world of outdoor has undergone an incredible increase and growth. Precisely because it reflects what people and tourists are looking for in this new phase, in which wide open spaces are preferred and less crowded.

On the contrary, the most attractive natural sites in Liguria are the natural landscapes linked to the sea. The protected natural areas of Liguria include one national park, eight regional parks and other smaller areas covering almost 12% of the regional territory.
Just like Liguria, the southern regions of Italy, therefore Campania, Calabria, Puglia, can offer tourists natural landscapes and a rich flora and fauna. In these areas are found some of the best sea areas and beaches of the whole nation. Here, however, tourism is mainly concentrated during the summer season, in which the regions fill up with tourists from all over Italy and abroad.

Sardinia is the natural site par excellence of Italy, it offers sea and uncontaminated nature; it is a real paradise for outdoor fans. Sardinia can count on a vast territory scarcely anthropized and for this reason rich in uncontaminated charm. Entire naturalistic areas reach from the hinterland to the

sea following the course of steep streams and the paths traced by the ancient charcoal burners. The coasts hide treasures unknown to most in the southwest and in Ogliastra while its capital, Cagliari contains the largest nesting site for flamingos in the entire Mediterranean. For this reason, the island, in addition to receiving a summer tourism linked to the sea, and for which the region is best known and frequented, is a destination for outdoor enthusiasts both amateur and competitive. The rocky walls are perfect for free climbing and sport climbing enthusiasts and the mysterious gorges are the base for canyoning and vertical progression training. The rivers lend themselves to kayaking and there is no shortage of those who choose it for water trekking. There are many marked hiking trails with various levels of difficulty for trekking, hiking and mountain biking. Thanks to this vast wilderness, the island offers many opportunities and different types of tourism.

Each territory in which the "Natural sites" category is predominant has some unique characteristic that leads it to have tourism related to that. Each territory, however, has its own natural characteristics and tourist attractions related to the place, tour operators must be able to enhance as much as possible what nature offers them.

"Cultural sides" is the most relevant macro category in three Italian regions: Piedmont, Friuli Venezia Giulia and Molise. In these regions, the most reviewed tourist attractions are cultural sites, mainly museums. In these areas, if we look at the thighs to do present on TripAdvisor, among the most reviewed, there are museums and foundations. In Piedmont, in fact, there is the Egyptian Museum, the oldest museum in the world, entirely dedicated to the Nilotic civilization and it is considered, for value and quantity of finds, the most important in the world after the one in Cairo.

Interestingly, "Tours and activities" is the most relevant macro in only one region, Valle d'Aosta. This category is not really a thing to do, but we can better define it as a tool to facilitate tourists. The area, despite being the smallest region in Italy, has very characteristic features and territory. It is located in the mountains and is home to Europe's highest mountain, Mont Blanc. This is why a cluster of tours and activities has been created here, which are offered to tourists in order to attract them and make the most of their visit. The tours and activities that are offered are varied and many, including those related to sport and outdoor activities, as well as those related to traditional gastronomy. It is also a region in the far north of Italy, on the border with France and Switzerland, which makes it easily accessible to both Italian and foreign tourists.

In general, the regions that have the same macro categories, in the graph above, have many things in common that attract tourists, although each territory has its own characteristics that distinguish it. Italy offers a wide range of territories and tourist attractions, here tourists can find what they are looking for and what satisfies their needs, what moves them to visit new places. This makes visitors look for what makes them want to visit a place on TripAdvisor and then leave a review after having seen, lived and experienced it.

Things to do correspond to the first motivation for users to visit a place, most of the time tourist attractions are what motivates the user to visit and then to go to a new location. This is what initially attracts the user to visit a place, but they must also find a series of services linked to the stay that allow them to stay in the designated location. This is why it is important that the things to do are highlighted with care and attention and that a network of services is created around them to offer tourists.

At this point we understand how the preferred macros are perceived by users and what quality is offered.
The average rating on TripAdvisor for Italian attractions is 4.3 stars. Through the analysis made we understand if this data is respected and the distribution of stars in each region for the most important, reviewed and visited attraction. The table below 3.2 shows the distribution in percentages of stars for the selected macro for each region.

As can be seen in general, in the selected macros, the highest percentages of star distribution are on the 4 and 5 stars.
And this indicates that most of the reviews left for the most visited destinations by users in the various regions of Italy are positive, and therefore left by users satisfied with their visit to the attraction. The stars, and therefore the review left on TripAdvisor is the summary of the whole experience lived in the tourist attraction.

We note immediately that the Piedmont region has 14% of reviews with a zero star rating, which is very negative and indicates the presence of many negative reviews. Compared to other regions where the percentage is very low and often equal to zero, in this case it has a remarkable value of attention. It can be seen as a wake-up call for tourism operators in this region as it indicates that although "Cultural sites" is the most searched, reviewed and visited macro category in the region and therefore a reason why people

choose to go there, they do not find the quality and services they are looking for.

This is the only very negative value that can be observed from the analysis in this paper. Also, the regions: Lombardy, Emilia Romagna, Sicily, Sardinia, Puglia and Calabria have a percentage between 1 and 3, in the 0 score value, so also in this case there is definitely something to do to improve the quality offered in order to bring the reviews with low score to zero, clearly however they are not as relevant as for Piedmont. So, all the regions that have reviews with 0, 1 or 2 reviews need to keep an eye on this data to ensure that they do not increase but on the contrary decrease and this can only be done by reading the reviews and understanding where the problems and critical issues were for which users have left such a low rating so that we can work on it and improve.

It must be considered that you can have many positive reviews with 5 stars, but it takes only a few negative ones, with 1 star, to change the users' opinion about the attraction. These negative reviews can happen (no one has ever gotten only positive reviews), but what matters is how they are handled by the structure directly in the platform and that they are taken into consideration to improve the service offered and in general the problem found in the review. If it is true that a negative review has much more value than a positive one, it is also true that a negative review that has an adequate response from the manager, is dampened and becomes an opportunity to do something more and be better known by users. In fact, it is significant that 89% of users read the answers to the reviews given by the managers of the structures. [32]

The percentages instead of 3 stars, indicate a neutral value, so not negative but at the same time not even positive, so most likely is to indicate that the user has not found anything special and exceptional in the place, and probably will not return and will not recommend. This is because they have not left anything amazing.

Most reviews of tourist destinations where the index of the analysis described above is higher have 4 and 5 stars. This indicates that the vast majority of the reviews are positive and that the users and therefore the tourists who visited the tourist attractions belonging to those macros had a very positive overall experience. This has led users to want to leave a very

[32] `https://www.brightlocal.com/home/`

positive review on the TripAdvisor portal, and this makes users who have to decide whether or not to visit that place by reading very positive reviews more enthusiastic and enticed to go there.

However, it should be noted that in most regions the concentration of stars is on a value of 4, indicating that very often the reviews left by users do not give the maximum possible. This may indicate that tourists who have visited these tourist destinations were not completely satisfied and enthused by the attraction and that there was something wrong with it or at least that made them put the maximum possible value of excellence.

Users want to be confident about where they go and the tourist destinations they choose to visit, because very often this decision involves spending money and time. Therefore, users inform themselves and make sure that it is worth it by reading what is said on the net, and TripAdvisor plays a central role in this, being the most important UGC site, as it is the most used and visited. Users prefer and choose to go where others have gone, because this gives a sense of security to the action. Reviews in this sense play a key role and contribute enormously to the trust that the tourist attraction is able to convey. The star rating is the first and most visually immediate thing a user sees when looking for a tourist attraction, so it is very important that the average of that thing to do is very high, because then the user will decide later to read in detail some reviews. Stars give an immediate idea and make the user who is browsing the site make a first decision. Therefore, it is essential to have the majority of reviews with a rating of 5 out of 5 stars.

Users prefer to share (and therefore recommend) products that have reviews. In fact, before deciding whether or not to read individual reviews, they look at the overall rating. If it is above three stars, they are likely to go deeper, reading the individual reviews. Below three stars, they will rate it negatively. Consumers read up to 10 reviews before deeming the business trustworthy. However, at least 40 reviews are needed to consolidate the rating.
86% of consumers admit to reading reviews. 91% of consumers identify word of mouth (and therefore reviews) as the key medium influencing their purchasing decisions. 57% consider businesses with reviews above 4 stars to be interesting. 85% of consumers disregard reviews older than 3 months.[33]

It is important to have a high score in the ratings, also because in this

[33] `https://www.brightlocal.com/home/`

way the tourist attraction will be reported and will come up first in users' searches. In this way, when users search for a location and the things to do there, those with the highest scores will come up first and make the user refer to those places. Normally only the first search pages are viewed by users, not all possible pages are looked at and therefore only those activities are preferred and then visited.
For all these reasons it is crucial to have a high score and a high number of reviews, so tourist attraction managers have to make sure that visitors who have seen and been to a location have an incentive to leave a positive review. In addition to high average stars, it is important to have many reviews, because users will have more confidence in the truthfulness and authenticity of positive opinions.

It is important to take into account the fact that this analysis is made and based on the reviews left by users on the TripAdvisor platform, and therefore certainly represents people's idea and what they perceived after experiencing and trying out the tourist attraction, but it may not reflect the totality of opinions on a given attraction. As, although increasingly widespread and used, it is not certain that all people are active players on the site and therefore always express their idea after visiting a place. So this data, which is useful to tour operators, should certainly be used and taken into account, but it should be enriched and incorporated with all the other data they have. Data from both the online and digital world, and from the offline world.

This analysis allows the interpretation of the data coming from TripAdvisor, and not only, because the same logic and analysis can be applied to any database coming from any online site, related to the tourism sector and not only. It allows us to transform Big Data information into precise knowledge about the true interests, needs and opinions of tourists. And this represents a great information resource for DMOs. On the basis of this information, expanded with information from the offline world, tourism operators can base and build their strategies. This huge amount of data, its processing, and its use must become much more used by DMOs, especially considering that most of the data they can use is available at low or no cost.

3.5 Policy implication

After having explained the analysis made and the results obtained, we will try to better understand how it can be used by tourism operators, because

as mentioned at the end of the previous chapter, tourism strategies can be based and built on the processing of data from online UGC platforms.

First of all, there is the fact that tourism in Italy is a regional matter. In fact, with the Constitutional reform of 2001, the subject 'tourism', which in the original version of Article 117 appeared among the subjects of shared competence between the State and the Regions, was transferred to the unwritten list of subjects of residual regional competence. In fact, this reform affirmed that the regions can make laws on tourism, laws that can replace state laws of principle.

At central level there is the Ministry of Tourism, which is responsible for 'the planning, coordination and promotion of national tourism policies, relations with the regions and projects for the development of the tourism sector, relations with the European Union and international tourism and relations with trade associations and tourism enterprises and consumer associations'. Over the years, this ministry has undergone many variations and integrations with other subjects. But in 2021 it was established by the last Prime Minister Mario Draghi, who separated it from the Ministry for Cultural Heritage and Activities.
This choice is very important, because as mentioned in the previous chapters, tourism plays a very important role in Italy and for the Italian economy, and it must be continuously encouraged, studied and given the importance it has. Also, because Italy is the country with the highest number of UNESCO heritage sites, 59 in all. It is important to understand, analyse and study the best strategy to make these sites become a resource for the country.

The aim of the ministry is to coordinate initiatives in the tourism sector, but then the management and decisions are left to the individual regions. This is because the Italian territory is different as is the type of tourism that each area hosts. Leaving the jurisdiction to the regions, means that each place can make the decisions and implement the strategies they deem most appropriate for their visitors, their characteristics and their tourism. Each region knows its territory well and can decide on strategies and direct investments on what it considers most appropriate for its territory. Regions must present a long-term program with a plan to boost their tourism and give a future vision of what they are aiming for.

It is important that the strategies that individual regions plan are based on data. Big Data plays a central role in this regard, but it is not enough to have a lot of data at one's disposal, as mentioned above, it must be processed

in order to make it readable and usable. The above analysis can be useful as a starting point because from here regions can understand what it is worth investing in, and therefore what tourists who decide to stay in that area are most attracted to. Using data from online platforms, in this case TripAdvisor, but the same analysis methodology can be used for any database obtained from UGC sites, where users leave evaluations and reviews, allows the competent bodies to understand the market in greater detail and depth, and above all the market demand. Understanding what people are looking for in a given location or region is fundamental to making sure that demand matches supply and especially that this is also done at an institutional level. Regions need to invest their tourism resources in what tourists are looking for in that area by helping and supporting the managers of tourist attractions.

In addition, regions can use this analysis to better understand the quality offered, in general, in regions by tourist attractions, and what people and users find and perceive after visiting the tourist attraction. This allows targeted investments to be made and helps the tourist attraction to increase and improve the quality offered, and in doing so the tourist attraction will increase its score, increase trust in reviews and appear earlier on the pages of the website. All this would lead to an increased flow of tourists and visitors to the tourist attraction.
The analysis of these online and offline big data allows the regional tourism boards to orientate themselves in the type of tourism most in demand in the area, and thus possibly understand if the region is going in the right direction or if it needs to change its tourism plan, to be more supportive of the sector and to accommodate the demands and offer requested.

After understanding how regions can use the analysis carried out, exploiting the Big Data provided by UGC sites, let's see how actors in the sector can benefit and use this.
The use of Big Data for players in the tourism sector is particularly useful because it allows predictive and behavioural analyses to be carried out. This is why the above analysis, as for regional tourism authorities, is very important when used to create strategies, to better understand the tourists attending their tourist attractions and in general to improve by understanding where visitors are affected and disappointed.

Therefore, DMOs and generally tourist attraction managers can use the analysis in this field to improve the quality offered; according to the table 3.2 with the results obtained, it is very important that this is done and considered especially by tourist attractions with percentages in the scores 0,1 and

2. So especially in the Piedmont region where there is a very high percentage in the most negative evaluations. Therefore, in these cases, attraction managers should carefully read the reviews, especially the negative ones, to understand what the inconveniences and the reasons are why visitors have left negative reviews and ratings. Because in addition to leaving negative reviews online, these tourists will certainly not spread the word offline and will not recommend this attraction, but they will give the buyer negative publicity by not recommending the visit. But in general, this data should be taken into account by everyone, even those who have received the most positive reviews, because within the reviews they can find ideas on how to always increase the quality offered.
So, the study of this analysis allows managers to understand the perception of their macro category in their region and in this way make future forecasts, on potential flows, on which potential investments can be based.

In general, processing and analyzing the big data coming from UGC sites has become a fundamental element to be able to make future forecasts and study market strategies in order to make supply and demand match more and more. It is very useful and important to have the tools to exploit this huge amount of data in favor of tourist attractions, so that we can better understand the visitors interested and the type of tourism on which it is more convenient to bet, invest and attract.

3.6 Final conclusions

After processing, analysing and interpreting the data, we return to the research question to try to give functional answers and draw conclusions from this study. First of all, thanks to the literature in the initial chapters and the analysis carried out, we have realised how crucial it is in the current online scenario to make use of the enormous amount of data coming from the various online sites. Being able to process the data and interpret it in the right way is the most important part of the process.

In this paper, analysis based on online reviews of tourist attractions in Italian locations is one type of analysis that can be done. This type of analysis is useful to understand the tourism typology of the different regions, i.e. what tourists look for in that particular location; and then the quality perceived by tourists. This is one of the many analyses that can be carried out starting from the database with the numbers of reviews of the various

macro locations in Italy left on TripAdvisor. The analysis carried out initially concentrates on the quantitative point of view of the reviews in the macros, and then analyses the results from a qualitative point of view. And as described above, these data, together with those coming from the classic feedback methods, can be used immediately by both the regional tourism authorities and the individual managers of tourist attractions.

Obviously, depending on the results you need, you can process the data differently and make analyses focused on certain aspects instead of others. From a single database you can obtain a variety of results, the important thing is to understand what you want to obtain in order to make a targeted analysis to obtain the results you are looking for.

Based on this analysis and the results obtained, further analysis can be done. First of all, a more in-depth analysis can be made for each individual region, because the preferred macro with the highest index is only one, but it would be necessary to understand how much the values of the others differ in order to have a global vision of tourism in the entire region. this can help DMOs understand how to manage and where it is appropriate to intervene with targeted strategies and investments. Moreover, it would be interesting to carry out the same analysis in a couple of years, to understand if the quality of the tourism offer has improved and if the type of tourism sought in each region has changed or not. Thus, redoing the analysis in the future would be useful to monitor the situation and check if any efforts made have been perceived and appreciated by tourists.

It would also be interesting to carry out the same analysis in other UGC sites in order to understand if the trends identified for TripAdvisor remain the same for other platforms or not. If not, it is the case to understand why the analysis of data coming from different platforms gives different results. In general, in order to have more truthful and reliable data, it would be necessary to take all these analyses and put them together with data coming from traditional channels. In this way you have a global overview of all the various types of tourists and their needs, needs and desires.

Region	Macro	Stars	Percentage
abruzzo	Landmarks and sites of interes	0	0%
		1	0%
		2	0%
		3	0%
		4	41%
		5	59%
basilicata	Landmarks and sites of interes	0	0%
		1	0%
		2	0%
		3	0%
		4	57%
		5	43%
calabria	Natural sites	0	2%
		1	0%
		2	0%
		3	7%
		4	43%
		5	48%
campania	Natural sites	0	0%
		1	0%
		2	4%
		3	6%
		4	63%
		5	27%
emilia romagna	Landmarks and sites of interes	0	3%
		1	0%
		2	0%
		3	6%
		4	69%
		5	22%
friuli venezia giulia	Cultural sites	0	0%
		1	0%
		2	0%
		3	5%
		4	76%
		5	18%
lazio	Landmarks and sites of interes	0	0%
		1	0%
		2	0%
		3	0%
		4	53%
		5	47%
liguria	Natural sites	0	0%
		1	0%
		2	0%
		3	0%
		4	52%
		5	48%
lombardy	Landmarks and sites of interes	0	3%
		1	0%
		2	0%
		3	3%
		4	53%
		5	41%
marche	Landmarks and sites of interes	0	0%
		1	0%
		2	0%
		3	3%
		4	62%
		5	36%
molise	Cultural sites	0	0%
		1	0%
		2	0%
		3	30%
		4	10%
		5	60%
piedmont	Cultural sites	0	14%
		1	0%
		2	0%
		3	3%
		4	51%
		5	32%
puglia	Natural sites	0	2%
		1	0%
		2	0%
		3	9%
		4	70%
		5	20%
sardinia	Natural sites	0	1%
		1	0%
		2	0%
		3	1%
		4	63%
		5	34%
sicily	Landmarks and sites of interes	0	2%
		1	0%
		2	2%
		3	0%
		4	63%
		5	33%
trentino alto adige	Natural sites	0	0%
		1	0%
		2	0%
		3	1%
		4	34%
		5	65%
tuscany	Landmarks and sites of interes	0	0%
		1	0%
		2	0%
		3	1%
		4	63%
		5	36%
umbria	Landmarks and sites of interes	0	0%
		1	0%
		2	0%
		3	0%
		4	56%
		5	44%
valle d aosta	Tour and activities	0	0%
		1	0%
		2	0%
		3	6%
		4	44%
		5	50%
veneto	Natural sites	0	0%
		1	0%
		2	0%
		3	4%
		4	44%
		5	52%

Figure 3.2: Distribution of stars.

Conclusions

The experiential nature of the tourist activity makes it particularly useful and important for a potential customer to have the testimony of those who have already lived this experience. The spread of User Generated Content has brought the issue of online reputation to the attention of tourism operators, completely changing the way the consumer's decision-making process works. Consumers are becoming prosumers in the new post-modern market, using social media and tourism portals to communicate online and recount their experience, constantly producing new content.

In his interaction with the commercial activity, he becomes an invaluable source of information on the perceived quality and the needs to be met: through his reviews, the relationship between supply and demand becomes two-way and the phases of service provision and the post-sales relationship become as relevant as the pre-sales phase. The credibility of reviews is very high since most Internet users consider peer-to-peer reviews reliable, one out of three tourists have planned their holidays using travel platforms and a purchase is often interrupted if the reader comes across a negative review [59].

The combination of big data and tourism has existed for a number of years now, an alliance that today has become indispensable for the recovery of a sector hard hit by the consequences of the pandemic. The post-Covid economic crisis has, in fact, taken the need for such information to the next level. The world in which the tourism industry is slowly structuring its revival is a new world.

When it recovers, international tourism will have a different face to the one we knew. At the same time, research, data and strategies used before the Coronavirus have lost their meaning and value. Faced with this scenario, the role of big data for the travel business and the induced activities generated is more than evident. The possibility of accessing real-time data, a multiplicity of information and performance indicators, makes it possible to understand

and predict the evolution of the market. In other words, the monitoring of analytics and statistics translates into data-driven relaunch strategies, i.e. effective insofar as they are based on real, concrete information. To fully understand the extreme value of such information, we need only think of the extent to which digitalisation has now penetrated all phases of travel. Booking accommodation and flights, restaurant and destination attraction reviews, maps, informative and interactive websites, proximity marketing: tourism facilities and companies have access to a huge amount of valuable data left by users more or less consciously.

On the one hand, as a direct consequence of the digital transition of which the pandemic was the accelerator. On the other, for reasons of force majeure: if tourism as we knew it is over, recovery cannot take place without a thorough understanding of its new face. Collecting and interpreting data is the only way to understand a sector that has changed dramatically: new methods and needs, new flows and behaviour, new players and power relationships.

In this report we have therefore explained how data from UGC platforms can be obtained, processed and analysed. Starting from the extrapolation of data from TripAdvisor with a Web Scraping software, we obtained a database with the numbers of tourist attractions in Italian cities up to April 2021.

Once the database was set up, according to the stable assumptions explained in the previous chapter, we moved on to the analysis phase, a first quantitative part, applying to the database an index related to the ratio between the reviews that each category of tourist attraction had obtained in the location and the number of tourist attractions of that specific category present. In this way we understood the types of tourism in the various Italian regions, highlighting the macro category with a higher index for each region. In general, which are the most visited and searched categories of tourist attractions in Italy and we tried to understand what is common to the areas where the same category has the highest index.

Based on these results we moved on to the qualitative phase of the analysis, understanding the distribution of stars, that is the ratings given by users in each review per tourist attraction, in the most relevant macro category for each region. At this point, with the results obtained from the analysis, we have tried to explain how the study of Big Data and the analysis of this thesis can be used both by tourism authorities and by all actors in the sector. This is important because it allows tourist attractions, in particular, to better understand market demand and to be able to implement strategies

and make the right investments to match supply.
According to the results, it can be said that the analysis allows us to understand everything and to transform the numbers from TripAdvisor into usable data for tourist attractions.

The monitoring, analysis and use of reviews left by visitors are a fundamental activity that every tourist attraction must do every day.
A complaint can be a useful source from which to extract suggestions for improvement and in the event that a long time after the publication of the complaint, another user reported the same defect, the attitude of silence by the manager would be counterproductive for his business. The behaviour to hold is to answer in a cordial way to the critical comment left by the dissatisfied reviewer, in such a way to strengthen his own reputation in the net and to make stable the relations with tourists. This is a good way to reinforce your reputation on the Web and stabilize your relationship with your visitors, sparking a positive and constructive debate.

In conclusion, we can say that this study confirms the need to integrate all the new technologies available in the tourism sector; using UGC sites and Big Data to their advantage in order to increasingly understand the needs and what tourists are looking for in the area to ensure that the offer fully meets the demand for each tourist attraction.

Bibliography

[1] George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier, 1978.

[2] Sinan Aral. The problem with online ratings. *MIT Sloan Management Review*, 55(2):47, 2014.

[3] Rodolfo Baggio and Jacopo A Baggio. Experiencing information asymmetries in tourism. In *4th Advances Tourism Marketing Conference, Maribor*, 2011.

[4] Annunziata Berrino. La storia del turismo in italia. *Nuova informazione bibliografica*, 8(3):539–554, 2011.

[5] Dimitrios Buhalis and Aditya Amaranggana. Smart tourism destinations enhancing tourism experience through personalisation of services. In *Information and communication technologies in tourism 2015*. Springer, 2015.

[6] MASSIMO CASA and GIOVANNI D'ALESSIO. Le statistiche della banca d'italia nell'epoca del coronavirus, 2020.

[7] Hyuk Jun Cheong and Margaret A Morrison. Consumers' reliance on product information and recommendations found in ugc. *Journal of Interactive Advertising*, 8(2):38–49, 2008.

[8] George Christodoulides, Colin Jevons, and Pete Blackshaw. The voice of the consumer speaks forcefully in brand identity: User-generated content forces smart marketers to listen. *Journal of Advertising Research*, 51(1 50th Anniversary Supplement):101–111, 2011.

[9] Pratibha A Dabholkar. Factors influencing consumer choice of a" rating web site": An experimental investigation of an online interactive decision aid. *Journal of Marketing Theory and Practice*, 14(4):259–273, 2006.

[10] Terry Daugherty, Matthew S Eastin, and Laura Bright. Exploring consumer motivations for creating user-generated content. *Journal of interactive advertising*, 8(2):16–25, 2008.

[11] Richard A D'Aveni, Giovanni Battista Dagnino, and Ken G Smith. The age of temporary advantage. *Strategic management journal*, 31(13):1371–1385, 2010.

[12] Andrea De Mauro, Marco Greco, and Michele Grimaldi. A formal definition of big data based on its essential features. *Library Review*, 2016.

[13] Chrysanthos Dellarocas. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, 49(10):1407–1424, 2003.

[14] Banca d'Italia. Indagine sul turismo internazionale, 2020.

[15] The Economist. The world's most valuable resource is no longer oil, but data. *The Economist: New York, NY, USA*, 2017.

[16] Organisation for Economic Co-operation and Development. *Data-driven innovation: Big data for growth and well-being*. OECD Publishing, 2015.

[17] Ulrike Gretzel, Marianna Sigala, Zheng Xiang, and Chulmo Koo. Smart tourism: foundations and developments. *Electronic markets*, 25(3), 2015.

[18] Dogan Gursoy, Giacomo Del Chiappa, and Yi Zhang. Preferences regarding external information sources: a conjoint analysis of visitors to sardinia, italy. *Journal of Travel & Tourism Marketing*, 34(6), 2017.

[19] Mariann Hardey. Generation c: content, creation, connections and choice. *International Journal of Market Research*, 53(6):749–770, 2011.

[20] Stephanie Hays, Stephen John Page, and Dimitrios Buhalis. Social media as a destination marketing tool: its use by national tourism organisations. *Current issues in Tourism*, 16(3):211–239, 2013.

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

[22] Juan M Hernández, Andrei P Kirilenko, and Svetlana Stepchenkova. Network approach to tourist segmentation via user generated content. *Annals of Tourism Research*, 73, 2018.

[23] Martin Hilbert. How much information is there in the "information society"? *Significance*, 9(4):8–12, 2012.

[24] Simon Hudson and Karen Thal. The impact of social media on the consumer decision process: Implications for tourism marketing. *Journal of Travel & Tourism Marketing*, 30(1-2), 2013.

[25] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[26] Nancy Keates. Deconstructing tripadvisor. *Wall Street Journal*, 1(4), 2007.

[27] Nikolaos Korfiatis, Elena García-Bariocanal, and Salvador Sánchez-Alonso. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217, 2012.

[28] Sandeep Krishnamurthy and Wenyu Dou. Note from special issue editors: Advertising with user-generated content: A framework and research agenda. *Journal of Interactive Advertising*, 8(2):1–4, 2008.

[29] Doug Laney et al. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

[30] Andres V Lerner. The role of'big data'in online platform competition. *Available at SSRN 2482780*, 2014.

[31] Jingjing Li, Lizhi Xu, Ling Tang, Shouyang Wang, and Ling Li. Big data in tourism research: A literature review. *Tourism Management*, 68, 2018.

[32] Alberto Lluch-Lafuente and Marco Righi. *Internet e web 2.0*. UTET università, 2011.

[33] Elena Marchiori and Lorenzo Cantoni. The role of prior experience in the perception of a tourism destination in user-generated content. *Journal of Destination Marketing & Management*, 4(3), 2015.

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

[35] Estela Marine-Roig and Salvador Anton Clavé. Tourism analytics with massive user-generated content: A case study of barcelona. *Journal of Destination Marketing & Management*, 4(3), 2015.

[36] Ana María Munar and Jens Kr Steen Jacobsen. Motivations for sharing tourism experiences through social media. *Tourism management*, 43, 2014.

[37] Peter O'connor. Managing a hotel's image on tripadvisor. *Journal of hospitality marketing & management*, 19(7):754–772, 2010.

[38] Mustafa Öz. Social media utilization of tourists for travel-related purposes. *International Journal of Contemporary Hospitality Management*, 2015.

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

[40] Simonetta Pattuglia and Sergio Cherubini. Social media marketing: consumatori, imprese, relazioni. *Social media marketing*, pages 1–152, 2012.

[41] Andrea Petrella, Roberto Torrini, Guglielmo Barone, Enrico Beretta, Emanuele Breda, Rita Cappariello, Giuseppe Ciaccio, Laura Conti, Francesco David, Petra Degasperi, et al. Turismo in italia: numeri e potenziale di sviluppo [tourism in italy: Figures and potential for development]. 2019.

[42] Alessandro Prunesti. *Social media e comunicazione di marketing. Presidiare la Rete, costruire relazioni e acquisire clienti con gli strumenti del web 2.0*, volume 36. FrancoAngeli, 2013.

[43] Tullio Romita. *Argomenti di sociologia del turismo*. Welcome Chiamami città, 2000.

[44] Michael Schroeck, Rebecca Shockley, Janet Smart, Dolores Romero-Morales, and Peter Tufano. Analytics: The real-world use of big data. *IBM Global Business Services*, 12(2012):1–20, 2012.

[45] Peter Sondergaard, Ronald Coase, and Tim Berners-Lee. " information is the oil of the 21st century, and analytics is the combustion engine. *Gartner Research, Datasciencecentral. com*, 2019.

[46] Shrihari Sridhar and Raji Srinivasan. Social influence effects in online product ratings. *Journal of Marketing*, 76(5):70–88, 2012.

[47] David W Stewart and Paul A Pavlou. From consumer response to active consumer: Measuring the effectiveness of interactive media. *Journal of the Academy of Marketing Science*, 30(4):376–396, 2002.

[48] European Data Protection Supervisor. *Privacy and Competitiveness in the Age of Big Data: The Interplay Between Data Protection, Competition Law and Consumer Protection in the Digital Economy*. European Data Protection Supervisor, 2014.

[49] Thales Teixeira and Leora Kornfeld. Managing online reviews on tripadvisor. 2013.

[50] Armando Travaglini, Simone Puorto, and Vito D'Amico. *Digital marketing turistico: e strategie di revenue management per il settore ricettivo*. LSWR, 2015.

[51] Catherine Tucker. Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility. *Review of Industrial Organization*, 54(4):683–694, 2019.

[52] Tracy L Tuten, Michael R Solomon, Luciano Pilotti, and Alessandra Tedeschi-Toschi. *Social media marketing: post-consumo, innovazione collaborativa e valore condiviso*. Pearson, 2014.

[53] Egbert Van der Zee and Dario Bertocchi. Finding patterns in urban tourist behaviour: A social network analysis approach based on tripadvisor reviews. *Information Technology & Tourism*, 2018.

[54] Egbert Van der Zee, Dario Bertocchi, and Dominique Vanneste. Distribution of tourists within urban heritage destinations: A hot spot/cold spot analysis of tripadvisor data as support for destination management. *Current Issues in Tourism*, 23(2):175–196, 2020.

[55] Camilla Vásquez. Complaints online: The case of tripadvisor. *Journal of Pragmatics*, 43(6):1707–1717, 2011.

[56] Zheng Xiang and Ulrike Gretzel. Role of social media in online travel information search. *Tourism management*, 31(2), 2010.

[57] Kyung-Hyan Yoo, Yoonjung Lee, Ulrike Gretzel, and Daniel R Fesenmaier. Trust in travel-related consumer generated media. *Information and communication technologies in tourism 2009*, pages 49–59, 2009.

[58] Benxiang Zeng and Rolf Gerritsen. What do we know about social media in tourism? a review. *Tourism management perspectives*, 10, 2014.

[59] Ziqiong Zhang, Qiang Ye, Rob Law, and Yijun Li. The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, 29(4):694–700, 2010.

[60] Bo Zhao. Web scraping. *Encyclopedia of big data*, 2017.