BIG DATA ANALYTICS IN PUBLIC ADMINISTRATION FOR TOURISM: A PRELIMINARY STUDY FOR TÜRKİYE[1]

Diego RODRIGUEZ[2] & Luis PINEDA[3]

Abstract

This preliminary study explores the use of Big Data analytics in the public administration of Türkiye's tourism sector, focusing on the extraction and analysis of visitor reviews from the TripAdvisor platform. The primary objective is to compile a comprehensive database of records for 262 museums and archaeological sites, managed by the Turkish Ministry of Culture and Tourism, distributed across 64 provinces. Using web scraping techniques, over 72,500 reviews written in English were collected, representing approximately 49% of the total available reviews for these attractions. he methodology involves sentiment analysis to assess visitor perceptions, highlighting both positive and negative feedback. The analysis identifies common themes such as the admiration for historical and cultural significance, and criticisms regarding logistical issues like ticket prices and crowd management. The study emphasizes the importance of such detailed feedback in informing public decision-making processes, aiming to enhance the quality and appeal of Türkiye's diverse tourism offerings. Findings from this analysis reveal that reviews are heavily concentrated in major tourist hubs such as Istanbul, Antalya, and Izmir, with attractions like the Ayasofia Museum/Mosque and Topkapı Palace receiving the most feedback. This geographical and thematic concentration of reviews provides valuable insights into visitor preferences and areas requiring improvement. The study concludes that integrating Big Data analytics into public administration can significantly benefit the tourism sector, leading to better-informed policies and strategic enhancements that could boost tourism revenues and visitor.

Keywords: Big Data, Tourism, Web Scraping, Sentiment Analysis, Türkiye.

KAMU YÖNETİMİNDE TURİZM İÇİN BÜYÜK VERİ ANALİTİĞİ: TÜRKİYE İÇİN BİR ÖN ÇALIŞMA

Öz

Bu ön çalışma, TripAdvisor platformundan ziyaretçi yorumlarının çıkarılması ve analizine odaklanarak, Türkiye'nin turizm sektörünün kamu yönetiminde Büyük Veri analitiğinin kullanımını araştırmaktadır. Birincil amaç, T.C. Kültür ve Turizm Bakanlığı tarafından yönetilen ve 64 ile dağılmış 262 müze ve ören yeri için kapsamlı bir kayıt veri tabanı derlemektir. Web kazıma teknikleri kullanılarak, bu turistik yerler için mevcut toplam yorumların yaklaşık %49'unu temsil eden İngilizce yazılmış 72,500'den fazla yorum toplanmıştır. Metodoloji hem olumlu hem de olumsuz geri bildirimleri vurgulayarak ziyaretçi algılarını değerlendirmek için duygu analizini içermektedir. Analiz, tarihi ve kültürel öneme duyulan hayranlık ve bilet fiyatları ve kalabalık yönetimi gibi lojistik konulara ilişkin eleştiriler gibi ortak temaları tanımlamaktadır. Çalışma, Türkiye'nin çeşitli turizm olanaklarının kalitesini ve cazibesini artırmayı amaçlayan kamusal karar alma süreçlerinin bilgilendirilmesinde bu tür ayrıntılı geri bildirimlerin önemini vurgulamaktadır. Bu analizden elde edilen bulgular, yorumların İstanbul, Antalya ve İzmir gibi başlıca turizm merkezlerinde yoğunlaştığını ve Ayasofya ile Topkapı Sarayı gibi cazibe merkezlerinin en çok geri bildirim alan yerler olduğunu ortaya koymaktadır. İncelemelerin bu coğrafi ve tematik yoğunlaşması, ziyaretçi tercihleri ve iyileştirilmesi gereken alanlar hakkında değerli bilgiler sağlamaktadır. Çalışma, Büyük Veri analitiğinin kamu yönetimine entegre edilmesinin turizm sektörüne önemli ölçüde fayda sağlayabileceği ve turizm gelirlerini ve ziyaretçileri artırabilecek daha iyi bilgilendirilmiş politikalara ve stratejik iyileştirmelere yol açabileceği sonucuna varmaktadır.

Anahtar Kelimeler: Büyük Veri, Turizm, Web Kazıma, Duygu Analizi, Türkiye.

---

[1] The first version of this paper is presented at the II Congreso del Caribe de la Asociación Latinoamericana de Estudios de Asia y África – ALADAA Colombia Conference in 2022.

[2] Diego Rodriguez, Instituto Distrital de Turismo - IDT – Bogotá, diarodriguezgu@gmail.com, ORCID: 0000-0002-1283-8412

[3] Luis Pineda, Instituto Distrital de Turismo - IDT – Bogotá, Lfpinedag.stat@gmail.com, ORCID: 0009-0000-2928-5706

Introduction

In Türkiye, the tourism sector represents 5.1% of GDP (2022), being one of the main drivers of the economy, due to the variety of tourism products offered (Beach Tourism; Culture; Health and Thermal; Winter; Mountain; Caves; Hunting; Golf; Yachting; Religion; Air Sports; Mountaineering; Rafting-Diving; and Birdwatching). Based on existing literature on the use of Big Data in public administration and public decision-making processes, this paper proposes research of this type applied to the tourism sector of Türkiye, particularly using the technique of "Web Scraping"[4]. The aim is to create a database of records for which a sentiment analysis will be carried out based on the perceptions visitors have of the different attractions that make up the Eurasian country's tourism offer. From this exercise it is hoped that the characteristics and conditions of this tourism offer of Türkiye will be known, and the result will be a very useful input for public decision-makers in the sector.

To this end, a list of 262 museums and archaeological sites was initially compiled, which are located in 64 of the 81 provinces of Türkiye, all of which have a common characteristic, namely that they are managed by the Turkish Ministry of Culture and Tourism[5]. For these 262 tourist attractions, more than 72,500 reviews were extracted, which are publicly accessible and are located on the TripAdvisor website, a platform that keeps the history of reviews and provides content related to the travel experiences of visitors. The collected reviews are written in English and represent about 49% of the total reviews available on the platform (approximately 155,000 reviews) for the selected tourist attractions.

On the other hand, this paper is framed in the digital transformation of public administration, for which a Big Data analysis is applied to the tourism sector, implementing technological developments that have been generated in recent decades, in the exercise carried out, through the positive reviews it is possible to identify the good practices that make an attraction popular or with an excellent rating, This would allow standards to be replicated in the attractions with the greatest lag, where, additionally, from the opinions with low scores, shortcomings/adverse effects are identified, which can guide well-informed improvement actions supported by objective quantitative evidence, by public entities, concluding that having more and high quality information, will be reflected in better decision-making.

In the first chapter of this study, a literature review was conducted on the use of Big Data in the tourism sector. In the second part, the importance of the tourism sector for Türkiye was analysed based on TUIK[6] data. In the third chapter of the study, the characteristics of the selected tourism attractions in Türkiye were emphasised. In this last chapter the reviews written on TripAdvisor were analysed by web-scraping and through sentiment analysis co-occurrence network graphs were made.

1. Literature

Tourism is one of the most studied sectors both in the world and in Türkiye. In addition to the fact that the sector is multidimensional, it is an important factor that also affects the politics and economy of countries. From this point of view, studies generally focus on different topics such as hotels, restaurants, archaeological sites, travel agencies and transportation. Additionally in recent years, the use of Big Data has begun to become widespread in works on tourism and its effect on different topics. For example, analysing reviews made by visitors on different platforms is an important example of the use of Big Data in tourism analysis.

According to the literature reviewed for this study (Table 1), it was determined that TripAdvisor was used as the main source in the studies where visitor reviews were analysed. On the other hand, the studies reviewed focuses on different areas of the service sector and especially on tourism activities. Some of the studies discussed tourist attractions, while others were about tourist restaurants and travel guides.

Table 1. Studies in which visitor reviews are used for the analysis of tourist attractions.

Author(s)	Source	Approach	Data extraction Method	Type of reviews	Total number of reviews analysed
Doğan & Demirtaş (2021)	TripAdvisor	Hatay, Türkiye	Not available	Restaurants	282

---

[4] Also called "web data extraction" or "web data mining".

[5] Taken from 2019 Museum and Archaeological Site Visitor Statistics.

[6] Türkiye İstatistik Kurumu (Turkish Statistical Institute)

Balıkoğlu et al. (2020)	TripAdvisor	Türkiye	Not available	Tour guides at archaeological sites	15,217
Sunar (2021)	TripAdvisor	Uzungöl, Türkiye	Not available	Tourist Attraction	450
Ünal (2020)	TripAdvisor	Patara, Türkiye	Data Mining	Tourist Attraction	5,288
Mancı & Tengilimoğlu (2021)	TripAdvisor	Göbeklitepe, Türkiye	Manually	Tourist Attraction	518
Valdivia et al. (2017)	TripAdvisor	Spain	Not available	Tree Tourist Attraction	45,303
Almeida & Abrantes (2018)	TripAdvisor	Minas Gerais, Brasil	Web scraping	Tourist Attraction	1,384 attractions
Fang et al. (2015)	TripAdvisor	New Orleans, United States	Not available	Tourist Attraction	41,061 reviews for 106 Attractions
Flôres Limberger et al. (2014)	TripAdvisor	Brasil	Not available	Hotels	660 reviews
Agüero-Torales et al. (2019)	TripAdvisor	Granada, Spain	Web scraping	Restaurants	33,594 reviews

Source: Authors' own elaboration.

As can be seen in Table 1, for example, Doğan and Demirtaş (2021) examined 282 reviews of restaurants in Hatay province of Türkiye, while Balıkoğlu et al (2020) examined 15,217 reviews about tourist guides in archaeological sites in Türkiye. On the other hand, Sunar (2021), who also conducted a study on Türkiye, focused on 450 reviews about tourist attractions in Uzungöl. On the other hand, Agüero-Torales et al (2019) and Valdivia et al (2017), who conducted studies on Spain, examined 33,594 and 45,303 reviews, respectively. What is noteworthy in these two studies is that Valdivia et al (2017) analysed only three tourist attractions, namely La Alhambra, Sagrada Familia and Cordoba Mosque. Another one is that Agüero-Torales et al (2019) focused on reviews of restaurants in Granada province. As can be seen from this, in some cases, even if what is analysed is limited, the recognition of the place can lead to a large number of comments from users. This increases the richness of studies using similar analysis methods.

2. Methods

For this study, the reviews were scraped from TripAdvisor, a widely used platform for travel reviews, ensuring a large and diverse dataset. Only English-language reviews were selected for scraping. This choice was made to streamline the sentiment analysis process and avoid the complexities associated with multi-language processing. Additionally, a comprehensive dataset was built by extracting over 72,500 reviews for 262 museums and archaeological sites in Türkiye. This represents approximately 49% of the total reviews available for these attractions on TripAdvisor. On the other hand, these selected tourist attractions are located in 64 out of the 81 provinces of Türkiye, ensuring a broad geographic representation of the country's tourism sector. On the other hand, all selected tourist attractions are managed by the Turkish Ministry of Culture and Tourism. This criterion was set to maintain consistency in the types of attractions analysed and to focus on those under direct governmental oversight.

Data Collection Process

The web scraping process was conducted using Python libraries such as Selenium, which is well-suited for extracting data from web pages. The reviews were collected programmatically, ensuring that data extraction was efficient and scalable. Additionally, the scraped reviews included fields such as review text, rating, date of visit, and reviewer profile information (e.g., travel group composition). Moreover, NLP techniques were employed to process and analyse the textual data from the reviews. This included tokenization, lemmatization, and stop-word removal to prepare the text for sentiment analysis using a software called KH coder for the efficiency of the process. Which is a lexicon and rule-based sentiment analysis tool that is particularly effective for social media texts. It was used to determine the polarity (positive, negative, neutral) and intensity of the sentiments expressed in the reviews.

Data Analysis

Additionally using the same software KH Coder, co-occurrence network graphs were created to visualize the relationships between words and phrases in the reviews. This helped identify common themes and sentiments associated with different attractions. Thus, by analysing positive reviews, the study identified good practices and features that contribute to high visitor satisfaction. These insights can be used to replicate successful strategies across other attractions. On the other hand, negative reviews were scrutinized to identify common complaints and areas needing improvement. This data-driven approach enables targeted interventions to enhance visitor experiences. Moreover, the analysis also included categorizing reviews based on visitor demographics (e.g., couples, friends), providing valuable insights into the preferences and behaviours of different tourist segments. By employing these web scraping selection criteria and sentiment analysis techniques, the study provides a robust framework for understanding tourist perceptions and improving the management and promotion of Türkiye's diverse tourism attractions.

Validations

Thus, for the further validations, the results were checked for consistency across different subsets of the data. For instance, the sentiment analysis results for Istanbul were compared with those for Antalya and Ephesus to ensure that similar patterns and trends were observed in comparable contexts. Additionally, the co-occurrence network graphs created with KH Coder were validated by comparing them with manually identified key themes and relationships in the reviews. This step ensured that the automated text mining accurately represented the underlying sentiment and themes. By implementing these validation techniques, the study ensured that the sentiment analysis results were robust and reliable. The insights gained from these validated results provided a solid foundation for making informed policy decisions and strategic initiatives in the tourism sector. The validated sentiment analysis results allowed for the identification of actionable data points, such as common complaints about entrance fees and waiting times, as well as positive aspects like historical significance and aesthetic appeal. These insights are crucial for targeted improvements and enhancing the overall visitor experience. Regular validation and recalibration of sentiment analysis models are essential for maintaining accuracy over time. This continuous improvement process ensures that the Ministry of Culture and Tourism can rely on up-to-date and precise data to guide their decision-making processes.

3. Insight Results
3.1. The importance of the Tourism sector in Türkiye

The service sector, especially tourism, is one of the most important resources of the Turkish economy. Tourism revenues correspond to approximately 5% of the country's gross domestic product. At the same time, when employment is analysed according to economic activities in Türkiye, it is observed that the service sector is the sector that generates the most employment.

According to data published by the Turkish Statistical Institute between the years 2018-2023 (see graph 1), more than 50% of the country's formal workers are employed in the service sector. In 2018, this rate was 54%, but as of 2023, it has increased to 57%. On the other hand, although the number of employees in the service sector remained stable in 2020, the year when the Covid-19 pandemic was prevalent, and in 2021, when the effects of the pandemic continued, it has generally increased each year. Industry and agriculture, which follow the service sector, have an employment rate around 15%-20%. However, while there has been an increase in the percentage of those employed in the industrial sector over the years, the situation is the opposite for the agricultural sector. In 2018, the proportion of agricultural sector employees was around 20%, but by 2023, it had decreased to 15%. Finally, the construction sector has an employment rate of less than 10%. Similar to the agricultural sector, the number of employees in the construction sector has also shown a decline over the years. Consequently, the service sector has become the most significant sector with a steadily increasing number of employees. In other words, the service sector is one of the most important sectors in the country. Therefore, it can be said that the Turkish economy is highly dependent on the service sector due to the large employment gap with other sectors. Although it is not reflected in this graph, the tourism sector experienced a significant setback following the aircraft crisis with Russia in 2016, which had both political and economic repercussions.
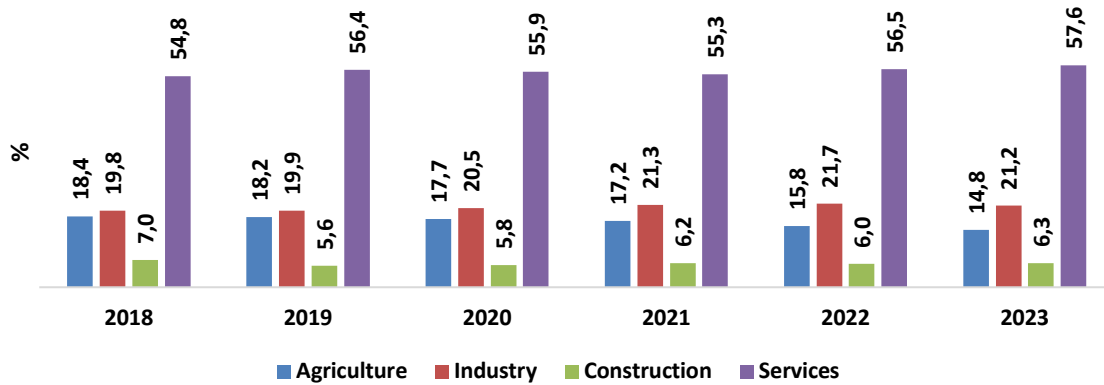
Figure 1. Employment by economic activity (2018-2023) (%)

Source: Own elaboration with data from TUIK, 2023.

Offering diverse tourism products such as Beach Tourism; Culture; Health and Thermal; Winter; Mountains; Caves; Jazz; Golf; Yachting; Religion; Air Sports; Mountaineering; Rafting-Diving and Bird Watching, Türkiye is among the most visited countries in the world. As shown in Figure 2, the number of visitors in 2001 was approximately 13 million, while by 2019 this number exceeded 50 million. The number of visitors to Türkiye generally has an increasing trend over the years. However, there were declines, especially in the years 2016 and 2020. The main reason for the decline in 2016 was, as mentioned earlier the aircraft crisis between Türkiye and Russia. The decline in 2020 is due to the COVID-19 pandemic, which affected all tourism worldwide. However, with the opening again in 2021, it can be observed that the number of the international visitors have been increasing. In 2020, the number of visitors decreased to 12 million, but it rose to 24 million in 2021, 44 million in 2022, and approximately 49 million in 2023.
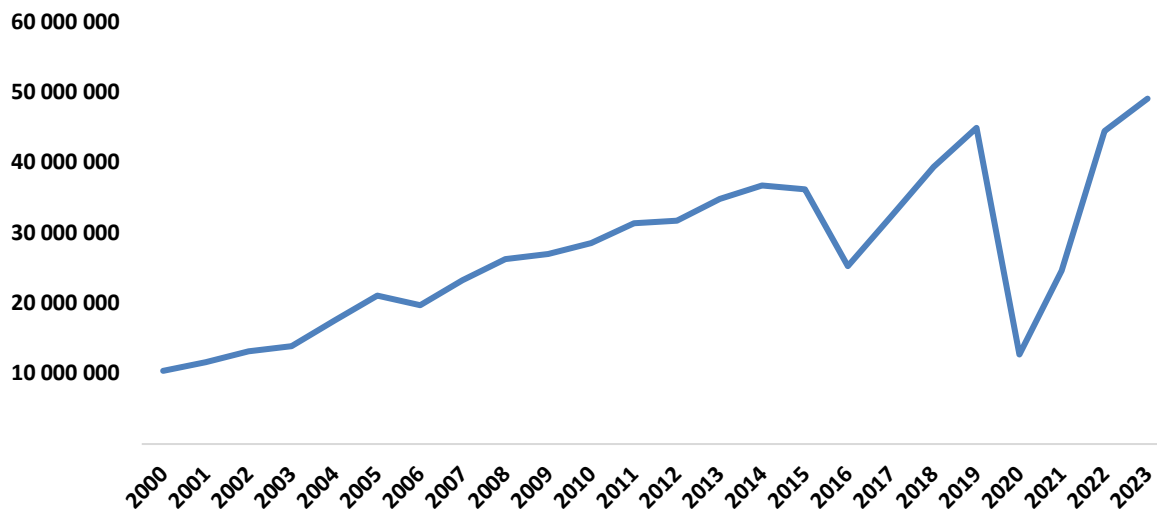
Figure 2. Number of international visitors (2001-2023)

Source: Own elaboration with data from TUIK, 2023.

As tourism is one of the most important sectors of the country's economy, tourism revenues are also significantly important for Türkiye. According to TUIK data, tourism revenues were over 10 million USD in 2001 and it is around 55 million USD in 2023. In 2021, after the pandemic, while the country and the sector were still in recovery process, tourism revenues were around 30 million USD, however, these revenues of $30 million are relatively low. When examining Figure 3, it would not be wrong to say that this value is the average. Especially since the country earned around US$41 million from tourism in 2014 and US$42 million in 2019, and these are the highest values earned in the last 20 years before the pandemic. Of course, the

effect of the COVID-19 pandemic on the decline in 2021 is undeniable. After the pandemic in 2020, the tourism sector was still trying to regain its former momentum. As can be seen in the figure, the recovery process was completed rapidly, and Türkiye's revenue from tourism was measured at approximately 49 million dollars in 2022 and 55 million dollars in 2023.
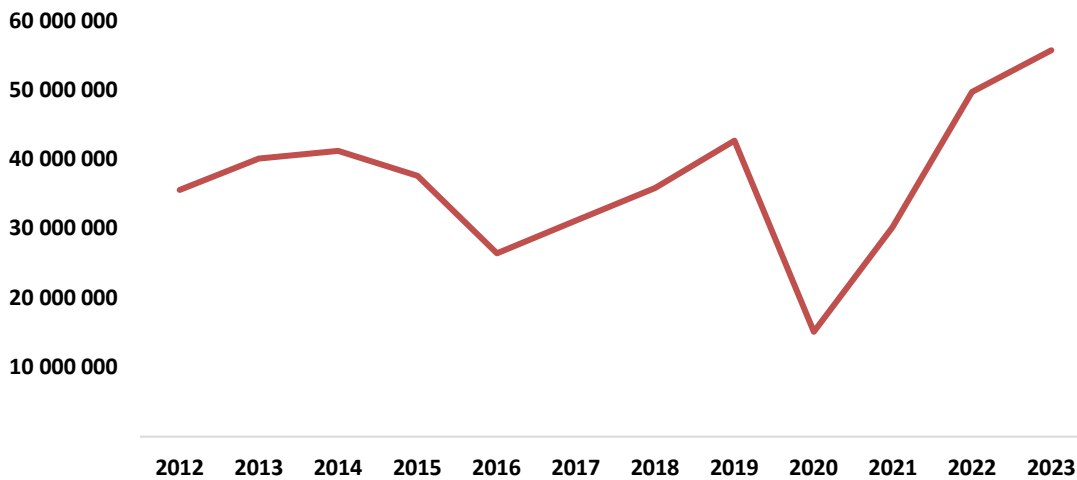
Figure 3. Tourism revenues (2012-2023) (Dollars)

Source: Own elaboration with data from TUIK, 2023.

Although changes in the country's tourism receipts over the years are clearer, the average expenditure of inbound tourists did not change much between 2012-2023. As shown in Figure 4, the average per capita expenditure was between US$750-1,000. Among the years analysed here, the year with the lowest average expenditure was 2018 with 787 dollars. Even in 2016, when the aircraft crisis with Russia took place, the average expenditure was USD 846. However, the year with the highest spending was 2013 with $1,024. On the other hand, from 2013 onwards, there is a steady decline in the measured value, bottoming out in 2018 as mentioned earlier. This suggests a continuous negative trend over these years. After hitting the lowest point in 2018, the average expenditure per capita shows a gradual recovery over the next two years, increasing to 958 dollars by 2020. Additionally, there is a sharp and significant rise after the pandemic, reaching the highest value around 1,032 dollars of average expenditure per capita. This indicates a rapid improvement or growth after the Covid-19. However, following the peak in 2021, the value drops again in 2022, though it remains higher than the lowest points in previous years, around 971 dollars. From 2022 to 2023, the average expenditure shows a slight upward trend, stabilizing at around 970 dollars by 2023.

Figure 4. Average expenditure per capita (2001-2023) (Dollars)

Source: Own elaboration with data from TUIK, 2023.
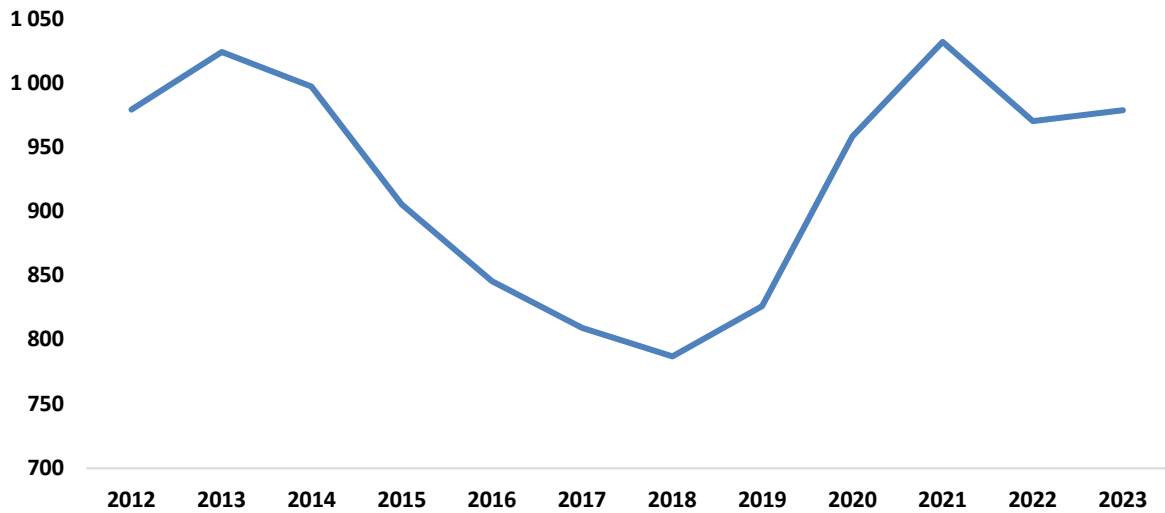
3.2. Characterisation of attractions in Türkiye

Although Türkiye is often noted for its beach tourism, the things the country has to offer in the name of tourism are not limited to only summer tourism. With its deep history and culture, the country is also home to many museums and archaeological structures. According to a list published by the Ministry of Tourism and Culture in 2019, Türkiye is home to a total of 262 museums and archaeological sites located in 64 of the country's 81 provinces that are controlled by the state. Of course, all these cities and the cities that appear on the map as having 0 attraction centres are home to many more attraction centres than mentioned here. Some of them are run by the city municipalities, while others are run by the private sector. However, since only the tourism attractions operated by the ministry are considered in this study, the others are not included in the map. The reason for considering the list published in 2019 in the study is that the ministry has not activated the list since 2019. For this reason, although it is known that the operation of some attraction centres has been transferred to the ministry, especially after the pandemic, they are not included in this study since the list has not been renewed.

According to this list, as can be seen in the first map, the province with the most attractions in the country is Muğla, a city on the Aegean coast of the country, with 25 attractions. This province is followed by Antalya, the country's most important coastal tourist province, with 24 attractions. Additionally, there are 16 attractions in Nevşehir, which is home to the Cappadocia region, one of the country's most prominent tourist areas as well as Mersin province which is located in Mediterranean region of Türkiye. Finally, as it can be seen in the following map Istanbul, Türkiye's best-known city has 12 attractions, which are under the management of the Ministry of Tourism and Culture in 2019.

Map 1. Number of tourist attractions per province (2019)

Source: Own elaboration with data from the Ministry of Culture and Tourism of Türkiye (2019).

Considering the annual number of visitors per city, it can be seen that Istanbul, which has fewer attractions than many cities, ranks first with approximately 8 million visitors per year. Istanbul is followed by Konya with just over 3.5 million visitors and 10 attractions. Konya, one of the largest provinces in the Central Anatolia region, is also one of the most conservative provinces in the country. At the same time, the fact that the tomb of the world famous Mevlana Celaleddin Rumi is located in this city makes Konya a religious tourist destination. The third most visited city in the country is Antalya, which has 24 attractions and one of the most important summer tourism destinations of Türkiye as also it can be seen on Map 2 Antalya.

Map 2. Number of visitors to tourist attractions by province (2019)

Source: Own elaboration with data from the Ministry of Culture and Tourism of Türkiye (2019).

3.3. Application of Big Data analysis to Türkiye's tourism sector

For this paper, web scraping is the methodology used for data collection, which can be defined as "the construction of an agent (or bot) to download, analyse and organise data from the web in an automated way" (vanden Broucke & Baesens: 2018). In these section that can be considered as results section, only the information of the 262 Museums and archaeological sites which are managed by the Ministry of Culture and Tourism of Türkiye and are located in 64 of the 81 provinces of Türkiye as presented in Map 1 earlier will be taken into account.

Map 3 presents the number of reviews or comments in English on TripAdvisor for the selected attractions, which were mapped according to the province in which they were located. Thus, attractions in Istanbul were the most commented on TripAdvisor with approximately 43,000 reviews. Among them, the Hagia Sophia Museum has about half of the reviews made. On the other hand, İzmir follows Istanbul as the province with the most reviews with more than 6,500, while Antalya, one of the most visited provinces especially for its location on the Mediterranean coast and for having 24 tourist attractions, ranks third with more than 6,000 reviews. On the other hand, Cappadocia, another important attraction centre, follows Antalya with approximately 4,300 reviews by TripAdvisor users.
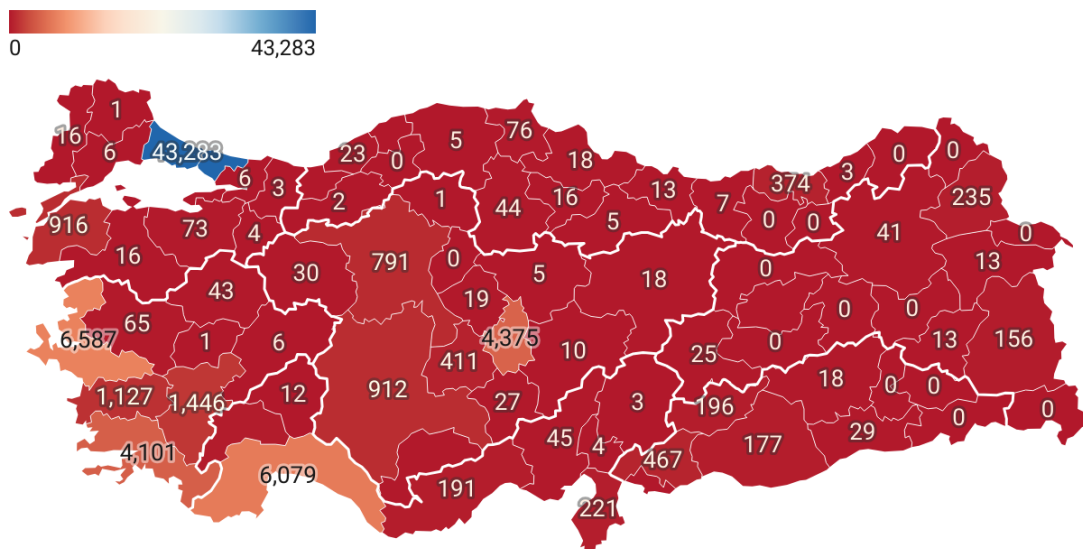
Map 3. Number of reviews of tourist attractions by province (2005-2022)

Source: Own elaboration with TripAdvisor 2022 data.

As mentioned above the tourist attraction with the most reviews on the Trip Advisor platform for Türkiye is: Ayasofia Museum/Mosque which is located in Istanbul and has a total of 22,652 representing 31% of the total reviews collected in this paper, Topkapı Palace Museum also located in Istanbul is the second attraction with the most reviews with about 14,000. Figure 5 presents the top 8 tourist attractions with the highest number of reviews, which represent 70.5% of the total number of reviews collected. It was also possible to identify that 205 attractions have less than 100 reviews, representing only 4.4%. In this first characterisation by province and by attraction, it can be observed that the reviews are mainly concentrated in 5 provinces and 8 attractions, which gives an indication of the preferences of tourists and visitors.
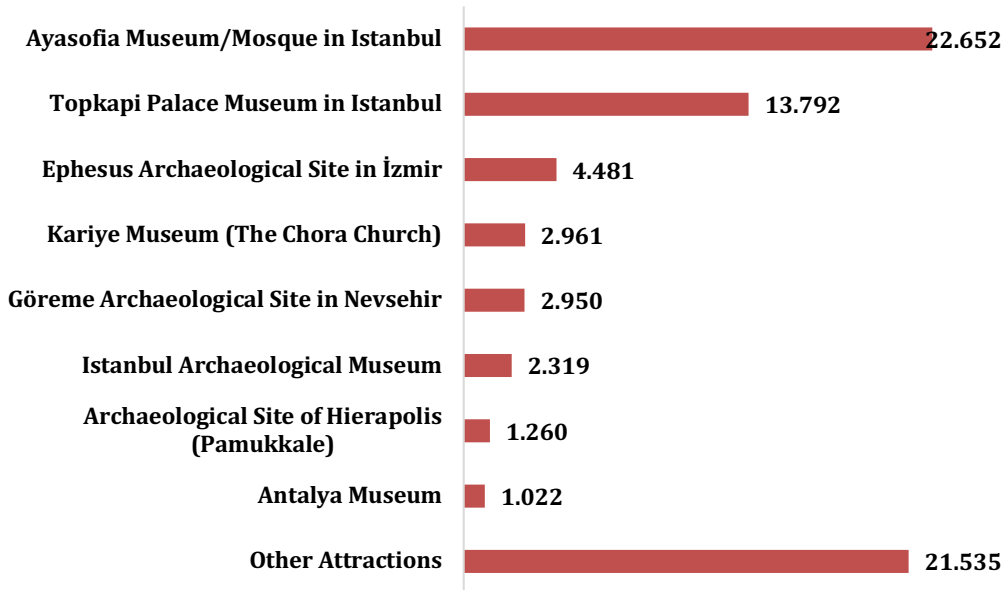
Figure 5. Attractions with the highest number of reviews (2005-2022)

Source: Own elaboration with TripAdvisor 2022 data.

On the hand Figure 6 presents the behaviour over time of the number of reviews for the selected attractions in Türkiye, where it is observed that the period between 2012 and 2016 is where the platform reached the highest popularity and more than 8,000 reviews were identified per year, with 2015 being the year where the most reviews were observed, reaching a little more than 15,000. However, as can be seen in the figure, it started to decline in the following years. So much so that reviews made after the pandemic is around 2,000 reviews per year.
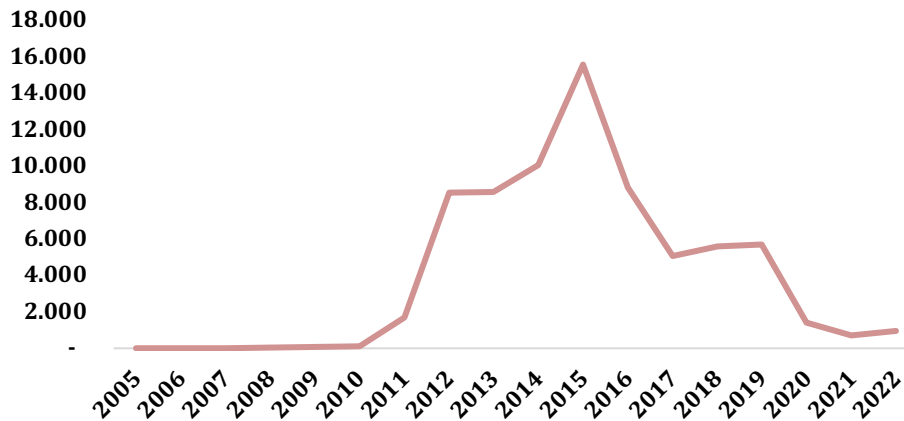
Figure 6. Number of reviews per year (2005-2022)

Source: Own elaboration with TripAdvisor 2022 data.

Regarding the travel group or people with whom the tourist or visitor visits the attraction, 42.6% of the reviews are mainly made by people who travel with their partner, followed by those who visit with friends (22.1%) and visitors who attend with their family (21.4%). Finally, and only representing 11.3%, are those attending alone or travelling alone. On the Trip Advisor platform, the information on the business motive is located within the travel group information, although this motive is not related to the different travel groups.
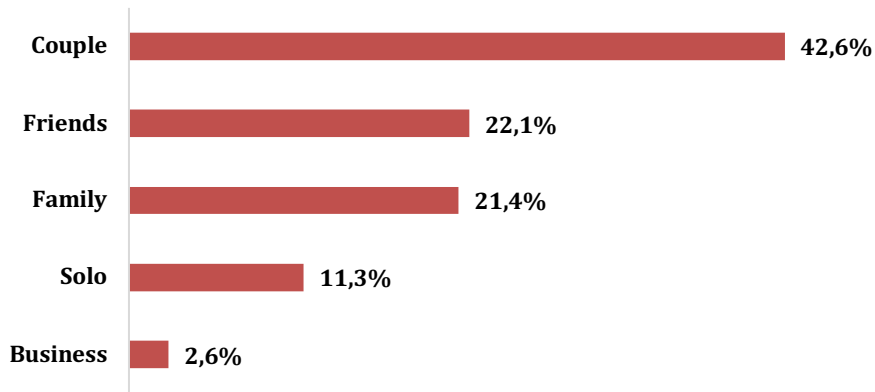
Figure 7. Visitor type (2005-2022)

Source: Own elaboration with TripAdvisor 2022 data.

The ratings given by users of the platform to each attraction, which are associated with the reviews, allow us to categorise the comments, which for this paper are considered as follows: values greater than three as positive reviews (3<x), values of three as neutral reviews (3=x) and values less than three as negative reviews (3>x). The distribution of the reviews is found to be 90.6% positive reviews, 7.3% neutral reviews, and 2.1% negative reviews, which would present an imbalance in the database.
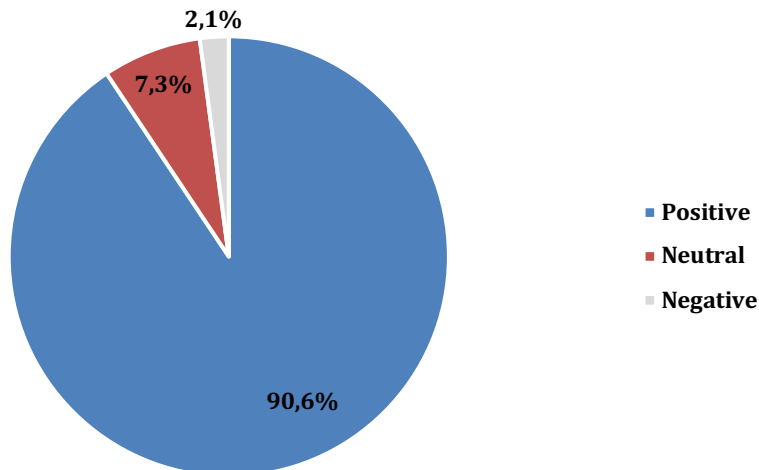
Figure 8. Classification of reviews by type (2005-2022)

Source: Own elaboration with TripAdvisor 2022 data.

Finally, to be more specific at the Turkish level, for the provinces of Istanbul and Antalya and for the attractiveness of the Ephesus Archaeological site, a text analysis of the reviews was carried out using Natural Language Processing (NLP) techniques, where words associated with positive and negative reviews are identified, and co-occurrence network graphs, a method for analysing text that includes a graphical visualisation of potential relationships between people, organisations or concepts, are used.

The different co-occurrence networks were created with KH Coder, which is a free software for quantitative content analysis or text mining, it is also used for computational linguistics and can analyse text in several languages such as English, Spanish, French, German, Italian, Japanese, Korean, Portuguese, Russian and others.

At the country level and for the more than 66,600 positive reviews, it was identified that the words "place", "beauty", "visit", "history", "museum" and "amazing" are the most relevant, are mainly associated

with the city of Istanbul and present another perspective on the importance of this city in tourism in Türkiye. As Istanbul accounts for more than 50% of the reviews extracted, its attractions such as Hagia Sophia, the Mosques and Topkapı Palace play an important role in the country-wide analysis. On the other hand, "tours" and "guides" are frequently mentioned in the reviews, especially in connection with Istanbul.

Figure 9. Co-Occurrence Network of Positive Reviews for Türkiye (2005-2022)

Source: Own elaboration with TripAdvisor 2022 data.

Regarding the negative reviews at the country level, the words "time", "payment", "entrance", "extra-payment", "buy-tickets", "money", "closed", "renovation" and "restoration", among others, are identified. This indicates that issues such as the high price at the time of buying tickets tend to have a negative outlook on visitors. However interestingly some words that have been identified in the positive reviews also identified in the negative reviews such as "museum" and "visit", "time". Negative connections of these words are often associated with time that people wait for the entrance, money that they spent or waste for the place, and finally how the places are full of people.

Figure 10. Co-Occurrence Network of Negative Reviews for Türkiye (2005-2022)

Source: Own elaboration with TripAdvisor 2022 data.

On the other hand, a more specific analysis has been made for Istanbul, Antalya and Izmir, although this has not been graphicalised. Therefore firstly, possible co-occurrence networks from the positive and negative reviews extracted is generated for the province of Istanbul. And not surprisingly, when only Istanbul is analysed, it is observed that the results obtained are quite similar to the country-wide results, mainly due to the number of reviews of the city's attractions. On the other hand, where the results differed from the country-wide results were the positive comments made to the Blue Mosque among the mosques in Istanbul.

In the following phase, it was analysed the negative and positive reviews of the province of Antalya, which is located on the Mediterranean Sea and is the second province with the highest number of archaeological sites, museums, and tourist attractions. According to the findings obtained for Antalya, words such as "ruins", "Perge", "Aspendos", "history", "visit", "amazing", "beautiful", "view" and "beach" were the most prominent in positive reviews made by TripAdvisor users. On the other hand, the frequency of negative comments about Antalya is quite low compared to the those for the country as a whole or for Istanbul. In fact, the comments generally concentrated on issues such as the warm sea water, the fact that some archaeological sites accept visitors on self-book, and the fact that the entrance to some archaeological sites is paid.

Finally, the reviews about the ancient city of Ephesus in Izmir, a city on the shores of the Aegean Sea, were examined in detail, and the co-occurrence networks of positive and negative reviews were analysed. Accordingly, the words that attracted the most attention in positive reviews were "amazing", "history", "ancient time", "visit", "ruins" and "bike". On the other hand, in the negative reviews, although the frequency is quite low, the comments about the entrance process and the entrance fee are noteworthy.

Lastly it can be said that regular validation and recalibration of sentiment analysis models are essential for maintaining accuracy over time. This continuous improvement process ensures that the Ministry of Culture and Tourism can rely on up-to-date and precise data to guide their decision-making processes. On the other hand, by thoroughly validating the sentiment analysis results, the study provides a

reliable and comprehensive understanding of tourist perceptions, enabling more effective and data-driven improvements in Türkiye's tourism sector.

Conclusion

In this exercise of approaching new technologies and methodologies, we applied them to the reviews and comments of visitors to 262 tourist attractions (museums and archaeological sites) administered by the Ministry of Culture and Tourism of Türkiye. These attractions are part of the tourist offer of the Eurasian country. It was possible to determine the perception of tourists and visitors using Natural Language Processing (NLP) techniques, in addition to generating a characterization (travel group and preferred attractions) of tourists and visitors.

A group of 5 provinces was identified: Istanbul, Izmir, Antalya, Muğla and Nevşehir, which contain about 80% of the reviews towards the attractions used in the study, with Istanbul having the most reviews (43,000 reviews) among them. On the other hand, a total of 8 attractions accounts for 70.5% of the total number of reviews collected on the Trip Advisor platform used, with the Hagia Sophia Museum/Mosque, followed by the Topkapı Palace Museum, both located in Istanbul, having the highest number of reviews. Additionally, a total of 205 attractions were identified as having less than 100 reviews, representing only 4.4% of the total number of reviews collected.

On the other hand, it is found that 42.6% of the reviews are made by people travelling with a partner, followed by those visiting with friends with 22.1%. Moreover, although it has been determined that the reviews made by Tripadvisor users are generally positive, it has been observed that the frequency of negative reviews is also quite high, especially in the country in general and in Istanbul in particular. It has been observed that these negative reviews are generally related to renovations that take place in the ruins or museums, entrance fees, ticketing processes and long waiting times.

Finally, the use of new methods such as Big Data analysis and application of data science tools such as data mining/web scraping in tourism research at the national level plays an important role, as their application and execution produce important information which cannot be collected through survey application processes or manual collection. Therefore, information is generated that can guide well-informed improvement actions supported by objective quantitative evidence, by public entities (in this case, the Ministry of Culture and Tourism of Türkiye, the entity that manages the attractions analysed), concluding that having more and high-quality information will be reflected in better decision-making for the future.

On the other hand, firstly to leverage the full potential of Big Data and web scraping techniques, the Ministry of Culture and Tourism should develop and implement a comprehensive digital strategy for tourism. This strategy would include establishing a dedicated data analytics team within the Ministry. This team would continuously monitor and analyse online reviews and other relevant data sources. By doing so, the Ministry can gain real-time insights into tourist satisfaction and emerging trends, allowing for more responsive and adaptive management of tourist attractions. Following these steps, the Ministry should implement a targeted improvement program for attractions that have fewer than 100 reviews. Using sentiment analysis to identify common issues in these underperforming attractions, resources can be allocated to address specific areas such as enhancing visitor experience, improving facilities, or promoting these sites more effectively. This approach can help elevate the overall quality and appeal of a broader range of attractions, potentially distributing tourist traffic more evenly across the country. As well, to manage visitor flow and reduce congestion at high-traffic sites like Hagia Sophia and Topkapı Palace, the Ministry should develop policies such as introducing timed entry tickets, optimizing ticketing processes, and providing real-time updates on waiting times and crowd levels through a dedicated mobile app. Improving the visitor experience at these popular sites can help maintain their appeal and reduce negative reviews related to long waiting times and ticketing issues.

Additionally, the Ministry should adjust its marketing strategies to cater to the predominant visitor demographics, such as couples and friend groups. By developing targeted marketing campaigns and packages that appeal specifically to these groups, offering tailored experiences and promotions, the Ministry can attract more tourists and improve satisfaction rates. Aligning marketing efforts with the preferences and behaviours of key visitor segments ensures more effective outreach and engagement. The Ministry should establish a rapid response protocol for addressing common complaints, such as issues with renovations, entrance fees, and ticketing processes. Creating a feedback loop where negative reviews are systematically reviewed and addressed by the relevant authorities within a set timeframe will proactively manage and respond to negative feedback. This approach can improve overall tourist satisfaction and reduce the recurrence of common issues. Integrating sustainable tourism practices into the management of

tourist attractions is essential. Also informed by data on visitor behaviours and preferences, the Ministry can implement eco-friendly initiatives such as waste reduction, energy-efficient facilities, and promoting lesser-known attractions to reduce environmental impact. Sustainable tourism practices help preserve cultural and natural heritage and appeal to a growing segment of environmentally conscious travellers'.

Lastly to attract a more diverse range of visitors and ensure everyone can enjoy Türkiye's rich cultural heritage, the Ministry should develop policies to improve accessibility and inclusivity at tourist attractions. By using data to identify barriers faced by various visitor groups, such as people with disabilities and elderly tourists, and implementing measures to enhance accessibility, the Ministry can make tourist sites more accessible and welcoming to all visitors. As conclusion it can be said that by implementing these policy changes and strategic initiatives, the Ministry of Culture and Tourism of Türkiye can leverage the insights gained from Big Data analysis to enhance the overall tourist experience, promote lesser-known sites, and ensure sustainable growth in the tourism sector.

References

Agüero-Torales, M. M., Cobo, M. J., Herrera-Viedma, E., & López-Herrera, A. G. (2019). A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor. Procedia Computer Science, 162, 392-399.

Almeida, R. de & Abrantes, R.M. (2018). The development of tourism indicators through the use of social media data: the case of Minas Gerais. Brazil Information Research, 23(4), paper 805.

Balıkoğlu, A., Ulusoy-Yıldırım, H., & Hacıoğlu, N. (2020). Turist Rehberlerine Yönelik Elektronik Yorumların İçerik Analizi: TripAdvisor Örneği. Gümüşhane Üniversitesi Sosyal Bilimler Enstitüsü Elektronik Dergisi, 11(2), 385-401.

Doğan, S., & Demirtaş, S. (2021). Tripadvisor.com'da Yer Alan Restoranlara Yönelik Yabancı Yorumların İçerik Analizi ile Değerlendirilmesi: Hatay İli Üzerine Bir Araştırma. Gastroia: Journal of Gastronomy and Travel Research, 5(2), 199-212.

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. Tourism Management(52), 498-506.

Flôres Limberger, P., dos Anjos, F. A., de Souza Meira, J. V., & dos Anjos, S. J. G. (2014). Satisfaction in hospitality on TripAdvisor. com: An analysis of the correlation between evaluation criteria and overall satisfaction. Tourism & Management Studies, 10(1).

Guerin, O. (2020). Hagia Sophia: Türkiye turns iconic Istanbul museum into mosque. Obtenido de BBC News, İtanbul: https://www.bbc.com/news/world-europe-53366307

Mancı, A. R., & Tengilimoplu, E. (2021). TripAdvisor Ziyaretçi Yorumlarının İçerik Analizi: Göbeklitepe Örneği. Türk Turizm Araştırmaları Dergisi, 5(2), 1525-1545.

Sunar, H. (2021). Çevrimiçi Değerlendirmelere Yönelik İnceleme: Uzungöl Örneği. International Journal of Tourism and Social Research, 6(7), 46-55.

T.C. Kültür ve Turizm Bakanlığı. (2019). Döner Sermaye İşletmesi Merkez Müdürlüğü 2019 Yılı Müze ve Örenyeri Ziyaretçi İstatistikleri. Obtenido de T.C. KÜLTÜR VE TURİZM BAKANLIĞI DÖNER SERMAYE: http://dosim.kulturturizm.gov.tr/

TUIK. (2021). Eğitim, Kültür, Spor ve Turizm. Obtenido de İstatistik Veri Portalı: https://data.tuik.gov.tr/Kategori/GetKategori?p=Egitim,-Kultur,-Spor-ve-Turizm-105

Ünal, A. (2020). Turistlerin Destinasyon Genel Yorumlarının Karma Yöntemle Değerlendirilmesi: Patara Örneği. En E. Karaçar, Turizm Alanında Güncel Konular ve Araştırmalar. Konya: Çizgi Kitabevi Yayınları.

Uslu, A., & Akay, B. (2019). Assessing The Effect Of International Relations On Tourism Demand In The Context Of Türkiye-Russia Aircraft Crisis. Journal of Tourism and Services, 10(18), 63-78.

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment analysis on tripadvisor: Are there inconsistencies in user reviews? En International Conference on Hybrid Artificial Intelligence Systems (págs. 15-25). Cham: Springer.

vanden Broucke, S., & Baesens, B. (2018). Practical "Web Scraping" for Data Science: Best Practices and Examples with Python. Apress.

Yücel, A. G., Köksal, C., Acar, S., & Icon, L. A.-A. (2022). The impact of COVID-19 on Türkiye's tourism sector: fresh evidence from the fractional integration approach. Applied Economics, 54(27), 3074-3087.