

Article

StreetLines: A Smart and Scalable Tourism Platform Based on Efficient Knowledge-Mining

Georgios Alexandridis ^{1,*}, Georgios Siolas ², Tasos Papagiannis ², George Ioannou ²,
Konstantinos Michalakis ³, George Caridakis ³, Vasileios Karyotis ^{2,4,*} and Symeon Papavassiliou ²

¹ Department of Digital Industry Technologies, National and Kapodistrian University of Athens, 34400 Psachna, Greece

² School of Electrical & Computer Engineering, National Technical University of Athens, Zografou Campus, 15780 Athens, Greece; gsiolas@islab.ntua.gr (G.S.); tasos@islab.ntua.gr (T.P.); geoioannou@islab.ntua.gr (G.I.); papavass@mail.ntua.gr (S.P.)

³ Department of Cultural Technology and Communication, University of the Aegean, University Hill, 81100 Mytilene, Greece; kmichalak@aegean.gr (K.M.); gcari@aegean.gr (G.C.)

⁴ Department of Informatics, Ionian University, 49100 Corfu, Greece

* Correspondence: gealexandri@uoa.gr (G.A.); karyotis@ionio.gr (V.K.)

Abstract: Identifying and understanding visitor needs and expectations is of the utmost importance for a number of stakeholders and policymakers involved in the touristic domain. Apart from traditional forms of feedback, an abundance of related information exists online, scattered across various data sources like online social media, tourism-related platforms, traveling blogs, forums, etc. Retrieving and analyzing the aforementioned content is not a straightforward task and in order to address this challenge, we have developed the StreetLines platform, a novel information system that is able to collect, analyze and produce insights from the available tourism-related data. Its highly modular architecture allows for the continuous monitoring of varying pools of heterogeneous data sources whose contents are subsequently stored, after preprocessing, in a data repository. Following that, the aforementioned data feed a number of independent and parallel processing modules that extract useful information for all individuals involved in the tourism domain, like place recommendation for visitors and sentiment analysis and keyword extraction reports for professionals in the tourism industry. The presented platform is an outcome of the StreetLines project and apart from the contributions of its individual components, its novelty lies in the holistic approach to knowledge extraction and tourism data mining.

Keywords: tourism platform; knowledge mining; sentiment analysis; keyword extraction; recommender systems; semantic enrichment



Citation: Alexandridis, G.; Siolas, G.; Papagiannis, T.; Ioannou, G.; Michalakis, K.; Caridakis, G.; Karyotis, V.; Papavassiliou, S. StreetLines: A Smart and Scalable Tourism Platform Based on Efficient Knowledge-Mining. *Digital* **2024**, *4*, 676–697. <https://doi.org/10.3390/digital4030034>

Academic Editor: Elpiniki I.

Papageorgiou

Received: 18 June 2024

Revised: 1 August 2024

Accepted: 8 August 2024

Published: 11 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The digital age has revolutionized travel planning, with an unprecedented wealth of tourism-related information readily available online. From comprehensive travel booking platforms offering detailed listings of accommodation, attractions, and activities to social media platforms where users share personal experiences and recommendations, travelers are inundated with a vast array of options. These resources empower individuals to conduct in-depth research, compare prices, and curate personalized itineraries tailored to their preferences.

Tourism-related reviews and information available online are invaluable not only to travelers but also to businesses and authorities. For businesses, especially those in the hospitality and tourism sectors, these reviews serve as real-time feedback mechanisms, providing insight into customer satisfaction, preferences, and areas needing improvement [1]. Positive reviews can enhance a business's reputation and attract more customers, while negative feedback highlights potential issues that need addressing. Furthermore, businesses can

analyze trends in reviews to adapt their services to meet changing consumer demands. For authorities, online tourism data are crucial for informed decision-making and strategic planning [2]. By examining reviews and social media discussions, the authorities can measure the success of tourism initiatives, identify emerging tourist trends, and understand the economic impact of tourism in specific regions. These data help in the allocation of resources, the development of infrastructure, and the formulation of policies that enhance the tourism experience while ensuring sustainable development. Additionally, the authorities can use this information to promote lesser-known destinations, thereby distributing the economic benefits of tourism more evenly. Overall, the digital footprint of tourism-related information empowers both businesses and authorities to make data-driven decisions that enhance the overall travel experience and foster economic growth.

Crawling online resources for tourism-related reviews and information presents significant challenges due to the vastness and diversity of the data sources [3]. The sheer volume of content spread across multiple platforms, such as TripAdvisor, Google Places, travel blogs, forums and online social media, requires substantial effort and planning in collecting and processing it. Additionally, the aforementioned platforms often use varied application programming interfaces (APIs) and data formats, further complicating data extraction and standardization [4]. Reviews may contain unstructured text, images, ratings and metadata, each necessitating different parsing techniques. The dynamic nature of the web, with frequent updates and the introduction of new content, introduces an extra challenge, as crawlers must continuously monitor and retrieve up-to-date data to maintain relevance. Moreover, many websites employ anti-scraping measures, such as CAPTCHAs and IP rate limiting, to protect their content, which can hinder automated crawling efforts. Consequently, despite the potential wealth of available sources, crawling and effectively utilizing tourism-related reviews and information online demand the development of robust methodologies, as well as careful consideration of ethical practices.

The proposed StreetLines platform, an outcome of the StreetLines project [5], tries to address the issues discussed above through the development of a versatile and modular information system that can harvest data from a growing variety of open platforms of touristic content and online social media, to perform the required analysis and to automatically produce insights for the strategic design of actions and policies that are expected to bring about better and more adequately distributed support of tourist destinations and services. In the context of this project, the following appropriate methodologies have been designed, along with software tools that have been developed and integrated into a platform: (i) efficient data collection from websites and online social media, (ii) functional representation and storage of multi-modal data, (iii) innovative knowledge extraction algorithms, (iv) sentiment analysis methodologies on free-text reviews and (v) algorithms for generating recommendations.

The generated insights can be used in many ways, such as to identify the current habits/desires of visitors to a particular place, the type of content and the communication channels they prefer in order to determine the most efficient promotion strategy (advertising and promotions) of a tourist service or product. It can also be exploited to analyze current and emerging trends (accommodation, diet, events, etc.), consumer interests and passions (what excites them, what are their priorities, e.g., environment, pets, etc.), and their concerns (e.g., what they fear most about a trip) in order to define and personalize the products offered accordingly.

The rest of this paper is structured as follows. In Section 2, relevant literature on tourism platforms is presented. Section 3 outlines the overall platform architecture, illustrating the flow of information between the provided services. Section 4 presents the constituent parts of the proposed platform, with each component being analyzed in terms of its operation and innovation. In Section 5, an initial evaluation of the platform as a whole, as well as some of its components, is conducted. Section 6 discusses limitations and possible extensions and finally the paper is concluded in Section 7.

2. Related Work

In the current section, we first review tourism-related data mining platforms in Section 2.1. More specifically, we examine tourism-related platforms in terms of their data sources, and whether they come from a single source or are aggregated from multiple sources. Then, in Sections 2.2 and 2.3 we place the emphasis on two of the core analytics services offered by the StreetLines platform: Keyword Extraction and Sentiment Analysis, respectively.

2.1. Tourism-Related Data Mining Platforms

Even though many works in the literature mine tourism-related data for analysis purposes, the vast majority rely on a single source, raising concerns about how representative and valid the obtained results are [6]. For example, many works focus their analysis on specific online social media, ignoring other sources (other online social media, the web, etc.). In [7], the authors only examined Twitter (now X) when considering the popularity of POIs and proceeded to apply text analytics methodologies to extract useful insights from the data. Additionally, they produced a number of visualizations that were helpful in their analysis. In [8], tourist destinations were mined from geo-tagged Flickr photos in order to identify POIs, in an effort to quantify visitor travel experience and identify preferences.

Platforms that extract tourism and culture-related information from heterogeneous data sources do exist, but quite often they place the emphasis of their analysis on specific aspects of the available data, not offering a holistic approach with regard to specific areas or regions. A typical example is BITOUR [9], which, despite integrating data from four online sources (Twitter, Openstreetmap, Tripadvisor and Airbnb), performs only basic data transformations, processing and visualizations; further analysis is limited to sentiment extraction from the visitor-generated textual reviews, as well as the visualization of frequently visited places.

There also exist cases of platforms that both consider heterogeneous data sources and perform more thorough analysis, employing big data techniques, but they limit their focus to specific cities or regions, failing to produce more robust tools. In [10], the authors describe a pipeline for big data analytics for the city of Barcelona in Spain. They crawled online social media and travel blogs, retrieving more than 100,000 textual reviews written in the English language by visitors and tourists over a period spanning a decade. After cleaning and pre-processing the available data, they focused their analysis on a single landmark, La Sagrada Familia church. In [11], the authors crawled data from TripAdvisor for the city of Pokhara in Nepal for a period of two years. Then, they applied unsupervised learning techniques in order to locate clusters in the data that would help them in correlating them with other dimensions of their collected data, such as ratings and reviews.

2.2. Keyword Extraction

The domain of keyword extraction in touristic data has seen significant attention in recent years due to the expanding reliance on digital platforms for travel-related information. Various works [12,13] have focused on the extraction of key terms and phrases from travel reviews, forums, and booking platforms to facilitate information retrieval, sentiment analysis and content summarization in the tourism domain.

The authors of [14] proposed a Long Short-Term Memory (LSTM)-based neural network model in order to extract key phrases from reviews collected from various online sources. In this approach, the raw data are initially preprocessed and labeled to be able to train the network in a supervised manner. The input is first passed through a pretrained BERT (Bidirectional Encoder Representations from Transformers) model used as an encoder and the output embeddings are then fed to a bidirectional LSTM. The model effectively combines the BiLSTM architecture with Conditional Random Field as the final layer to produce relevant keyphrases from the review texts.

Methodologies not relying on deep learning have also been employed in the context of mining useful information from travel reviews. In [15], keyword extraction is

performed on tourist reviews in Hokkaido. An initial pre-processing step is carried out in this case, and irrelevant information and duplicates are also excluded. Subsequently, two different techniques are employed to extract the main points of each review, namely the term frequency–inverse document frequency (TF-IDF) and TextRank algorithms. The results indicated that even more lightweight approaches like TF-IDF may extract quite useful information in terms of the topics discussed and the main spots of attention in touristic sights.

In contrast to the approaches outlined above, in the StreetLines platform, keyword extraction is performed by transformer-based models (Section 4.2) and more specifically by the DistilRoBERTa model [16]. Even though transformer-based approaches have been used on tourism-related data before [17], none of them, to the best of our knowledge, has used distilled versions of larger models, whose main advantages are the reduced model size and the faster inference.

2.3. Sentiment Analysis

When delving into sentiment analysis within touristic data, researchers have explored a spectrum of methodologies to discern opinions and emotions expressed in user reviews. Numerous studies [18] have investigated the sentiment detection of traveler feedback, aiming to decipher the nuanced perspectives and experiences shared across various destinations and accommodations. These analyses have employed diverse approaches, leveraging machine learning algorithms, natural language processing techniques, and clustering methods to extract sentiments and key insights from the vast volume of user-generated content in the tourism domain.

In [19], the authors explore sentiment analysis in hotel reviews sourced from a prominent travel site within the tourism domain. This highlighted the increasing importance of social media as a substantial data source for user opinions. The study introduces an automated sentiment detection approach using the Fuzzy C-means clustering algorithm to analyze hotel reviews. Several machine learning techniques, including Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Logistic Regression and Random Forest, were applied to classify sentiments within these reviews. Additionally, an ensemble learning model combining these classifiers was proposed and evaluated, offering a comparative analysis of their performance. Overall, the paper presents methodologies to automatically assess sentiments in hotel reviews from a significant travel platform, employing diverse machine learning techniques and an ensemble approach for classification.

Another attempt in which several machine learning algorithms are used to extract sentiment from tourists' reviews is presented in [20]. Traditional machine learning and deep learning techniques are employed in order to predict both the reviews' sentiments and ratings (in a discrete form) from a set of reviews crawled from TripAdvisor. Among the tested architectures, the bidirectional LSTM achieved the highest accuracy in both tasks, outperforming Naive Bayes, SVMs, 1-D CNNs and simple LSTMs, highlighting the advantages of deep learning algorithms over classical methodologies. Additionally, the importance of the training dataset size on the overall performance of the algorithm is emphasized, indicating the need for automated extraction tools on review platforms.

Unlike current approaches and following the reasoning of the previous subsection (Section 2.2), in the current work, sentiment analysis is carried out by DistilRoBERTa [16], a transformer-based model. Its main advantages for the task at hand are (i) the attention mechanisms that uncover complex dependencies between review text and annotated sentiment and (ii) the compact size of the transformer, which speeds up training and inference.

3. System Architecture

Prior to the introduction of the overall system architecture and its individual components, the theoretical framework supporting the tourism mining platform needs to be discussed. In principle, tourism mining is an interdisciplinary subject involving many areas, such as data science, information retrieval, natural language processing (NLP) and

tourism studies, with the objective of systematically collecting, analyzing and interpreting tourism-related information. Data can be either user-generated (e.g., reviews in free-text form, photos and videos posted online), device-generated (e.g., GPS positioning, mobile) or transactional (e.g., web searches, online booking) [13]. In the realm of the current research, the emphasis is placed on user-generated content (UGC).

The main functionality of the presented platform would be to utilize advanced data mining techniques to crawl and aggregate data from diverse sources of UGC, such as review sites, social media, blogs and forums. NLP algorithms would be employed to process unstructured text data, extracting key themes, sentiments and trends. Machine learning models could be used to classify and predict tourist preferences and behaviors, providing personalized recommendations and insights. Additionally, the framework would incorporate data integration tasks to ensure that the most relevant and high-quality information is available to all interested parties.

From a tourism studies perspective, the framework should aid in understanding traveler behavior, destination image formation and service quality evaluation, ensuring that the data mining processes align with the specific needs and contexts of the tourism industry. The platform would also emphasize user interface design to present data in an intuitive and actionable manner, allowing businesses and policymakers to easily interpret and leverage insights. Ethical considerations, such as the prevention of biased or misleading information, would be integral, ensuring the platform’s credibility and user trust. By integrating these multidisciplinary approaches, the theoretical framework would enable a robust, efficient and user-centric tourism mining platform capable of transforming vast digital data into valuable, actionable insights for enhancing the travel experience and supporting strategic decision-making in the tourism sector.

3.1. The StreetLines Platform

Figure 1 provides a high-level overview of the overall StreetLines architecture. At its core, it is a modular architecture based on processing queues that provide decoupling capabilities; producers do not need to know anything about consumers. Additionally, queues provide an ideal way to create asynchronous data flow channels, providing an interface to the outside world in the form of an API gateway.

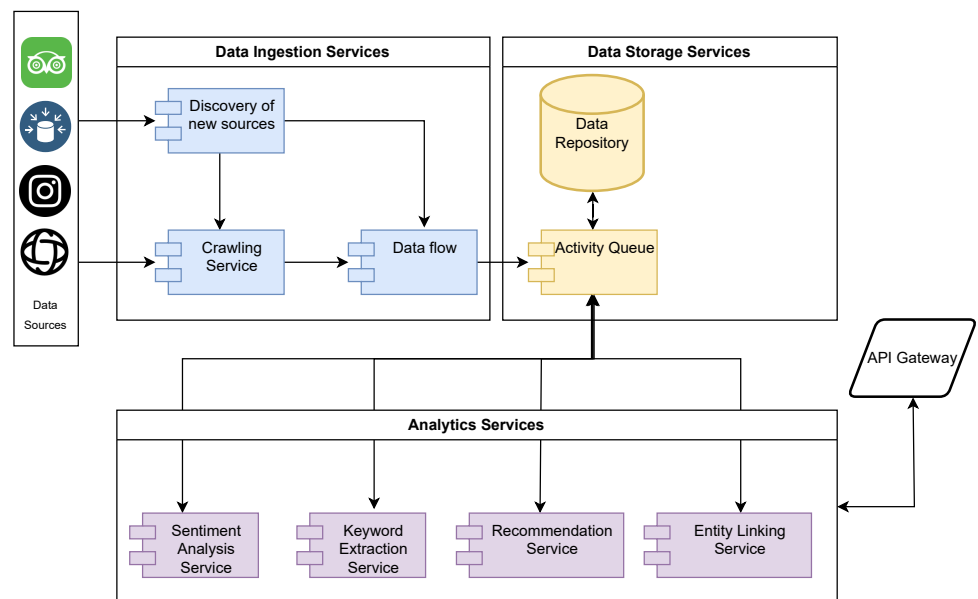


Figure 1. StreetLines platform architecture.

The processing queues are implemented as groups of interconnected services, data sources and repositories. There exist three groups of services: (i) data ingestion services (in light-blue color) (ii) data storage services (in light-yellow color) and (iii) analytics services (in light-purple color). As explained in the previous section, data sources include the worldwide web, online social media, forums, blogs, etc.

Data ingestion services include the crawling service (described in detail in Section 4.1), the discovery of new sources service and finally the data flow service. There are two ways for new data sources to be included in the crawling procedure: either manually or having been discovered after preprocessing the crawled data. In the latter case, possible sources include hyperlinks, mentions in social media posts, etc. Candidate sources are stored in a separate location in the data repository, to be later evaluated for their significance by human experts. Both the crawling and the discovery of new source services feed data to the data flow service, which ensures seamless communication with the data storage services and more specifically the activity queue service that, in this case, stores the crawled data to the data repository in a structured manner.

The activity queue service plays a central role in the proposed architecture, as it is called by the analytics services, when the relevant data analysis is requested by the users of the platform via API calls. These include the sentiment analysis service (Section 4.3), the keyword extraction service (Section 4.2), the entity linking service (Section 4.4) and finally the recommendation service (Section 4.5). It should once again be noted that the architecture of Figure 1 is highly modular and easily extendible; indeed, plugging in a new analytics service requires minimal effort as, apart from the API gateways it should implement, it would only need to communicate with the activity queue service to read from the data repository.

3.2. Implementation Details

All of the services presented in the previous subsection were implemented using the Python programming language and the relevant modules for each service. In this subsection, however, the development of the API Gateway is discussed in more detail, as well as the web-based GUI, where all information is aggregated and visualized, in order to gain various insights. In both cases, two popular and suitable Python frameworks have been employed: FastAPI [21] and Streamlit [22].

FastAPI [21] is a high-performance Python framework for creating APIs. It stands out for its remarkable speed, thanks to its asynchronous capabilities and efficient design. Furthermore, its straightforward design provides a flexible interface for developers. FastAPI utilizes Python data types for automatic data validation and generates interactive documentation based on OpenAPI standards. Supporting asynchronous programming and WebSockets, FastAPI offers flexibility for a variety of web application needs and has become the most popular framework for creating APIs and web services.

Streamlit [22], on the other hand, is an open-source Python library that simplifies the process of building web applications and interactive charts. It has been designed with a focus on simplicity and user-friendliness and is primarily used for data visualization and the rapid generation of web applications. Interactive applications created with Streamlit are based on Python scripts that include graphical elements like bars, buttons and text input fields. These graphical components enable users to interact with data and view real-time results. Streamlit applications can be employed for a wide range of purposes, including data visualization, machine learning model development and prototyping.

The combination of FastAPI and Streamlit in the platform provides an efficient and user-friendly environment for visualizing data and interacting with the available information. Figure 2 displays the visualizations of the Sentiment Analysis module, as they appear on the StreetLines platform. In the data repository (Figure 1), we have already inserted the POIs for the region of interest (in this example, Attica, Greece) and therefore we are able to select them from the drop-down menu ('Acropolis' has been selected in the example of Figure 2). This results in the relevant service being executed asynchronously (Sentiment

Analysis, see Figure 1) for the POI in question; data are requested from the Activity Queue and are subsequently provided as input to the trained model behind the Sentiment Service. Model predictions are then transferred to the API Gateway through FastAPI calls, where Streamlit is used to produce the chart in Figure 2.

Street Lines: Knowledge and Sentiment Extraction Tool

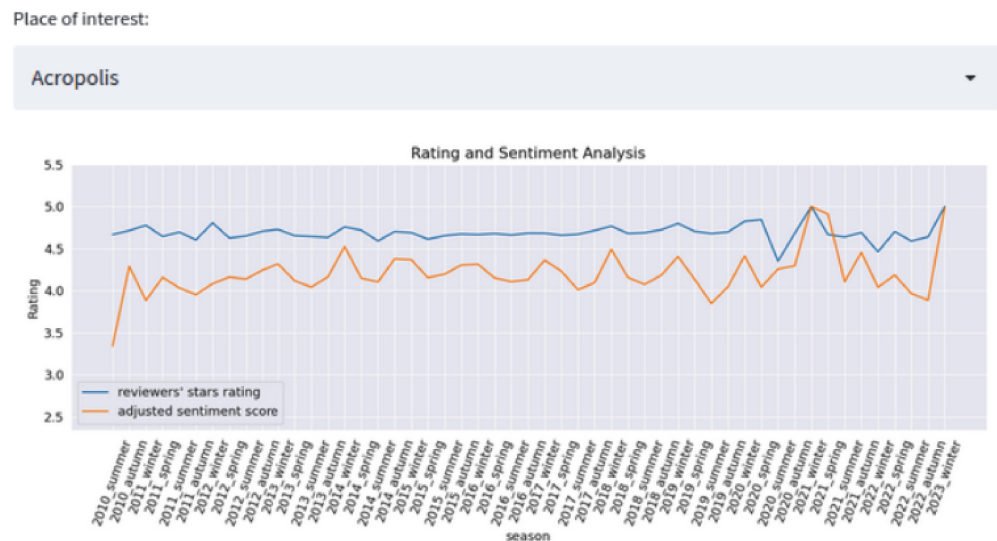



Figure 2. The sentiment analysis module of the StreetLines platform.

4. Components

In this section, the various components of the proposed platform are presented; starting with the crawling process (Section 4.1), we proceed with the keyword extraction methodology (Section 4.2), followed by the Sentiment Analysis procedure (Section 4.3). Then, Entity Linking (Section 4.4) is presented, along with the Recommendation module (Section 4.5).

4.1. Data Sources and Crawling

Two of the main services that store reviews of visitors and tourists in various locations around the world are TripAdvisor [23] and Google Maps [24]. TripAdvisor started as a travel website in 2000 and has now evolved into the largest travel social network, operating in 40 countries, available in 20 languages, and containing 1 billion reviews for approximately 8 million locations worldwide. In addition to the website, it has applications for both popular mobile phone platforms (Android and iOS), where the user can register either with a Google account or with their email. Registered users leave personalized reviews on points of touristic and cultural interest, both in the form of stars and in free text. Other users can interact with them by liking them, following the user or sending them a direct message. Also, certified POI owners (hotels, restaurants, etc.) can respond to user reviews. However, a drawback is that the reviews are not verified in practice; that is, it is not certain that the user who leaves a review for a place has actually visited it. Figure 3, below, contains excerpts from reviews of a restaurant in Athens, Greece, as seen on the TripAdvisor website.



Sandra Z
154 reviews

★★★★★ Reviewed 2 days ago via mobile

Good place to eat

We visited here based on previous good reviews and were not disappointed. The food was tasty and nicely served. We particularly enjoyed the wine selection.


Date of visit: November 2023

Helpful? 👍

This review is the subjective opinion of a Tripadvisor member and not of Tripadvisor LLC. Tripadvisor performs checks on reviews as part of our industry-leading trust & safety standards. Read our [transparency report](#) to learn more.

IOANNIS ILIOPOULOS, Ιδιοκτήτης at Elaëa Dine & Wine, responded to this review
Responded 2 days ago

We make great efforts to provide top quality cuisine, wines and services to all our guests . Your satisfaction, recognition and review rewards us 😊❤️ Thank you for dedicating your time to post on TripAdvisor !!



Barbara S
1 review

★★★★★ Reviewed 3 days ago via mobile

Do not miss when visiting Athens

The best food in Athens, best waiter Spiros. The food is excellent. The fish is amazing and the tastes are so well combined. Totally worth the price. Very friendly stuff. Also we got a complimentary desert.

Date of visit: November 2023

Helpful? 👍

This review is the subjective opinion of a Tripadvisor member and not of Tripadvisor LLC. Tripadvisor performs checks on reviews as part of our industry-leading trust & safety standards. Read our [transparency report](#) to learn more.

IOANNIS ILIOPOULOS, Ιδιοκτήτης at Elaëa Dine & Wine, responded to this review
Responded 3 days ago

All the Elaëa crew go out of their way to provide guests with honest hospitality and a special epicurean experience !! Thank you for the recognition and terrific review 😊 Your satisfaction rewards all our efforts ❤️

Figure 3. Review excerpts on TripAdvisor for a restaurant in Athens, Greece.

Google Maps [24], a mapping and navigation application first introduced in 2004 by Google Inc., has a built-in Places Service, which displays various POIs on a map, including those related to culture and tourism. As part of Google Services available for desktop and mobile devices, Google Maps are used by a very large number of people worldwide, exceeding 1 billion. Similar to TripAdvisor, registered users leave personalized reviews both in the form of stars and in free text. Certified business owners can respond to the reviews, while other users can have basic interactions through a like button, as well as by following the user in question, forming a primitive social network. In this case, as well, the main drawback is that reviews are not verified in practice. Figure 4, below, depicts an excerpt of reviews as they appear on Google Maps for the same restaurant in Athens, Greece, as in Figure 3. Obviously, the reviews in both cases have been made by different sets of users, which, however, may partly coincide.

Both of the aforementioned services provide APIs, which can return the reviews for a given POI. However, they limit the results to either the five most recent (TripAdvisor [25], Places API [26]) or the five most “relevant”, as determined by the service itself [26]. For this reason and in order to be able to collect a sufficient amount of data for the analysis described in the following sections, we chose to extract the reviews directly from the websites of the two services, applying web scraping techniques [27].

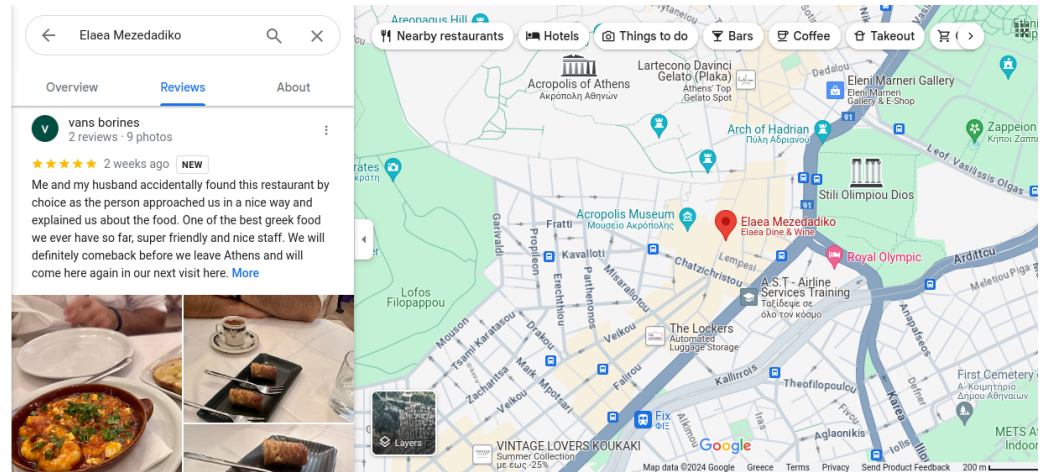


Figure 4. Review excerpts on Google Maps for the same restaurant in Athens, Greece, as in Figure 3.

Web scraping is a technique for extracting data from the Web and storing it in a file system or database for later retrieval or analysis [27]. Typically, web data are collected using by software known as web crawlers. Due to the fact that a huge amount of heterogeneous data are continuously generated on the Web, web scraping is widely recognized as an effective and powerful technique for collecting large volumes of data [28]. To adapt to a variety of scenarios, current web data harvesting techniques have been adapted from smaller ad hoc, assisted processes to the use of fully automated systems that can turn entire websites into a well-organized dataset. Modern web harvesting tools are not only capable of parsing markup languages or JSON files, but can also incorporate elements of visual analysis, as well as natural language processing techniques, to simulate how people browse web content [29].

The process of harvesting data from the Web can be divided into two steps: (i) acquiring Web resources and (ii) extracting the desired information from the acquired data. Specifically, a harvester begins by making an HTTP request to obtain resources from the target website. This request can be formatted as either a URL containing a GET query or an HTTP message body containing a POST query. Once the request has been successfully received and processed by the target website, the requested resource is retrieved and then returned to the harvester. The returned resource can be of many types, such as web pages in HTML format, data streams, XML or JSON files, or multimedia data such as images, audio and video. After receiving the web data, they are analyzed, reformatted and organized in a structured manner.

There exist two types of harvesters, depending on how they post-process the transferred resources: (i) those that treat web page content as pure text and multimedia and (ii) those that try to recreate the Document Object Model (DOM) of the web page [30], to varying degrees. The former type includes programs (e.g., wget [31] and curl [32]) as well as programming libraries (e.g., urllib [33] and requests [34] for the Python programming language), while the latter type also involves automation tools such as Selenium [35]), apart from programming libraries (e.g., BeautifulSoup [36]). In general, harvesters of the second type are more complex to handle and program, but they are also the most suitable for the objectives of the current work, as the examined webpages are not static and their DOM is recreated by the browser asynchronously with the use of scripting languages such as JavaScript and data transfers based on AJAX calls. In this respect, the harvester developed in the context of the proposed architecture is based on the Python APIs provided by the Selenium automation tool [35].

The aforementioned tool acts as a browser, in the sense that user interaction with a website is simulated programmatically. Therefore, in order to fully function, it requires the existence of a browser (e.g., Google Chrome or Mozilla Firefox), which is operated

in the so-called “marionette” mode [37], where user actions (scrolling, clicking, etc.) are performed by the harvester itself.

4.2. Keyword Extraction

Keyword extraction is performed on the free-text reviews that tourists make in online platforms, like those discussed in Section 4.1. The objective is to extract the most significant words and phrases out of the reviews for each POI, in order to quantify visitor opinion. To this end, transformer-based models [38] are employed; more specifically, the DistilRoBERTa model [16].

Transformers [38] constitute a category of neural networks developed for natural language processing tasks. Their architecture allows models to comprehend the full semantics of words in a text and understand the relationships between them. This makes them particularly suitable for tasks such as sentiment analysis, language translation, text generation and other applications that demand advanced text comprehension. BERT is trained on large text corpora and possesses the capability to “comprehend” the complete semantics and context of words. This means it can predict the next word that is likely to follow in a sentence, taking into account the entire context. This ability makes it suitable for various applications, such as sentiment analysis and keyword extraction.

Prior to model training, data pre-processing is performed in order to boost performance. Initially, all stop words (articles, punctuation marks, etc.) are removed from the reviews, as they do not provide substantial information to the model and complicate the process. Subsequently, reviews are sorted chronologically and grouped by year and season, in an effort to evaluate the evolution of visitor opinion over time. Each review is then divided into tokens: word or small phrases (consisting of up to four words) using n -grams of corresponding lengths.

The aforementioned tokens form the model input, which produces the respective embeddings, essentially converting review text into numerical vectors. Following this, an embedding is obtained out of the whole review text and is subsequently compared to the embeddings of the individual keywords via an appropriate metric (i.e., cosine similarity). Finally, the words/phrases whose embeddings exhibit the highest similarity score to the embedding of the review are returned as the most relevant keywords.

In the framework of the current project, tokens were generated for n -grams of lengths from 1 to 4, meaning that the candidate key phrases considered had a minimum length of 1 (single words) and a maximum of 4 (four-word phrases). The chosen DistilRoBERTa consists of six blocks, and each attention layer includes twelve heads. The obtained embeddings are vectors of size 768. Figure 5 displays keywords extracted from visitor reviews of the Acropolis of Athens, Greece, retrieved from TripAdvisor.



Figure 5. Keywords extracted from visitor reviews of the Acropolis of Athens, Greece.

4.3. Sentiment Analysis

DistilRoBERTa, outlined in Section 4.2, has also been employed for the sentiment analysis task. In this setting, however, the model has been pretrained on five different text datasets (BookCorpus [39], Wikipedia, CC-News [40], OpenWebText [41], Stories [42]) so as to learn fundamental correlations between words and sentences that may appear in text. Following that, every review undergoes a preprocessing stage, where it is tokenized in words or sub-words, with each token being represented as a numeric vector (embedding). The aforementioned embeddings are subsequently fed into the model that predicts the sentiment polarity of the review (positive/negative), along with a confidence score.

In this specific application, the model is comprised of 6 transformer blocks, each of which incorporates a multi-head attention layer with 12 heads. This design enables the model to capture a wide array of dependencies and intricate relationships within the text, enhancing its understanding of the context. In terms of the embeddings, they possess a dimensionality of 768, implying that each token is encoded with a vector of 768 values before being fed into the model. This rich embedding scheme empowers the model to process and interpret the textual content effectively.

The process begins with the separation of reviews into phrases and the assignment of emotional meaning to each of them. By utilizing information from the pre-trained model, we can understand whether the reviews express positive or negative sentiments, along with the intensity of each emotion, as it is represented by the confidence score. This procedure helps in understanding how tourists perceive various destinations, what advantages and disadvantages are mentioned, and how these emotions can influence travelers' choices. Furthermore, it is also possible to monitor the evolution of emotions via the analysis of the relevant patterns and trends that may emerge, depending on seasons or events. This perspective assists in mitigating negative aspects and enhancing travelers' experiences.

The primary objective is to uncover overarching patterns in the expressions used and to discern the sentiment prevalent in each comment. The analytical approach is twofold, delving into both long-term trends and seasonal dynamics. In the former perspective, the aim is to identify persistent linguistic patterns that emerge across diverse sets of reviews, shedding light on the recurring themes and sentiments. Simultaneously, we seek to pinpoint specific periods when reviews tend to turn negative, thereby exploring the correlations between sentiment and the time of the year. The comprehensive analysis results in valuable insights with respect to the expressed opinions, enhancing our understanding of the factors influencing visitor experiences over time.

To accomplish this task, the reviews are categorized according to the season they pertain to, and their sentiment scores, which fall within the $[-1, 1]$ range, are extracted for the corresponding seasons. To assess the reliability of the generated metric, these sentiment scores are linearly mapped to the $[0, 5]$ range, facilitating a direct comparison with user ratings (stars) during the relevant time frames, as depicted in Figure 6 for the Acropolis of Athens, Greece. A more detailed analysis has been published in a previous work by some of the authors [43].

4.4. Entity Linking

User reviews often contain references to other POIs, or more generally, named entities, as for example depicted in the review of Figure 7, where the visitor references in his/her review another POI (Syntagma Square), a property of the visited POI (a museum containing collections of Cycladic Art), and an event (the recent COVID-19 pandemic). It is obvious that being able to identify relevant entities in text is of pivotal importance for the objectives of the proposed platform, as it has the potential for uncovering hidden patterns in the factors that affect visitor preferences to provide valuable insight into their attitudes.

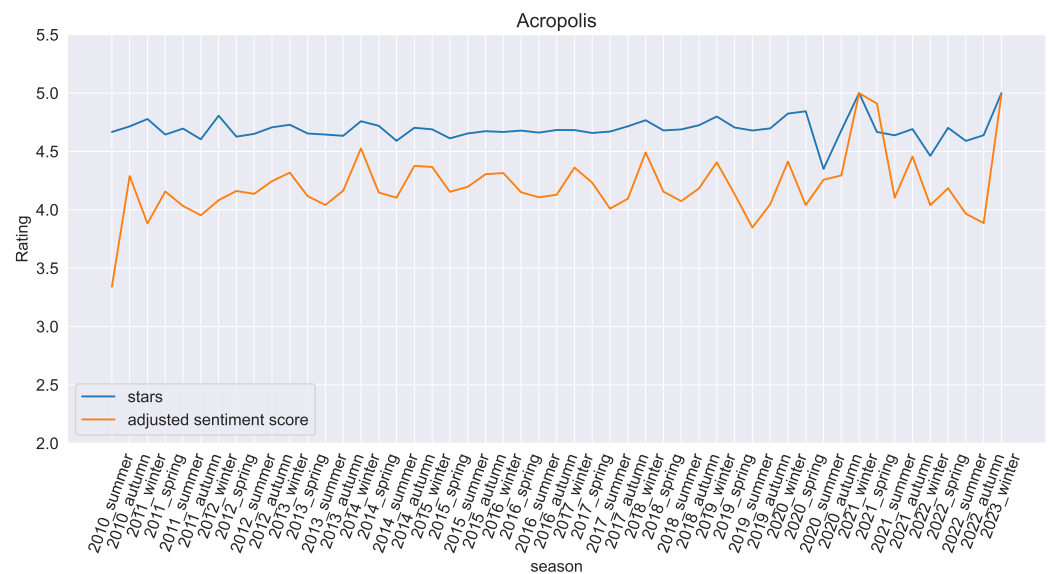


Figure 6. Sentiment analysis per season for the Acropolis of Athens, Greece.

The museum is within walking distance of *Syntagma Sq.* It is an intimate private museum featuring *Cycladic artifacts* and sculptures. There were 4 floors with a one way system in place due to *Covid* restrictions...

Figure 7. User review excerpt for the Museum of Cycladic Art in Athens, Greece (posted on TripAdvisor).

In the NLP realm, the procedure described above is known as Named Entity Recognition (NER); this is the process of searching for and locating entity references in texts, as well as assigning to them a label characterizing the type of entity, usually a unique Uniform Resource Identifier (URI) from a Knowledge Base (KB). This technique is extensively used in an effort to locate patterns in textual data, under the constraint of acceptable computational complexity [44].

One of the main issues of NER is that of ambiguity, where a candidate entity may correspond to more than one category, or otherwise be identified by more than one URI. In the example of Figure 7, “Syntagma Sq.” refers both to the square itself and the metro station underneath. “Covid” may refer to the virus itself or the ensuing pandemic. This problem is formally known as Entity Disambiguation (ED); that is, the process of matching the nominal entities of unstructured text to entities in the NB. In contrast to NER, ED identifies the entities in the knowledge base and, by extension, the meaning to which the named references correspond [45].

Regarding the issue of ER, the existing detection and disambiguation tools do not offer satisfactory solutions with easy parametrization, which is considered particularly important for the design and implementation of more specialized applications [46,47]. In this respect, automatically recognizing and supporting a new class of entities or determining how to link entities to KBs is a particularly difficult task. A possible solution can be the use of the Semantic Web as a KN, which has been explored in the literature in recent years, as a dynamically parametrized technique [48–51]. In particular, semantically organized information from available KBs on the Internet (e.g., DBpedia) can be used to more precisely clarify the discovered entities.

In this work, ED is treated in the framework of Entity Linking (EL), a common approach followed in the literature [52–55]. In EL, all possible candidate entities are identified and matched to at least one named entity in the respective KB. This is achieved by determining the semantic affinity of all “reference in text”–“candidate entity” pairs, taking into account both the coherence between candidate entities in the same text passage and the degree of similarity between the contexts of a reference and the text with which a candidate entity is described.

More specifically, we follow the approach described in [52]. Initially, review words are converted into a suitable vector representation (embeddings) that preserves their semantic meaning. Following that, they are provided as inputs to an entity connective model, whose integral component are Bi-directional Long Short-Term Memory (BiLSTM) networks. The aforementioned networks are considered to be within the state-of-the-art methods for capturing the spatial relationship between words and the order in which they appear in a text sequence [56].

For each reference entity, its embedding is expanded with neighboring words in the text, so that the semantic context in which it occurs is taken into account during clarification, thereby obtaining context-aware embeddings. Then, for each candidate named entity reference in the text under consideration, the most likely candidate entities are selected from KBs such as Wikipedia, Crosswikis and Yago, based on probabilistic algorithms [55]. For each candidate entity, a similarity metric (similarity score) is calculated, while the final vector of candidate entities for each reference in the text is used for the processes of both locating and clarifying entities.

The embeddings are subsequently presented to a fully connected, feed-forward neural network that consists of three layers, (i) one responsible for the final vector representation of the candidate nominal entities detected in the text, (ii) another one for computing the similarity between candidate entities and the final entity interconnection and (iii) the final one for their disambiguation. Finally, the semantic context of all occurrences of the same candidate named entity is taken into account for ED; that is, not only in the specific location in which the candidate entity has been located, but throughout the text. The disambiguation process is based on the most frequently mentioned entities in text, whose similarity is calculated against all the remaining entities left to disambiguate.

Figure 8 depicts the disambiguated text of Figure 7, where the text in brackets designates the semantically annotated entities located after the application of the ED process described in this section. In this example, each annotation serves as an identifier to the DB-Pedia KB (e.g., [Syntagma_Square] points to the https://dbpedia.org/page/Syntagma_Square entry of the KB, accessed on 29 February 2024). In this specific example, we can see that the ED model correctly identified the reference to the Syntagma Square of Athens, Greece, as well as the nature of the artifacts in the museum (Cycladic art). However, it failed to distinguish between the COVID-19 pandemic mentioned in the review (identifier [COVID-19_pandemic]) and the COVID-19 virus (identifier [COVID-19]), thereby exhibiting certain limitations of the approach that need to be further resolved by a human annotator.

The museum is within walking distance of [Syntagma_Square]. It is an intimate private museum featuring [Cycladic_art] and sculptures. There were 4 floors with a one way system in place due to [COVID-19] restrictions...

Figure 8. Semantically annotated user review excerpt for the Museum of Cycladic Art in Athens, Greece (posted on TripAdvisor).

4.5. Recommendation

Recommender Systems (RS) are software tools used to propose new, unseen items to users, according to their needs and preferences [57]. Items can be of various kinds: physical (e.g., products, books) or intangible (music, films, etc.). In the context of the current work, objects are exclusively POIs, while the users are the visitors to an area. In this respect, the objective of the RS is to suggest new and interesting POIs to visitors; that is, locations that the visitor has not visited before (“new”), which are expected to be a match for his/her tastes (“interests”).

RSs try to model visitor taste based on his/her past interactions, and for this reason machine learning methodologies are extensively used in this realm. ML methodologies typically model visitor taste by creating a “profile” for him/her, based on his/her “history” (set of POIs already visited). Then, on the basis of the constructed profile, the algorithm produces new recommendations. The type of feedback provided by the user can either be

explicit, such as the rating of a POI on a 5-star or a binary scale of like/dislike, or implicit, where no ratings or reviews are involved, but the presence of a visitor to a POI is designated to be an indication of interest.

In the context of the current work, Bayesian Personalized Ranking (BPR) [58] has been chosen because it supports both of the aforementioned feedback types and additionally it has been successfully realized several times in the past in culture- and tourism-related recommendations. BPR produces a ranked list of recommended items, where each item’s score is determined by the maximization of the probability (likelihood maximization) of observing the specific training data.

In BPR, each user and item is represented by a vector in a latent feature space, which encodes user preferences and item characteristics, respectively. Vector values are learned from the history of each user, with the objective function being optimized on both positive and negative feedback pairs. The former are user-item pairs encountered in the training dataset, while the latter are not. Therefore, the training objective is to learn to rank positive pairs higher than negative ones.

Like most supervised learning algorithms, BPR uses stochastic gradient descent to optimize the user and item latent representations. Since, in most practical RSs, negative pairs far outnumber positive ones, BPR samples negative pairs during the training process. Additionally, in order to avoid complex representations, a regularization term is also added to the objective function. Once training is complete, the model can predict user preferences on items he/she has not encountered yet.

In the example of Table 1, a personalized recommendation list produced by BPR for a sample user is displayed. The training data have been obtained for the city of Athens, Greece, according to the crawling procedure described in Section 4.1. POIs are returned in a ranked list, where the rank number is determined by the BPR score (3rd column of Table 1). Naturally, this architecture is extensible and may be provided as a service to the visitors of an area, should they choose to subscribe to a relevant service.

Table 1. Set of recommended POIs in Athens, Greece, for a sample user.

Result	POI	BPR Score
1	Benaki Museum	0.1390
2	Church of Kapnikarea	0.1107
3	Arch of Hadrian	0.0880
4	Attiko Metro	0.0765
5	Anafiotika	0.0761

5. Evaluation

This section discusses a preliminary evaluation of the StreetLines platform. The said evaluation can be performed at two levels: at the platform level (Section 5.1) and at the service level. In Section 5.2, an initial assessment of the Sentiment Analysis service is provided.

5.1. Platform

The StreetLines platform implements a comprehensive approach to extract valuable insights from users’ reviews by integrating two layers of services: Sentiment Analysis and Keyword Extraction at the first layer, and Entity Linking and Recommendation at the second. The culmination of these analyses is presented in the form of charts and plots within a user-friendly web application. This visualization approach enhances the interpretability of the results, empowering policymakers and interested individuals to quickly grasp overarching trends and sentiments.

Despite meeting the requirements to comprehensively analyze and present insights derived from users’ reviews on the internet, the current limitations should also be discussed. Firstly, as the platform is designed to crawl data from online services, it is sensitive to data format alternations or sudden policy changes of service owners (e.g., blocking crawlers).

Regarding keyword extraction, keywords consisting of up to 4 words can be extracted by the model, although this is an implementation choice that was adopted in order to preserve the balance between efficiency and extracted knowledge quality. Finally, in terms of sentiment analysis, the final score can be interpreted as negative, neutral or positive, indicating the overall users' satisfaction without explicitly focusing on specific aspects.

5.2. Sentiment Analysis Service

In order to be able to evaluate the performance of the model behind the sentiment analysis service (Section 4.3), we treat the task as a supervised learning problem. In this respect, there is a limitation arising from the lack of ground-truth labels from the extracted reviews. This means that the polarity and sentiment score of the reviews is not known beforehand. Therefore, it is not possible to directly compare the model's predictions with actual sentiment labels and derive system performance metrics. For this reason, we proceeded with two different evaluation protocols.

In the first case, an annotated dataset of reviews from TripAdvisor [59] was employed, on which standard supervised learning metrics had been computed. Since the crawled data were also tourism-related and came from similar platforms, the model's performance on the selected dataset was expected to be similar to its performance on the crawled data, thereby exhibiting its generalization capability.

Table 2 summarizes the precision, recall, F1-score and accuracy scores of the model on the TripAdvisor dataset [59]. With the exception of the recall metric on the negative class, all other metrics demonstrated efficient operation, which means that the proposed model is able to identify the actual sentiment of each review in the vast majority of cases. The model's performance on the negative sentiment class is somewhat below that of the positive one, and this behavior is attributed to the class imbalance between the two categories, something that should be taken into account in future model updates.

Table 2. Sentiment classification performance on the TripAdvisor dataset [59].

	Precision	Recall	F1-Score	Support
NEGATIVE	0.83	0.72	0.77	61
POSITIVE	0.90	0.95	0.92	169
Accuracy	-	-	0.89	230
Macro Avg	0.87	0.83	0.85	230
Weighted Avg	0.88	0.89	0.88	230

The second evaluation protocol is based on the use of the reviews' ratings as a measure of comparison with the sentiment score of the model. Specifically, the mean score of the ratings from the corresponding POI is used as the ground truth to evaluate the efficiency of the model per place. To achieve this, the sentiment confidence score (initially in the $[-1, 1]$ range) is linearly scaled to the $[0, 5]$ range, mapped to the range of values user ratings (stars) can take. Subsequently, the mean squared error (MSE) and mean absolute error (MAE) for each location were calculated separately, as well as their averages, in order to assess the model's effectiveness.

The average MAE and MSE scores are 0.5563 and 0.8390, respectively, indicating the model's capability of capturing the general trend of the reviews and depicting the overall satisfaction of the users. It also becomes evident that the model performs very satisfactorily, given the constraints and the level of difficulty of the specific problem, successfully identifying correlations between the review text and the extracted sentiment.

In order to delve into a more thorough evaluation, we performed a detailed analysis of the errors encountered during the application of sentiment analysis. Our investigation reveals the existence of two distinct and noteworthy types of errors, shedding light on the intricacies of the sentiment prediction process. The first type of error emerges when reviewers assign positive star ratings, indicating an overall favorable experience,

while the corresponding review texts yield a contrary result by expressing negative sentiments. This incongruity highlights the nuanced nature of sentiment expression within the textual content.

Conversely, the second type of error manifests when reviewers provide negative star ratings, suggesting a less favorable encounter, while the sentiment analysis predicts a positive sentiment based on the textual analysis. To visually articulate and emphasize these findings, the prevalence and distribution of these two types of errors are presented in Figure 9 for 7 touristic places from our dataset. This Figure serves as a visual aid in elucidating the implications of sentiment analysis errors in the context of touristic reviews. It is shown that there is a tendency among reviewers to overscore the experience of places, while the written review text usually focuses on things they disliked. On the other hand, the figure highlights that the percentages of underscored reviews are close to zero, meaning it is quite uncommon to negatively evaluate a place (in terms of star rating) while describing a positive experience in the review text.

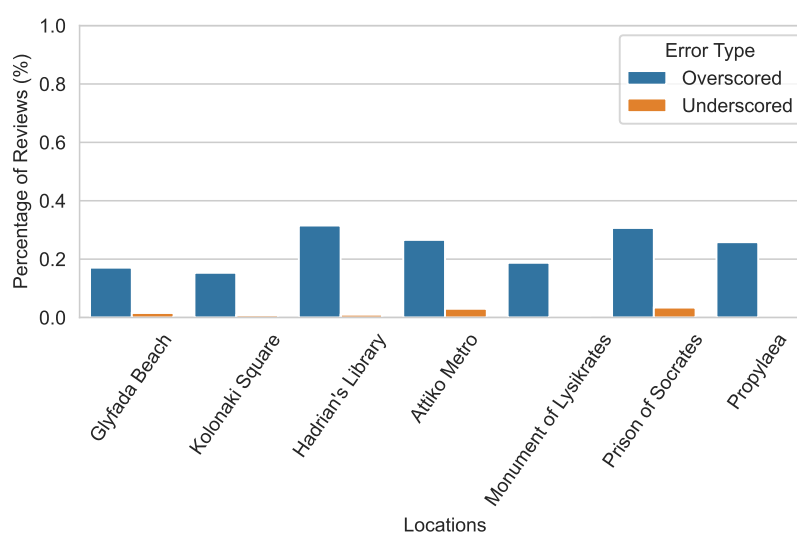


Figure 9. Percentage of overscored and underscored reviews.

6. Discussion

In this section, we provide an overview of the proposed platform (Section 6.1), summarizing the findings of the current work, analyzing the novelty of the platform and discussing its theoretical and practical implications. Additionally, we considered the limitations of tourism mining in general and of the proposed solution in particular (Section 6.2), along with some possible extensions (Section 6.3) that can further expand the functionality of the StreetLines platform.

6.1. Overview

The StreetLines platform is a highly modular and extensible data mining platform for tourism-related tasks. Data sources may be easily added or removed and analytics services may be carried out. It can also ingest data from a multitude of online sources and it can be used for an arbitrary number of regions, cities and areas. The StreetLines platform addresses issues related to data heterogeneity, which are common for all related attempts.

The novelty of the proposed platform also extends to its constituent components, and more specifically the Keyword Extraction and Sentiment Analysis services. By employing a DistilRoBERTa, a transformer-based large language model equipped with attention mechanisms, more complex relationships between words are modeled, which permit enhanced keyword extraction and sentiment classification. Additionally, the distilled nature of the model guarantees a reduced network size and faster inference, without compromising performance.

The main theoretical implications of the StreetLines platform revolve around market efficiency, because by analyzing data in real-time, the platform can help businesses in the tourism domain to improve service offerings and target market efforts, for example, more precisely. Additionally, insights from the platform can be used to support policies that promote balanced regional development by highlighting, for instance, under-visited destinations that have potential for growth, thus helping to distribute tourist traffic more evenly.

On the other hand, the main practical implications of the StreetLines platform have to do with bolstering innovation and contributing to technological evolution in the tourism sector. By providing a wealth of data, the platform can spur innovation in tourism-related services, such as personalized travel experiences, dynamic tour packages and immersive virtual tours. In addition, the platform may serve as a testbed for emerging technologies like augmented reality (AR) and virtual reality (VR), offering new ways to experience and market tourism.

6.2. Limitations

Despite its promising results, the introduced platform faces certain limitations that might hinder its effectiveness and reliability, the most prominent of which is related to its sole reliance on UGC (Section 3). This limitation is not only related to the fact that device-generated or transactional data can possibly extend existing analysis tools, but also has to do with the quality of the platform-generated insights, which might lack contextual depth, as quantitative data alone cannot fully capture the nuanced human experiences and cultural factors influencing tourism behaviors [60].

Moreover, online reviews and social media posts can be biased, fabricated, or influenced by individual subjective experiences, leading to skewed insights [61]. Another challenge is the multi-modality of the crawled data, which vary in format, structure and language. This diversity, and especially the latter case, require advanced NLP techniques, as automatic translation into a reference language (e.g., English) is not necessarily the best choice [62]. The aforementioned limitations underscore the need for ongoing research into technology, data governance and methodological approaches to enhance the effectiveness of tourism mining platforms.

6.3. Extensions

In this subsection, we discuss some possible extensions of the proposed platform with regard to context and personalization (Section 6.3.1), as well as Recommendation (Section 6.3.2), that can not only improve its performance, but also its capacity and broader added-value.

6.3.1. Context and Personalization

A possible extension of the proposed platform is to add context-awareness and personalization capabilities. The former is the ability of the system to adapt based on various parameters of its operational environment. A context-aware system is able to customize content delivery according to the specifics of each situation. In the examined case, context mainly refers to visitor profiles, which can be explicitly or implicitly collected. Registered accounts may store their preferences, profile and history of actions, while guest accounts build a temporary profile based on previous actions of the current session.

External contextual data may also be exploited, such as environmental parameters (temperature, humidity, etc.) or real time data retrieved from POIs (such as crowd congestion for popular attractions). The fusion of profiled and external contextual data streams can direct the customization process to adapt not only to visitor preferences, but also to the current availability and suitability of attractions. Finally, time and space can also be taken into account, recommending suitable attractions for the time of the day and based on the location and proximity to them.

Adding personalization aspects to a context-aware system results in improvements to the quality of the produced recommendations, and it also aids in the data analysis of

touristic data collected from online social media. Taking context into account, sentiment analysis could also be optimized by adding more weight to those opinions that are relevant to the user. Finally, the extended platform could exploit real-time contextual data for the detection of immediate actions required by the persons in charge of an attraction (e.g., by identifying mass complaints on an emergency issue).

6.3.2. Recommendation

One of the components of the platform that can be modularly improved is the recommendation engine. Various mechanisms can be employed and tested with the currently demonstrated solution, comparing their performance in terms of accuracy, speed and scalability metrics. There are multiple recommendation approaches that could be a potential fit in our case as well, but here we envision a more tailored one, based on complex network analysis. Our previous work [63] proposed a path-based recommendation approach, where users and items are represented in a two-layer network graph, and recommendations for one particular user are determined by solving a shortest path problem between the user and the items he/she has not encountered yet. Such items can be selected from a user's peers/friends. This problem is solved rather efficiently by performing network embedding over a two-dimensional hyperbolic metric space, and then computing simple hyperbolic distances (with complexity $O(1)$) over the embedded nodes. This approach allows for tremendous search speeds between user-item pairs, even though it bears an initial computational cost for the embedding. Furthermore, it was shown that this approach is a rather suitable one for networks exhibiting hierarchical structure, e.g., those of scale-free structure. It is noted that such two-layer graphs associating users (in terms of friendships or memberships in online social networks) and the items available in such platforms bear such a hierarchical structure, which they inherit due to the scale-free nature developing in affiliation and human-activity-related networks.

In our case, the availability of knowledge on involved tourists and relevant services/events of interest, which can potentially be represented in the form of a two-layer knowledge graph, makes it suitable when employing the hyperbolic network embedding approach for recommendation decisions. The knowledge collected via our platform concerns multiple types of information, which, however, can be efficiently represented by hierarchical multi-layer graphs seamlessly, also incorporating information on the interaction of the piece of available information collected through various means. The approach [63] can be transparently applied over such network graphs, and reveal hidden, emerging patterns which in turn will lead to more targeted recommendations. In addition, the proposed approach ensures efficient scaling, since network embedding has been shown to scale up much better than Euclidean network embedding or other recommendation approaches. Intuitively, this can be verified by the fact that path distances are computed as coordinate distances and not as shortest paths on graphs (hop-distances). In addition, the hyperbolic space, bearing a fractal-like structure, allows for the packing of more points (corresponding to network nodes or information) over the same space compared to Euclidean spaces. For instance, 2-dimensional embedding in the hyperbolic space might be sufficient for the developed knowledge graph, when 8-dimensional or even higher Euclidean space would be required.

7. Conclusions

The StreetLines platform, an innovative tool for touristic data analysis and insights, has been introduced in this paper. The proposed platform exploits machine learning techniques in order to extract meaningful information about data collected from popular online social media (such as Tripadvisor and Google Places), as well as other online sources. More specifically, automatic data collection is followed by data analysis which extracts keywords, enriches the entities semantically and performs sentiment analysis on the review text. The tool also supports recommendations to the users on touristic attractions. In this respect, the presented platform can significantly contribute to future studies in the field by

providing a rich, empirical foundation for research. Its modular architecture can be easily extended to incorporate even more data sources and analytics services, while an update to existing modules does not affect its overall functionality. Therefore, the StreetLines platform outlines the feasibility of a tourism-mining architecture that consolidates heterogeneous data sources, as well as its applicability to any city, region or country.

Overall, the platform can support the vivid touristic sector by providing an automated analysis of touristic data, which are generated in huge volumes on the Web. The initial evaluation of the StreetLines platform exhibited good performance and precision of the underlying technologies, allowing for accurate information delivery. Thus, interested organizations and institutions that support or provide touristic attractions can identify emerging product trends, consumer interests and concerns, while also allowing for personalized recommendations. Nevertheless, the platform needs to be further evaluated, and also regarding its other components. Finally, two extensions have been identified, one exploiting hyperbolic network embedding for more efficient and targeted recommendations and a second on contextual personalization, which could provide a basis for further enhancing the platform and providing added value.

Author Contributions: Conceptualization, G.A. and V.K.; methodology, K.M.; software, G.A., T.P. and G.I.; validation, G.A., T.P. and G.I.; formal analysis, V.K.; investigation, K.M. and G.S.; resources, G.C. and S.P.; data curation, G.A.; writing—original draft preparation, K.M. and V.K.; writing—review and editing, G.S.; visualization, G.A., T.P. and G.I.; supervision, G.C.; project administration, S.P.; funding acquisition, G.C. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the “Research and Innovation Synergies in the Region of Attica”, realized within the framework of ESPA 2014-2020, co-financed by Greece and the European Union (European Regional Development Fund) under the title “Smart Tourism Recommendations based on Efficient Knowledge Mining on Online Platforms” (operation code: MIS 5185025).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors due to copyright issues.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
AR	Augmented Reality
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory networks
CNN	Convolutional Neural Network
DOM	Document Object Model
ED	Entity Disambiguation
EL	Entity Linking
ER	Entity Recognition
FFNN	Feed Forward Neural Network
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
KB	Knowledge Base
LOD	Linked Open Data
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MSE	Mean Squared Error

NER	Named Entity Recognition
NLP	Natural Language Processing
POI	Point of Interest
REST	Representational State Transfer
RS	Recommender System
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
UGC	User-Generated Content
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VR	Virtual Reality
XML	Extensible Markup Language

References

- Buhalis, D.; Amaranggana, A. Smart Tourism Destinations Enhancing Tourism Experience Through Personalisation of Services. In Proceedings of the Information and Communication Technologies in Tourism, Lugano, Switzerland, 3–6 February 2015; Tussyadiah, I., Inversini, A., Eds.; Springer: Cham, Switzerland, 2015; pp. 377–389.
- Lu, W.; Stepchenkova, S. User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software. *J. Hosp. Mark. Manag.* **2015**, *24*, 119–154. [\[CrossRef\]](#)
- Iorio, C.; Pandolfo, G.; D’Ambrosio, A.; Siciliano, R. Mining big data in tourism. *Qual. Quant.* **2020**, *54*, 1655–1669. [\[CrossRef\]](#)
- Han, S.; Anderson, C.K. Web Scraping for Hospitality Research: Overview, Opportunities, and Implications. *Cornell Hosp. Q.* **2021**, *62*, 89–104. [\[CrossRef\]](#)
- StreetLines. StreetLines Project. 2023. Available online: <https://streetlines.gr/> (accessed on 20 January 2024).
- Lyu, J.; Khan, A.; Bibi, S.; Chan, J.H.; Qi, X. Big data in action: An overview of big data studies in tourism and hospitality literature. *J. Hosp. Tour. Manag.* **2022**, *51*, 346–360. [\[CrossRef\]](#)
- Raj, S.; Kajla, T. Tourism analytics: Social media analytics framework for promoting Asian tourist destinations using big data approach. *J. Glob. Bus. Adv.* **2018**, *11*, 64–88. [\[CrossRef\]](#)
- Zhou, X.; Xu, C.; Kimmons, B. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Comput. Environ. Urban Syst.* **2015**, *54*, 144–153. [\[CrossRef\]](#)
- Bustamante, A.; Sebastia, L.; Onaindia, E. BITOUR: A Business Intelligence Platform for Tourism Analysis. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 671. [\[CrossRef\]](#)
- Marine-Roig, E.; Anton Clavé, S. Tourism analytics with massive user-generated content: A case study of Barcelona. *J. Destin. Mark. Manag.* **2015**, *4*, 162–172. [\[CrossRef\]](#)
- Wenan, T.; Shrestha, D.; Gaudel, B.; Rajkarnikar, N.; Jeong, S.R. Analysis and Evaluation of TripAdvisor Data: A Case of Pokhara, Nepal. In Proceedings of the Intelligent Computing & Optimization, Hua Hin, Thailand, 27–28 October 2022; Vasant, P., Zelinka, I., Weber, G.W., Eds.; Springer: Cham, Switzerland, 2022; pp. 738–750.
- Álvarez Carmona, M.A.; Aranda, R.; Rodríguez-Gonzalez, A.Y.; Fajardo-Delgado, D.; Sánchez, M.G.; Pérez-Espinosa, H.; Martínez-Miranda, J.; Guerrero-Rodríguez, R.; Bustio-Martínez, L.; Díaz-Pacheco, Á. Natural language processing applied to tourism research: A systematic review and future research directions. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 10125–10144. [\[CrossRef\]](#)
- Li, J.; Xu, L.; Tang, L.; Wang, S.; Li, L. Big data in tourism research: A literature review. *Tour. Manag.* **2018**, *68*, 301–323. [\[CrossRef\]](#)
- Le Huy, H.N.; Minh, H.H.; Van, T.N.; Van, H.N. Keyphrase extraction model: A new design and application on tourism information. *Informatica* **2021**, *45*. [\[CrossRef\]](#)
- Liu, Z.; Masui, F.; Ptaszynski, M. Supporting Inbound Tourism in Hokkaido: Keyword Extraction and Focus Point Analysis from Spot Reviews. In Proceedings of the 2021 International Workshop on Modern Science and Technology, Hangzhou, China, 16–18 July 2021; The International Center of National University Corporation Kitami Institute: Kitami, Japan 2021; Voluem 2021, pp. 151–156.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
- Qian, H.; Tang, Z.; Ren, Y.; Li, Q.; Zeng, D. A Transformer-based Approach for Identifying Target-oriented Opinions from Travel Reviews. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–7. [\[CrossRef\]](#)
- Mehraliyev, F.; Chan, I.C.C.; Kirilenko, A.P. Sentiment analysis in hospitality and tourism: A thematic and methodological review. *Int. J. Contemp. Hosp. Manag.* **2022**, *34*, 46–77. [\[CrossRef\]](#)
- Anis, S.; Saad, S.; Aref, M. Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cairo, Egypt, 19–21 October 2020; Hassanien, A.E., Slowik, A., Snašel, V., El-Deeb, H., Tolba, F.M., Eds.; Springer: Cham, Switzerland, 2021; pp. 227–234.
- Puh, K.; Bagić Babac, M. Predicting sentiment and rating of tourist reviews using machine learning. *J. Hosp. Tour. Insights* **2023**, *6*, 1188–1204. [\[CrossRef\]](#)

21. FastAPI. Available online: <https://fastapi.tiangolo.com/> (accessed on 15 June 2024).
22. Streamlit. Available online: <https://streamlit.io/> (accessed on 15 June 2024).
23. TripAdvisor. Over a Billion Reviews & Contributions for Hotels, Attractions, Restaurants, and More. Available online: <https://www.tripadvisor.com/> (accessed on 29 February 2024).
24. Google Maps. Available online: <https://maps.google.com/> (accessed on 29 February 2024).
25. Tripadvisor Content API. Available online: <https://tripadvisor-content-api.readme.io/reference/getlocationreviews> (accessed on 29 February 2024).
26. Places Details | Places API | Google for Developers. Available online: <https://developers.google.com/maps/documentation/places/web-service/details> (accessed on 29 February 2024).
27. Zhao, B. Web Scraping. In *Encyclopedia of Big Data*; Schintler, L.A., McNeely, C.L., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 1–3. [CrossRef]
28. Bar-Ilan, J. Data collection methods on the Web for infometric purposes—A review and analysis. *Scientometrics* **2001**, *50*, 7–32. [CrossRef]
29. Yi, J.; Nasukawa, T.; Bunescu, R.; Niblack, W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL, USA, 19–22 November 2003; pp. 427–434. [CrossRef]
30. W3C. *Document Object Model (DOM) Level 3 Core Specification*; Technical report; World Wide Web Consortium. 2004. Available online: <https://www.w3.org/TR/DOM-Level-3-Core/> (accessed on 29 February 2024).
31. Free Software Foundation. *GNU Wget*; GNU Project; Free Software Foundation: Boston, MA, USA, 2021.
32. Curl—A Tool to Transfer Data from or to a Server. 2024. Available online: <https://curl.se/> (accessed on 29 February 2024).
33. Urllib—URL Handling Modules. Available online: <https://docs.python.org/3/library/urllib.html> (accessed on 29 February 2024).
34. Reitz, K. Requests: HTTP for Humans. In *Requests Documentation*; 2013. Available online: <https://requests.readthedocs.io/en/latest/> (accessed on 29 February 2024).
35. The Selenium Browser Automation Project. Available online: <https://www.selenium.dev/> (accessed on 29 February 2024).
36. Richardson, L. Beautiful Soup. 2004. Available online: <https://www.crummy.com/software/BeautifulSoup/> (accessed on 29 February 2024).
37. Introduction to Marionette. Available online: <https://firefox-source-docs.mozilla.org/testing/marionette/Intro.html> (accessed on 29 February 2024).
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; 2017; pp. 5998–6008.
39. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In Proceedings of the The IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
40. Hamburg, F.; Meuschke, N.; Breiting, C.; Gipp, B. News-please: A Generic News Crawler and Extractor. In Proceedings of the 15th International Symposium of Information Science, Berlin, Germany, 13–15 March 2017; pp. 218–223. [CrossRef]
41. Gokaslan, A.; Cohen, V. OpenWebText Corpus. 2019. Available online: <http://Skylion007.github.io/OpenWebTextCorpus> (accessed on 29 February 2024).
42. Trinh, T.H.; Le, Q.V. A Simple Method for Commonsense Reasoning. *arXiv* **2019**, arXiv:1806.02847.
43. Papagiannis, T.; Ioannou, G.; Michalakis, K.; Alexandridis, G.; Caridakis, G. Analyzing User Reviews in the Tourism & Cultural Domain - The Case of the City of Athens, Greece. In Proceedings of the Artificial Intelligence Applications and Innovations. AIAI 2023 IFIP WG 12.5 International Workshops, León, Spain, 14–17 June 2023; Maglogiannis, I., Iliadis, L., Papaleonidas, A., Chochliouros, I., Eds.; Springer: Cham, Switzerland, 2023; pp. 284–293.
44. Cao, L.; Luo, C.; Zhang, C. Agent-mining interaction: An emerging area. In Proceedings of the International Workshop on Autonomous Intelligent Systems: Multi-Agents and Data Mining, St. Petersburg, Russia, 3–5 June 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 60–73.
45. Mohit, B. Named entity recognition. In *Natural Language Processing of Semitic Languages*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 221–245.
46. Derczynski, L.; Maynard, D.; Rizzo, G.; Van Erp, M.; Gorrell, G.; Troncy, R.; Petrak, J.; Bontcheva, K. Analysis of named entity recognition and linking for tweets. *Inf. Process. Manag.* **2015**, *51*, 32–49. [CrossRef]
47. Shen, W.; Wang, J.; Han, J. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 443–460. [CrossRef]
48. Guo, Z. Towards an Accurate, Robust, and Scalable Named Entity Disambiguation System. Ph.D. Thesis, University of Alberta, Edmonton, AB, Canada, 2018.
49. Fafalios, P.; Baritakis, M.; Tzitzikas, Y. Exploiting linked data for open and configurable named entity extraction. *Int. J. Artif. Intell. Tools* **2015**, *24*, 1540012. [CrossRef]

50. Ristoski, P.; Paulheim, H. Rdf2vec: Rdf graph embeddings for data mining. In Proceedings of the The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, 17–21 October 2016; Proceedings, Part I 15; Springer: Berlin/Heidelberg, Germany, 2016; pp. 498–514.
51. Frontini, F.; Brando, C.; Ganascia, J.G. Semantic web based named entity linking for digital humanities and heritage texts. In Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, Portorož, Slovenia, 1 June 2015.
52. Kolitsas, N.; Ganea, O.E.; Hofmann, T. End-to-end neural entity linking. *arXiv* **2018**, arXiv:1808.07699.
53. Ji, Y.; Tan, C.; Martschat, S.; Choi, Y.; Smith, N.A. Dynamic entity representations in neural language models. *arXiv* **2017**, arXiv:1708.00781.
54. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-end neural coreference resolution. *arXiv* **2017**, arXiv:1707.07045.
55. Ganea, O.E.; Hofmann, T. Deep joint entity disambiguation with local neural attention. *arXiv* **2017**, arXiv:1704.04920.
56. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
57. Resnick, P.; Varian, H.R. Recommender systems. *Commun. ACM* **1997**, *40*, 56–58. [[CrossRef](#)]
58. Rendle, S.; Freudenthaler, C.; Gantner, Z.; Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June 2009; pp. 452–461.
59. Wang, H.; Lu, Y.; Zhai, C. Latent aspect rating analysis on review text data: A rating regression approach. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 783–792.
60. Kusumasondjaja, S.; Shanka, T.; Marchegiani, C. Credibility of online reviews and initial trust: The roles of reviewer’s identity and review valence. *J. Vacat. Mark.* **2012**, *18*, 185–195. [[CrossRef](#)]
61. Choi, S.; Mattila, A.S.; Hoof, H.B.V.; Quadri-Felitti, D. The Role of Power and Incentives in Inducing Fake Reviews in the Tourism Industry. *J. Travel Res.* **2017**, *56*, 975–987. [[CrossRef](#)]
62. Mariani, M.M.; Borghi, M.; Okumus, F. Unravelling the effects of cultural differences in the online appraisal of hospitality and tourism services. *Int. J. Hosp. Manag.* **2020**, *90*, 102606. [[CrossRef](#)]
63. Papadis, N.; Stai, E.; Karyotis, V. A path-based recommendations approach for online systems via hyperbolic network embedding. In Proceedings of the 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 3–6 July 2017; pp. 973–980. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.