







## Article

# A Machine Learning Framework for Harvesting and Harmonizing Cultural and Touristic Data

Kimon Deligiannis \*, Christos Tryfonopoulos , Paraskevi Raftopoulou , Costas Vassilakis , Vassilis Kaffes  and Spiros Skiadopoulos 

Department of Informatics and Telecommunications, University of the Peloponnese, GR-22100 Tripoli, Greece; trifon@uop.gr (C.T.); praftop@uop.gr (P.R.); costas@uop.gr (C.V.); vkaff@uop.gr (V.K.); spiros@uop.gr (S.S.)

\* Correspondence: deligiannis@uop.gr

## Abstract

Cultural and touristic information is increasingly available through a multitude of heterogeneous sources, including official repositories, community platforms, and open data initiatives. While prominent landmarks are typically covered across sources, less-known attractions are also documented with varying degrees of detail, resulting in fragmented, overlapping, or complementary content. To enable integrated access to this wealth of information, harvesting and consolidation mechanisms are required to collect, reconcile, and unify distributed content referring to the same entities. This paper presents a machine learning-driven framework for harvesting, homogenizing, and augmenting cultural and touristic data across multilingual sources. Our approach addresses entity resolution, duplication detection, and content harmonization, laying the foundation for enriched, unified representations of attractions and points of interest. The framework is designed to support scalable integration pipelines and can be deployed in applications aimed at tourism promotion, digital heritage, and smart travel services.

**Keywords:** Machine Learning (ML); Named Entity Recognition (NER); web scraping; data homogenization; data augmentation; trajectory extraction; cultural heritage; tourism analytics; digital heritage; social media analysis



Academic Editors: Sokratis Katsikas and Vincenzo Moscato

Received: 27 September 2025

Revised: 7 November 2025

Accepted: 14 November 2025

Published: 28 November 2025

**Citation:** Deligiannis, K.; Tryfonopoulos, C.; Raftopoulou, P.; Vassilakis, C.; Kaffes, V.; Skiadopoulos, S. A Machine Learning Framework for Harvesting and Harmonizing Cultural and Touristic Data.

*Information* **2025**, *16*, 1038. <https://doi.org/10.3390/info16121038>

**Copyright:** © 2025 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

In the era of big data, cultural institutions and tourist organizations increasingly rely on large-scale data collections to support regular and strategic decision making, advancing beyond individual perceptions, instincts, or general observations [1–5]. A wide array of structured, semi-structured, and unstructured data is now accessible through various digital channels, including Application Programming Interfaces (APIs) [6], open data platforms [7], manually collected survey records [8], and digital libraries [9]. Among modern data acquisition practices, web scraping has emerged as a popular and effective method, enabling the automated extraction of information from specific web pages or broader sections of the public web [10]. Additionally, social media platforms offer rich sources of user-generated content, such as reviews and personal experiences, which can provide valuable insights across numerous sectors. In recent years, numerous solutions have leveraged social media data harvesting, often employing web scraping techniques, to extract, analyze, and repurpose this type of content for research and operational purposes [11–30].

Harvesting heterogeneous data on cultural points of interest (PoIs) from multiple sources often leads to overlapping or duplicated records referring to the same attraction.

These redundancies impose the need for data cleaning and harmonization, while complementary information from different sources must also be effectively integrated [31]. The challenge, commonly known as entity resolution, has become increasingly prominent in systems managing large volumes of diverse data. To provide comprehensive and reliable information, such systems must detect and reconcile duplicate entries, ensuring that data referring to the same PoI are consolidated into a unified representation. Although various methods have been proposed to address entity resolution [32–46], most existing approaches are tailored to specific homogenization tasks and often lack general applicability.

In search of useful insights and extraction of fundamental knowledge, various stakeholders in the cultural and tourism sectors (such as tourists, small and medium enterprises, transport providers, etc.) exploit Machine Learning (ML) algorithms and models to analyze the data they possess [47–49]. The quality of these data plays a crucial role, as cleaner and more structured datasets lead to more accurate and efficient ML outcomes [50]. ML approaches are also extensively applied in Cultural Heritage (CH) conservation, including the recognition of cultural patterns [51], the digitization and dissemination of historical documents [52], and the monitoring of valuable architectural structures and cultural sites [53]. Natural Language Processing (NLP) is a common technique applied to textual data that integrates computational linguistics, such as rule-based modeling of physical language, with statistical and ML models, allowing computing devices to recognize and understand human-generated text [54]. Named Entity Recognition (NER), in turn, constitutes a core component of NLP systems, enabling the identification and classification of entities, based on predefined categories of objects, that may appear in a body of text [55]. These categories may include, but are not limited to, names of individuals, organizations, locations, time expressions, quantities, medical codes, monetary values, and percentages. NLP and NER are considered together essential components and are widely leveraged in many research fields, including the classification of human-written game review texts [56], multilingual information retrieval (IR) chatbot implementation [57], processing of foreign languages and historical handwritten documents [58–60], cybersecurity domain projects [61–63], fake news detection [64], proposals in medical and biomedical sectors [65–69], geographic and geolocation tasks [70–72], as well as social media content analysis [73–77]. Nevertheless, with the emergence of Cultural Informatics (CI), both NLP and NER techniques hold significant potential for application in tourism and Cultural Heritage (CH), offering deeper insights and a refined understanding of socio-technological phenomena [78–80]. Sentiment analysis is another NLP technique used to estimate the hidden emotions expressed by the author in a piece of text, determining whether textual data is imbued with positive, negative, or neutral sentiments [81]. Applying sentiment analysis to user-generated content (UGC) helps organizations make data-driven decisions, improve visitor satisfaction, while enhancing companies' long-term sustainability [82]. To this end, sentiment analysis is extensively applied in social media mining, having diverse applications [27,83–87].

At the same time, based on the literature in the aforementioned scientific fields, we can conclude that a system combining the above methodologies, such as (i) the acquisition of crowd-sourced information from social media and from any other available online source, (ii) the homogenization of the harvested outcome, and (iii) the augmentation of the homogenized data through NLP procedures in a way that produces semantically richer knowledge, would provide a key technological factor driving the CH and tourism sectors forward. To the best of our knowledge, existing approaches neither consistently exploit these advances nor integrate them into a single, end-to-end framework that provides a holistic methodology and toolset; such a framework would yield synergistic capabilities that exceed those of its individual components and deliver actionable benefits for tourism and cultural operators. In this work, we present an innovative, cross-linked framework

supporting the acquisition, storage, homogenization and augmentation of heterogeneous touristic and cultural data. The proposed infrastructure, with all provided technological components (explained in detail in the following sections), is developed entirely by the authoring team using open source tools, components and libraries. The storage component is designed with the concept of a data lake architecture [88] in order to accommodate and efficiently exploit heterogeneous multilingual information registered in diverse formats and data structures, sourced from various online sources ranging from social media platforms (for instance, Facebook [89] and TripAdvisor [90] and geolocation services (like Google Maps [91] to cultural portals (such as Odysseus [92], SearchCulture [93], the Greek Ministry of Culture [94] and Athens Culture Net [95]). From this perspective, the suggested framework mainly aims to meet the operational needs triggered by the CI sector, empowering interested parties within the CH and tourism domains to harvest, stockpile, homogenize and extract significant insights generated effortlessly.

Therefore, this paper presents a novel unified framework that enables stakeholders in the cultural and tourism sectors to easily (i) initialize and launch data acquisition services, such as web scraping procedures on social media platforms and cultural websites and thematically focused crawlers that explore the clear web to discover additional digital cultural data points, (ii) store and organize the harvested outcome in data collections tailor-designed for specific tasks, (iii) homogenize the collected multilingual data, by identifying duplicate records and overlapping information, and merging the complementary ones, by applying normalization techniques, (iv) augment the homogenized data employing NLP processes such as NER, in order to find additional sites of interest, while performing sentiment analysis to achieve opinion mining and uncloak concealed emotions in user generated reviews and comments, and (v) automatically extract personalized touristic attraction trajectories adjusted to visitors' requirements.

Figure 1 provides a high-level view of the architecture of the proposed framework, the different services and functions implemented, their conceptual orchestration, and the data flow passing through each interconnected unit. The contributions of this work are summarized below.

- (i) We propose an innovative, integrated data lake framework capable of meeting several requirements posed by the CI, CH and tourism sectors, such as (a) acquiring cultural data from diverse online endpoints, (b) storing and managing information in properly configured NoSQL data collections, (c) applying data homogenization methods combining multiple string similarity measures with geolocation material, (d) identifying additional PoIs by employing NER language models, and (e) performing opinion mining on user reviews applying sentiment analysis pipelines. The framework exploits the augmented information to derive touristic and cultural routes based on personalized experiences of people who have already traveled there.
- (ii) We introduce the architectural approach behind the offered framework, discuss the technological means used to produce the aforementioned services, and illustrate the conceptual coordination of each encapsulated module. In addition, we provide detailed descriptions of the developed services, including the machine-driven data harvesting, the streamlined data homogenization, and the ML-powered augmentation of the data.
- (iii) We conduct variable experiments to assess the performance of the data homogenization stage, to evaluate the improved NER language models against baselines, demonstrating their robustness. Sentiment-analyzed data are used to create well-informed visualizations, in order to assess their potential to generate useful insights.

The proposed system designed to acquire, store, and harmonize data, while simultaneously deriving valuable insights such as sightseeing routes, constitutes a highly beneficial

asset for a wide range of cultural heritage and tourism stakeholders. These include policy and decision makers (such as ministries and regional authorities), tourism professionals and organizations (e.g., travel agencies and tour operators, mountaineering or urban sightseeing clubs etc.), as well as individual travelers seeking to discover remarkable sites and locations.

The rest of the paper is structured as follows. Section 2 presents related work on data collection and homogenization, with emphasis on the fields of culture and tourism; related work on NLP is also surveyed, emphasizing on the application of NER and sentiment analysis to non-formal texts. Section 3 presents the approach and methods used to harvest information from social media and also perform thematic harvesting. Section 4 presents the methodology used to homogenize the harvested data, as well as the homogenization result. Section 5 illustrates the data augmentation strategy, describing the four-level process implemented to extract tourist trajectories from the data. Section 6 provides the discussion on the overall data workflow, validation, and limitations of the proposed contribution. Finally, Section 7 concludes this article and provides future research directions.

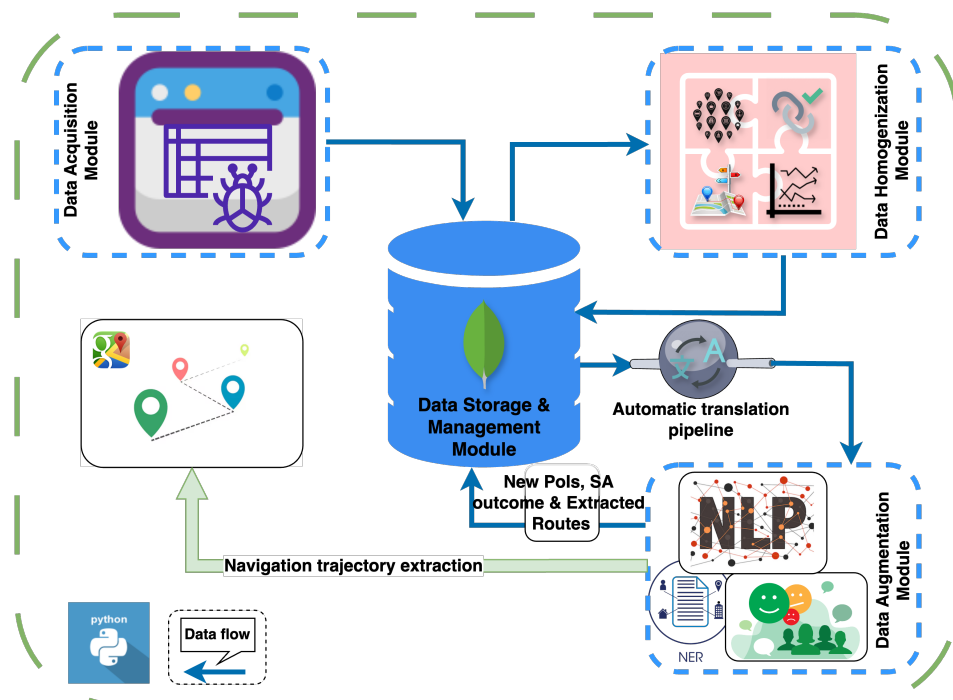


Figure 1. The architecture of the proposed framework.

## 2. Related Work

In this section we focus on the most prominent related approaches published in the literature concerning (a) data collection, (b) knowledge homogenization and integration, and (c) feature and entity extraction, emphasizing the culture and tourism domains.

### 2.1. Data Collection Employing Crawling Methods

Exploring the Web and social media networks for cultural and touristic data is a crucial task in the tourism sector, as it provides valuable insights into cultural sites and PoIs. In recent years, plenty of research works, which perform cultural and touristic content extraction by using a range of mining techniques, have been proposed. To start with, refs. [13,14] present systems able to generate data analytics on cultural heterogeneous content. Moreover, ref. [13] illustrates an online data lake, coined Hydria, that offers users without prior IT experience a tool to (i) harvest cultural information, (ii) store and organize

the gathered information, (iii) share whole/partial datasets and collaborate with other participants, (iv) administer users and enforce access control on the stockpiled data.

Two prominent paradigms concerning focused crawling research field are presented in [15,16]. Initially, the authors of [15] put forward a focused crawling procedure on immense amounts of cultural heterogeneous linked metadata appearing on Europeana digital platform with the view to enhance crawling effectiveness and precision, whereas UCrawler in [16] is a learning-based focused crawler, which applies a method that blends both content and URL analysis practices aiming to strengthen the accuracy and relevance of the topic with the initial page subject. Furthermore, ref. [17] introduces a couple of food image-caption pairs datasets, coined Kenya104K and KenyanFood13, hatched by implementing respectively location-based and keyword-based crawling strategies on Instagram.

The studies in [18,19] address geolocation and spatial data extraction from social media using crawling techniques. Specifically, ref. [18] enhances Flickr and YouTube crawling precision and recall to inform citizens about nearby emergency events, while ref. [19] introduces GeoLOCK, a lightweight search engine that integrates web and social media data through consecutive link crawling and API access. Moreover, SOUrCe [20] is based on website structures to navigate through social media pages and harvest data without supervision.

The article in [21] describes a supervised social media crawling approach capable to gather data from Facebook posts; it is used to aggregate a wide range of user interactions without crawling the whole post content. The studies in [22,23] involve Facebook users' profile crawling and captures data analysis, in order to demonstrate users' behavioral aspects and their psychological prosperity. Moreover, the work in [24] presents a Python-based web scraping strategy for historical Twitter [96] data extraction beyond API limits, while ref. [25] proposes a crawler that collects and classifies politically motivated hate tweets, distinguishing them from non-political or non-hateful content. Similarly, the research in [97] demonstrates a methodology for identifying Major Depressive Disorder emotions in user posts, leveraging ML algorithms on both publicly accessible datasets collected and those collected using web scraping techniques from Twitter and Reddit [98]. Furthermore, prior studies [26,27] concerning the TripAdvisor platform have focused on capturing user-generated hotel review texts through customized web crawlers that parse the HTML, CSS, and JavaScript components. In [26], the extracted dataset is subsequently employed in a text classification model to predict travelers' objectives based on the semantic content of the reviews. In contrast, ref. [27] concentrates on applying sentiment analysis techniques to the collected data, in order to uncover patterns in user opinions and experiences.

The review studies in [28–30] examine and compare diverse web crawling techniques. In [28], the authors evaluate the advantages and limitations of algorithms underlying various crawling methods. The work in [29] classifies focused crawling strategies into five main categories and proposes an architecture that clarifies key design challenges in web scraping. Similarly, ref. [30] connects automated information retrieval (IR) with web crawling, categorizing existing approaches into four groups and providing comparative assessments across multiple IR methodologies with adjustable user constraints.

## 2.2. Homogenization Approaches

Data redundancy is a common problem, as the same information can be encountered and stored more than once in a database. To address this issue, several data homogenization approaches have been proposed. The most prominent works in this area, which combine various similarity metrics and leverage supervised ML algorithms, are presented in [32–35]. More specifically, refs. [32,33] introduce two similar supervised ML methods that exploit multiple textual similarity scores, applied to decision trees, to distinguish whether a pair



of records is identical or not. In the same spirit, ref. [34] incorporates different feature similarity estimates into a priority queue that detects replicate database records through the proposed MSRD algorithm. The work in [35] exploits unsupervised Random Forests (RF) in combination with data conditioning, such as bins and string encoding, to compute similarity weights as a matching criterion for pairwise classification of duplicate records in the RLdata500 dataset.

Furthermore, the emerging field of entity resolution focuses on detecting, cleansing, and integrating heterogeneous data that refer to the same real-world entity, and it plays a crucial role in data homogenization tasks [36–39]. The study in [36] introduces a similarity-based indexing method that efficiently identifies meaningful record matches in heterogeneous environments. In [38], the S-HER algorithm addresses schema diversity by combining synonym-based reasoning with semantic enrichment to align feature names and values, further applying a weighted bipartite graph matching algorithm for correlation estimation. Similarly, ref. [37] presents a standalone entity resolution framework leveraging record block partitioning and graph-based techniques. The work in [39] proposes MULTIBLOCK, a continuous manifold iterative approach that performs entity matching across multiple rounds, efficiently resolving large-scale record blocks.

Several studies [40–43] explore deep learning architectures and embedding models to address the challenge of cleaning and linking identical records from heterogeneous data sources. Specifically, ref. [40] introduces BertLoc, a model that identifies and merges location records written in diverse forms across datasets. Similarly, ref. [41] proposes a two-tier method for duplicate detection and record consolidation. The approach in [42] mitigates data redundancy in scientific and restaurant datasets using sentence embeddings combined with an SVM classifier, while CompanyName2Vec [43] employs a BiLSTM model for company name matching based solely on vectorized name representations.

Duplicate detection and data integration are critical applications in the Big Data domain [44–46], yet the inherent 3V characteristics (volume, velocity, and variety) make these tasks highly challenging. To address this, ref. [44] proposes the Property Entropy Grouping Clustering algorithm, which measures record equivalence through entropy and clusters data accordingly before applying a sorted-neighborhood matching method. In [45], modular ontologies and MongoDB are employed to integrate heterogeneous data sources into large-scale ontological structures. Similarly, ref. [46] tackles duplicate detection in massive power grid datasets by combining the Spark analytics engine with an enhanced Simhash algorithm (IPOP-Simhash), incorporating improved fingerprint generation, weighting, and indexing methods for efficient record matching. More recently, ref. [99] introduces a deep learning-based homogenization approach that leverages knowledge graphs and Large Language Models (LLMs) for enhanced entity disambiguation in data lakes, while ref. [100] presents AutoGAT, which employs hierarchical graph attention mechanisms for semantic feature extraction and entity alignment.

### 2.3. Feature and Entity Extraction Using NLP Techniques

Knowledge extraction from structured (e.g., relational databases) or unstructured (e.g., raw text) sources is crucial across research and industry, supporting applications like data discovery, decision-making, prediction, and security. Named Entity Recognition (NER) constitutes a key knowledge extraction technique that automatically identifies and classifies entities within text. Before examining domain-specific studies, comparative analyses in [55,56] evaluate several NLP frameworks (e.g., SpaCy, NLTK, StanfordNLP), revealing that tool performance depends strongly on dataset characteristics and tuning. Similarly, refs. [101,102] assess transformer-based and hybrid deep-learning NER methods for cultural data, improving entity recognition through attention and CRF-based models.

The SpaCy ecosystem [103] is among the most widely adopted NLP frameworks. In [57], it powers a multilingual chatbot for Android and web platforms that adapts answers to the user's language, whereas [73] leverages pretrained SpaCy models to predict events from tweets based on linguistic context. The system in [104] extends SpaCy through a parameterized NER mechanism that enables custom, user-specific training over existing models. SpaCy is also extensively applied in cybersecurity and Cyber Threat Intelligence (CTI) research [61–63], where it supports entity detection and corpus creation for malware, cyberattacks, and vulnerabilities. Likewise, ref. [105] employs SpaCy for automated entity extraction and knowledge graph generation for cultural heritage (CH) objects, while ref. [106] focuses on transfer learning of pretrained word2vec models and clustering for identifying semantic relations among intangible cultural heritage documents.

NLTK (Natural Language Toolkit) [107] remains a fundamental Python-based NLP platform. It has been applied in works such as [74] for identifying person (PER), organization (ORG) and location (LOC) entities in Twitter datasets; it has also been extended through modules like NLPyPort [58] for Portuguese NLP tasks. Another open-source framework, FLAIR [108], facilitates experimentation with pretrained sequence tagging and text classification models. Domain-specific variants like Med-Flair [65] and BioNerFlair [66] target the medical and biomedical domains, integrating Bidirectional LSTMs with CRF for enhanced sequence labeling.

Transformer-based methods, especially those based on BERT and its variants, dominate current NER research. Examples include address entity tagging [109], historical French document parsing via CamemBERT [60], span-oriented labeling with S-NER [110], and fine-tuned BERT models for material identification in scientific literature [111]. NER systems also play a major role in social media analytics, extracting valuable insights from large-scale Twitter data. Notable works such as UAMNer and CWI [75,76] combine textual and visual information for named entity disambiguation, while ref. [64] employs NER for detecting propaganda and validating information from trusted sources. Similarly, ref. [77] applies RNN architectures (LSTM, BiLSTM, GRU) on disaster-related tweets to detect PER, ORG, and LOC entities, enhancing response coordination.

In the tourism and cultural sectors, where the volume of online information is immense, NER and ML-based solutions facilitate information extraction and search. The study in [79] compares transformer models (BERT, RoBERTa, XLM-RoBERTa) on Moroccan tourism data, finding BERT to yield the most accurate results. SpaCy fine-tuned models are used in [59] to improve recognition of transliterated Arabic and Muslim names in English texts, while ref. [80] employs SpaCy and BERT on tourism-related data from TripAdvisor, Traveloka, and Hotels.com to identify entities such as names, locations, and facilities.

In biomedical research, domain-specific BERT adaptations (PubMedBERT, BlueBERT, BioBERT) are widely used [67,68]. However, ref. [67] notes that pretrained models often fail on unseen terminology, while ref. [68] proposes a hierarchical shared transfer learning approach to address this limitation. NER also supports healthcare analytics, as shown in [69], which evaluates supervised and unsupervised methods for detecting psychosocial risk factors in medical texts.

Geospatial research has similarly benefited from NER advancements. Studies such as [70–72] apply NLTK, SpaCy, and Flair to historical corpora to identify geographic entities. Specifically, ref. [70] combines these tools with TagMe on English-translated Armenian texts, ref. [71] presents LOCALE, a rule-based system for Latin documents, and ref. [72] develops a Mexican geoparser for geographic NER in Spanish-language corpora.

Finally, several studies highlight persistent NER weaknesses, including misclassification and bias. Ref. [112] compares Stanford, LBJ, and IdentiFinder annotators, revealing

confusion among overlapping entity types, while ref. [113] uncovers gender bias in classifying person entities, showing poorer recognition of female names compared to male ones.

### 3. Data Harvesting

The evolution of the World Wide Web constitutes an inexhaustible source of information on many topics and scientific fields. Similarly, the proliferation of social media platforms and the continuously increasing number of user registrations establish social media as an extensive source of information, as well. Through web sites and social media platforms, substantial amounts of data can be harvested, which can be exploited in several CH and tourist organizations. The nature of these data may vary, from basic information (e.g., location and title) to enhanced aspects including photos and videos, tags, personal opinions etc., and, correspondingly, their use in CH and tourism applications may vary. Moreover, some data may be offered in curated and well-structured form, while other data may need to be analyzed, cleansed, structured, and validated in order to be exploited thereafter. In order to cover this wide range of needs related to data acquisition in the CH and tourism domains, we present a versatile and robust data harvesting framework that is conceptually divided into two individual components:

1. The *Social Media Harvesting* unit, which traverses through the social media public pages, identifying the targeted information, extracting and storing it in a suitably structured database. It encompasses task-specific web scraping modules termed spiders, and a coordination unit which initiates scraping tasks, delegating each task to the appropriate spider, and receives the harvested data for analysis, structuring, and storage. If additional sources of information are identified in the retrieved content, the coordination unit launches additional scraping tasks. The Social Media Harvesting unit is described in detail in Section 3.1.
2. The *Thematic Harvesting* unit performs web scraping on the clear web, discovering additional assets related to a specified topic of interest (CH and tourism in our case), by implementing focused crawling. The Thematic Harvesting unit is discussed in detail in Section 3.2.

Figure 2 provides a high-level overview of the system infrastructure, depicting the main modules and their conceptual organization within the data harvesting environment. The numbering shown in the figure corresponds to the distinct phases of the web harvesting process described in Section 3.1. Furthermore, the Social Media Harvesting unit, which constitutes the core component of the process, is detailed in Figure 3.

#### 3.1. Social Media Harvesting Unit

The *Social Media Harvesting* unit provides user-driven semi-automated data acquisition from social media websites. At the current stage of development, this unit is able to perform crawling on two of the most widespread social media platforms, Facebook and TripAdvisor, while implementations for additional social media platforms are in progress. The core of the unit is built on top of the Scrapy engine [114,115], enhancing it with specialized functionality for scraping structured content from social media pages. Web scraping procedures commence by providing the Social Media Harvesting unit with the appropriate seed URLs of the social media websites, which will serve as starting points to the harvesting procedure. Then, the Scrapy engine processes these seeds, identifying for each one the social media platform and the information elements to be scraped (such as cultural sites, event locations, PoIs, user reviews). Subsequently, it orchestrates the web scraping software (known as *spiders*) and triggers the execution of the appropriate spider instance to execute the task. At this stage, it is important to note that, in order to ensure compliance with privacy-preserving principles and to mitigate any risks related to the exposure of sensitive



or personally identifiable information, the spiders were explicitly designed to collect only aggregated and anonymized data at an appropriate level of granularity. This implementation strategy ensures that no individual user can be identified, either directly or through inference, from the exported datasets. Figures 2 and 3 illustrate the operation of the Social Media Harvesting unit, showing the data and control flows among its components, while Algorithm 1 provides a pseudocode representation of the process. As discussed earlier, the Scrapy Engine orchestrates the overall workflow, coordinating component interactions and initiating the appropriate tasks. The procedure is briefly summarized below, where the numbering corresponds to that shown in Figures 2 and 3.

---

**Algorithm 1** Social Media Web Scraping Procedure
 

---

```

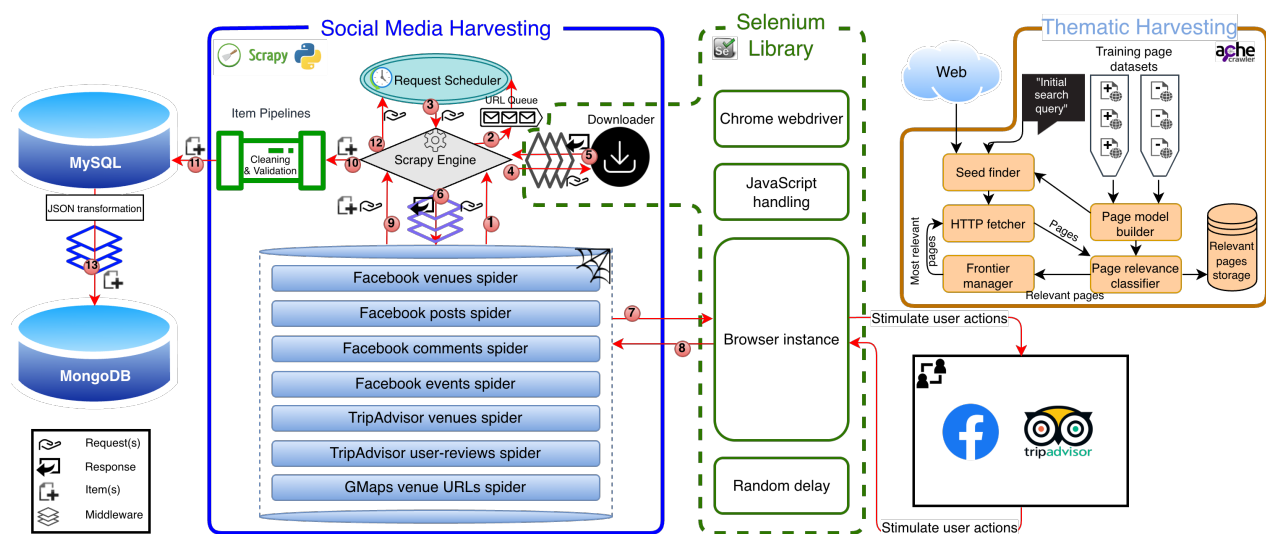
1: Input: Set of spiders  $\mathcal{S}$ , initial sets of URLs  $\mathcal{U}$ 
2: Output: Set of PoIs  $\mathcal{P}$ , set of reviews  $\mathcal{R}$ 
3: for each spider  $s_i \in \mathcal{S}$  do
4:   for each url  $u_j \in \mathcal{U}$  relevant to  $s_i$  do
5:     Scheduler.enqueue( $u_j$ )
6:   end for
7:   while Scheduler not empty do
8:      $request \leftarrow$  Scheduler.dequeue()
9:     Forward  $request$  to  $s_i$ 
10:     $s_i.extract(response)$ 
11:     $items, newReqs \leftarrow s_i.parse(response)$ 
12:    Store processed  $items$  in MySQL DB
13:     $\mathcal{P} \leftarrow \mathcal{P} \cup items$   $\triangleright$  Push the retrieved object in the set of PoIs
14:     $\mathcal{RU} \leftarrow items.get(names)$   $\triangleright$  Generate PoI review URL using unique venue name
15:    for each  $ru_k \in \mathcal{RU}$  do
16:       $s_k.extract(response)$ 
17:       $reviews, relationships \leftarrow s_k.parse(response)$   $\triangleright$  Also, find the relationships
        between PoI and PoI review
18:      Store  $reviews, relationships$  in MySQL DB
19:       $\mathcal{R} \leftarrow \mathcal{R} \cup reviews$   $\triangleright$  Push the retrieved review in the set of reviews
20:    end for
21:    Scheduler.enqueue( $newReqs$ )
22:  end while
23:  Transform  $reviews$ : columnar  $\rightarrow$  JSON
24:  Store transformed data in MongoDB
25: end for
  
```

---

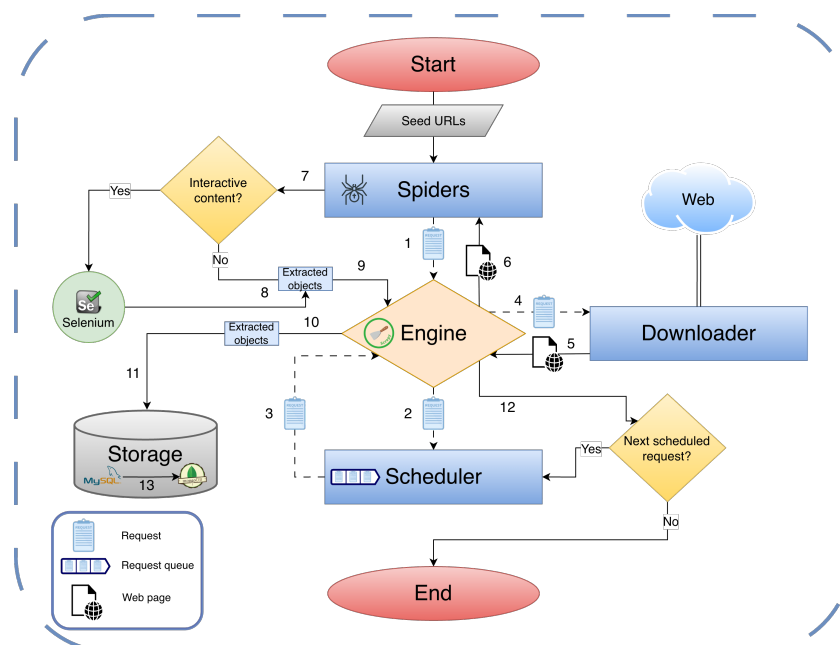
1. The *Scrapy Engine* starts by receiving the initial scraping requests from one of the available spiders. Each spider defines a customized crawling procedure, parses the website structure, and extracts the relevant data.
2. The Engine sends these requests to the *Request Scheduler*, which queues them and supplies them back as needed, ensuring a continuous flow of tasks.
3. The scheduled requests are then forwarded back to the Scrapy Engine.
4. The Engine passes the requests to the *Downloader*, which retrieves the web pages via the *Downloader Middleware*. This middleware manages all outgoing requests and incoming responses between the Engine and Downloader.
5. Once a page is retrieved, the Downloader sends the response back to the Engine through the Downloader Middleware.
6. The Engine routes the response through the *Spider Middleware* to the appropriate Spider. The Spider Middleware facilitates handling of responses, requests, and scraped items between the Engine and Spiders.
7. The Spider renders the page, simulates user interactions if necessary (e.g., scrolling down, to fetch additional content), navigates through the content, identifies the target

data fields, and extracts the information. The operation of spiders is described in more detail in Section 3.1.1.

8. The extracted data are processed and converted to structured item objects.
9. The Spider sends the item objects and any new requests back to the Engine via the Spider Middleware.
10. The Engine forwards the items to the *Item Pipeline*, a sequence of modules that validate and process each item before storage.
11. Processed items are initially stored in a MySQL relational database.
12. The Engine continues requesting the next items from the Scheduler, repeating the process until all scheduled requests are completed.
13. Finally, all items pass through the *Data Management Middleware*, which performs data transformation, converting the data from columnar form to JSON format. Once transformed, the data is stored in the MongoDB NoSQL database.



**Figure 2.** Web crawling architecture for data harvesting; the numbered steps correspond to the procedure described in Section 3.1.



**Figure 3.** Flowchart of the social media web scraping process; the numbered steps correspond to the procedure described in Section 3.1.

The crawling process is further illustrated in the high-level flowchart of Figure 3, where dashed lines denote message passing and solid lines represent data flow. In addition, Algorithm 1 presents a pseudocode overview of the procedure.

### 3.1.1. Spiders

Spiders are specialized modules that specify the crawling behavior for a given website or a group of websites. In particular, a spider defines how to interact with an HTML page by identifying the target data fields and by applying a traversal strategy across pages, such as following specific links (e.g., links from the main PoI page to the page listing related user reviews) [115]. All spiders developed within the framework are preconfigured for designated data sources. For each web scraping operation, the corresponding spider is instantiated and supplied with a list of URLs adhering to a predefined structural pattern. As each spider is specifically tailored to the HTML schema of its target website, input URLs must comply with the expected format. Consequently, if the supplied URLs do not match any of the existing spider configurations, the source cannot be processed with the current codebase, and a new spider must be implemented and executed for that task.

Overall, seven spiders have been developed and used. Each of the seven spiders was developed with a specific objective in mind, tailored to efficiently extract relevant information from its assigned platform; four of these spiders were configured to collect structured data from the Facebook network, while the next two were designed to crawl TripAdvisor. Another spider was created to crawl Google Maps, with the aim of collecting PoI websites that were not available on the aforementioned social networking platforms, complementing thus the available information. A detailed overview of all deployed spiders is provided in the following paragraphs.

All spiders were exploited in the context of European project TripMentor [116], which aimed to develop an interactive tourist guide for the region of Attica, Greece. The guide is accessible through the web as well as through a mobile application. In this context, spiders were used to cross-validate and populate elements of the database. Spiders have been also used for obtaining data for research on recommender systems.

#### Facebook Spiders

The four Facebook spiders were configured to navigate the structure of pages related to PoIs, automatically identifying and collecting key attributes such as geographical coordinates, opening hours, user reviews, posts, and comments. These spiders incorporate traversal strategies that allow them to follow links across multiple pages while avoiding unnecessary or redundant content.

The *Facebook venues spider* crawled Facebook in a time interval of about 48 h and collected approximately 10 K distinct PoIs of different categories (such as Arts and Entertainment, Breakfast and Brunch Restaurants, Cafe, Hotels, Landmarks, Museums, Parks and Outdoors and Restaurants) in the region of Attica, Greece. For each individual PoI, the spider navigated through the related public Facebook pages, detecting and harvesting the following data: the PoI name, the Facebook unique ID assigned to this PoI, its opening hours each day of the week, its website and its phone number, the enrolled email address, the PoI postal address, the geographical location of the venue (latitude and longitude), the total number of visitor check-ins in this place, the total review score generated by Facebook users who have visited this venue, and its category.

The *Facebook posts spider* exploited the venues' unique IDs exported from the previous spider, and generated the PoIs' profile page URLs, which would be used as the spider's initial seed URLs. To reduce both the time needed to collect the data and the number of requests against social media platforms, we exploited the number of PoI check-ins to

discern the most popular venues and we limited the data collection to these PoIs only. This filtering stage narrowed the number of seeds by approximately 83%, i.e., 1.7 K seed URLs qualified for the next data collection stage. In that stage, the spider extracted approximately 140 K posts related to the popular venues. Along with the captured post text, essential metadata for each post were extracted and stored, such as the post's unique ID on the social media platform, the profile origin and the PoI profile associated with the post, the post's URL, the date the post was uploaded on Facebook, and the reaction statistics for the post, i.e., the total number of user reactions, and the number of each reaction type (such as likes, loves, angers etc.).

In the same fashion, the *Facebook comments spider* was seeded with 587 profile page links of the most popular PoIs of Attica region, and the crawling process fetched approximately 240 K user comments on these venues. Similarly to the collection of posts, important metadata elements for comments were also retrieved, including the comment's posted date, the total number of user reactions to the comment, and the comment link address.

Subsequently, in order to gather social event information, the *Facebook events spider* was employed. Experts in the tourism and the event organization sectors manually identified PoIs profiles that are considered as principal event hosts, organizers and coordinators, and based on this information, a list of profile page URLs was compiled, which was employed as the initial seed for the events spider. The event harvesting process then commenced, resulting in the collection of information regarding a few hundreds of forthcoming events and their respective event postcards. The harvested event data included the event id, name and textual description, the date(s) of the event, the physical address of the location where the event takes place, the number of people who expressed their interest in attending the event, the URL of the event on Facebook, and the id of the PoI from which the event information was sourced.

### TripAdvisor Spiders

The two TripAdvisor spiders were designed to handle the unique layout and data organization of the TripAdvisor platform. They target similar PoI attributes to their Facebook counterparts, including location, ratings, reviews, and visitor comments, ensuring that the extracted data are both comprehensive and structured.

The *TripAdvisor venues spider* traverses the TripAdvisor site, identifying and extracting information related to PoIs in the Attica region. The process resulted in the collection of approximately 7 K data records in a broad range of PoIs of various categories such as monuments, landmarks, nature conservation areas, theme and aquatic parks, restaurants of different types, etc. The duration of the collection process was approximately 48 h. Each record encompasses fields including the PoI's name (both in English and Greek), the number of reviews for the PoI and its overall review score, the ranking order among the PoIs in the same category and the same district (e.g., "No 42 out of 2385 Restaurants—Athens"), the PoI's category(ies), the postal address and the phone number of the venue, and the URL of the venue on the TripAdvisor platform.

The *TripAdvisor user-reviews spider* implements the collection of TripAdvisor user reviews, and was used to retrieve user reviews for each of the 7 K aforementioned PoIs gathered by the TripAdvisor venues spider. Once the process concluded, after a period of approximately five days, approximately 300 K unique user reviews records were harvested. Each review record includes the title and the text of the user's review, the date that the review was uploaded on TripAdvisor, the review scores (the overall score, and -where available- the aspect-specific scores such as value, rooms, location, staff), as well as anonymized data fields about the review author. User-related data were collected with the aim of exploring and interpreting the background and the cognitive environment of

these individuals. It is noted however that the gathered user-related data do not contain personally identifiable information; the spider mainly collected generic fields, such as the place the user comes from (at country granularity), the user's number of votes, the total number of the user's contributions on TripAdvisor, the self-assigned user tags (such as "Like a local", "History lover", etc.), the self-assigned age range, the total number of cities the user has visited, the TripAdvisor user level (which indicates the experience level of the user with respect to their contributions on TripAdvisor), the total number of photos the user has posted on TripAdvisor, and the total numbers of review characterizations ("Wonderful", "Very good", "Moderate", "Poor" and "Very bad") they had set in places they visited.

#### GMaps Venue URLs Spider

The *GMaps venue URLs spider* harvests PoI URLs from the Google Maps geolocation service [117]. The need to establish this web scraping service came up in the information homogenization stage (described in detail in Section 4.1), when it was observed that numerous PoIs, sourced from Facebook and/or TripAdvisor, did not list a website address. Hence, the GMaps venue URLs spider attempts to find the respective PoI information on Google Maps, extract the website address and complement the PoI's record on the consolidated database. The execution of the GMaps venue URLs spider resulted in retrieving approximately 1.5 K missing URLs, which were stored in the respective venue records.

All spiders produce their results as structured item objects, which are subsequently passed to the Item Pipeline for post-processing. This stage performs data cleaning (e.g., removal of residual HTML tags), validation (ensuring that all required fields are present and correctly extracted), and duplicate detection and elimination (to remove any items that may have been scraped multiple times). These operations collectively ensure consistency and quality across data gathered from heterogeneous sources, facilitating accurate aggregation, analysis, and seamless integration into the TripMentor system.

#### 3.1.2. Web Scraping Issues

The crawling procedure follows the workflow of the Scrapy engine and has proven to be clear, flexible, easy-to-learn, and straightforward to implement. However, many sites and social media platforms (such as Facebook and TripAdvisor) make use of AJAX and JavaScript, to make their content more interactive, lightweight and user-friendly. These techniques support behavioral functionalities like text automatic translation, correlated with the country of origin of the user, infinite scrolling to confine the amount of data that needs to be transferred, as well as hidden content or dialogues that reveal information if the user interacts with them. The use of these techniques, however, encumbers the scraping procedure, since requests made by the scraping engine may retrieve partial information on some aspects (e.g., few user reviews only) or completely miss some others (e.g., hidden dialogues).

To address these issues, our crawler leverages the Selenium library [118,119], which enables automated control of web browsers. After successfully integrating the Selenium library into the Downloader Middleware and configuring the Chrome WebDriver, each spider encountering dynamic scripts on a social media platform was enhanced with the necessary Selenium extensions. This allowed the spider to simulate user interactions within the virtual Chrome browser, detect hidden or dynamically generated fields, and reliably extract the desired content, supporting the operations described in Steps 7 and 8 of the Social Media Harvesting unit (c.f. Section 3.1.1).

Another challenge was to ensure that the harvesting procedure did not issue an excessive number of requests to social media platforms, which would lead to blockings and



interruptions in the scraping process. To mitigate this, we implemented a strategy to vary the crawling speed through request throttling, i.e., introducing random delays between requests. By configuring the Downloader to pause for a random interval before retrieving consecutive pages from the same social media platform, the crawling behavior more closely resembled that of a human user, thereby reducing server load and minimizing the risk of blocking and interruption.

### 3.1.3. Ethical and Legal Concerns

To address potential ethical and legal implications arising from the application of web scraping techniques on social media platforms, we adopted specific practices to ensure the transparency and integrity of our data collection process. First, it is important to note that all the data used in this work are publicly accessible, and the examined sources provide such content without requiring user registration or authentication, even in the case of social media platforms. Thus, the information utilized was already available to the general public. Moreover, as discussed in relevant legal analyses [120], the web scraping of publicly accessible data does not violate the Computer Fraud and Abuse Act (CFAA), which protects the rights of researchers, academics, archivists, and journalists to access and analyze public information. To this end, web scraping has become a standard and widely accepted practice, even among major technology companies, which routinely collect public web content and online activity data [121]. Furthermore, GDPR [122] includes special provisions to allow the processing of data for research purposes (recital 50; article 14 par. 5b). Nevertheless, to ensure ethical compliance, particular care was taken to (i) preserve user anonymity, by avoiding the collection of personally identifiable information and removing any such data inadvertently acquired, and (ii) prevent excessive server load, through the implementation of throttling mechanisms in the spidering software to regulate the frequency of requests.

Regarding the practical applicability of our framework and its reliance on social media crawling, several online repositories of cultural content either explicitly permit web scraping (e.g., the Odysseus cultural portal [92] or offer open datasets and APIs providing relevant cultural data. For example, the Yelp Open Dataset [123] and the European Data Portal [124] are valid alternative or complementary sources that can be employed alongside, or in place of, direct web scraping.

## 3.2. Thematic Harvesting Unit

In order to further enrich the records of the database with additional PoIs as well as data elements on the already known ones, the Thematic Harvesting unit was developed to locate and extract this information from web sources; this unit employs adaptive crawling, implemented on top of the ACHE focused web crawler [125], to discover new cultural points on the open cyberspace and gather related information. The operation of the Thematic Harvesting unit commences by providing the appropriate seeds, through the accumulation of web pages that fulfill the predefined points of reference (e.g., pages that concern a given topic, such as “Cultural Heritage” or through a user-specified search query, for instance “Touristic destinations in Attica Greece”). Unlike the Social Media Harvesting, this unit employs a Page Relevance Classifier to distinguish relevant pages from irrelevant ones, according to predefined criteria.

More specifically, the initial seed URLs are inserted into the *Frontier Manager*, which functions as a priority queue. The *Thematic Crawler* module then dequeues URLs with the highest priority score from the Frontier Manager. The corresponding web page HTML elements are subsequently retrieved via the HTTP fetcher and passed to the *Page Relevance Classifier*, which determines whether the page is relevant or not to the given topic,

using ML methods. The configuration of the Page Relevance Classifier relies on applying `title_regex`, `url_regex`, and `body_regex` patterns. More specifically:

- the `title_regex` examines the HTML title tag of a page to check for matches with a given pattern;
- the `url_regex` tests the page URL against a list of predefined regular expressions; and
- the `body_regex` attempts to match patterns within the HTML content of the page.

Additionally, a blending method is supported, which combines all these regular expressions using Boolean operators (AND/OR) across multiple HTML fields (e.g., URL, title, content, and content type) to form a comprehensive Page Classifier.

To enhance the efficiency and reliability of regular expression (regex) patterns, a dedicated data preprocessing stage involving cleaning, normalization, and structuring of the raw HTML content can significantly improve page classifiers and strengthen the focused crawling process. This stage can be divided into two main phases: (i) URL normalization for link filtering and (ii) page content preparation for classifying pages as relevant or not. Regarding the `url_regex` patterns that determine which links are included or excluded, ACHE implements several built-in URL preprocessing and normalization methods prior to applying the patterns. These include case conversion (e.g., transforming URLs to lowercase), port number removal, trailing slash handling, fragment removal (text following a # in the URL), and domain filtering to prevent the crawler from deviating beyond the specified domains (via the `use_scope` feature). Duplicate URLs are also detected and removed before being added to the processing queue. To improve `title_regex` patterns, preprocessing steps such as case normalization, punctuation and special-character removal, and handling of common title structures (e.g., website names appended to the end of page titles) can be applied beforehand. Finally, preprocessing for the `body_regex` classifier is typically the most demanding step, as it requires transforming noisy HTML into clean text. Preparing the `body_regex` classifier involves: (i) extracting readable text from all HTML, JavaScript, and CSS elements, (ii) performing text normalization (including case-folding and whitespace or special-character handling), and (iii) removing repeated or template content such as headers, footers, navigation menus, and sidebars.

Beyond pattern matching, the Page Classifier also employs an ML-based text clustering approach using a Support Vector Machine (SVM) supervised learning model. This model is pretrained with a dataset comprising an equal number of on-topic (relevant) and off-topic (irrelevant) web pages, which were collected and stored by the *Page Model Builder* component. Pages classified as relevant are stored in the *Relevant Web Page* set, while irrelevant pages are discarded. To enhance theme coverage and discover additional topic-related pages, the *Seed Finder* [126] component is activated. It iteratively formulates new queries by synthesizing the initial search terms and the URLs of positively classified pages from the Page Model Builder, enabling thus the crawler to identify further seed URLs. This process continues until the predetermined maximum number of seed URLs is reached.

The output of the Thematic Harvesting unit consists of multiple websites with diverse HTML structures, unlike social media pages, which generally have a consistent format. These websites can range from cultural organization pages and travel and tourism magazines, to personal blog posts. To preserve this information, each retrieved page is cleaned by removing HTML tags, and its full page content is stored as raw text in a structured data repository within the NoSQL data lake [127] that we developed. The focused crawl successfully retrieved several culturally oriented pages. However, the desired content from most of these pages (approximately 85%) had already been captured by the spidering process. The most relevant content was harvested from the cultural portal Odysseus, where we were able to identify 147 historical PoIs in total, out of which 77 had not been retrieved from other sources. It has to be noted though that the majority of these PoIs were found to

be of very low tourist interest (despite their historical/archaeological value), being mainly “unspectacular” remains of walls or buildings that have sustained significant damage, which explains the fact that these PoIs were not found in other sources.

#### 4. Data Homogenization

The volume of information related to PoIs increases rapidly as data streams are continuously extracted from multiple online sources. During our analysis, we observed that information about the same cultural site often appears in different forms; some data are repeatedly recycled across sources, while others are complementary. Moreover, when examining the content of a single venue, we frequently uncovered references to additional PoIs, revealing hidden interconnections among them. Another complexity arises from language variation, as user reviews, posts, and comments are written by visitors from diverse backgrounds, who have experienced at least one cultural site in the Attica region of Greece. These contributions appear in multiple languages, primarily Greek and English, but occasionally in others, ranging from European to Asian languages such as Japanese. Consequently, the aggregated dataset is multilingual and highly heterogeneous. In addition, social media content is often noisy, further complicating integration tasks.

In this section, we present a method for the homogenization and consolidation of the information associated with related cultural sites, addressing the challenges of data unification and resulting in structured knowledge, as follows.

1. Harvested PoIs are grouped based on their website URL address.
2. For venues without an associated website, we apply a two-step process: (1) Initially, we identify the  $k$  nearest neighbors of a given PoI based on geographical coordinates. (2) Subsequently, for each candidate within the identified radius, we compute string similarity measures between venue names (using the method in [33]), selecting the most likely match(es) that correspond to the same PoI.

This approach enables us to link fragmented information across heterogeneous sources and build a more coherent representation of cultural heritage sites. The Website-Based Grouping and the Closest Neighbors Candidate Matching phases are described in the following subsections.

##### 4.1. Website-Based Grouping

In the Website-Based Grouping phase, harvested PoIs are grouped based on website URL address, under the rationale that places appearing in the database with similar or different names, yet having the same URL address must be treated as identical and can be merged. However, we noticed that numerous venue records (mainly corresponding to cafeterias and restaurants) do not have an associated website, while others sharing the same URL appear to be chain or franchise stores, owned and operated by larger companies, and are distributed in various geographical locations of the Attica region. Chain/franchise store venues should be excluded from merging, with their individual records maintained. Table 1 lists the 20 websites with the highest occurrence in the database.

In more detail, out of the total of 17,421 database entries, 6133 venue records do not list a website. As noted in Section 3.1.1, the GMaps venue URL spider was used to locate the missing information through the Google Maps service [117] and complement the database records. The crawler is provided with the list of venue names that do not have an associated website; for each venue name, the automated scraper uses Google Maps search bar to find the given venue; once the place has been located, the crawler copies the place’s website if available and updates the particular entry in the database. After approximately 24 h of Google Maps web scraping, the crawler identified approximately 1.5 K URLs and updated the corresponding venue records. Table 2 presents the new list with the 20 most

frequently occurring PoI websites in the database. To avoid erroneous assignment of websites to venues, the query against the Google Maps service was confined to consider only candidates within a physical distance range of 500 m.

**Table 1.** The 20 PoI websites (last accessed 31 December 2023) with the highest number of occurrences.

#	Venue's Website	Appearances
1	NULL	6133
2	<a href="https://www.coffeeisland.gr/">https://www.coffeeisland.gr/</a>	20
3	<a href="http://www.mikelcc.gr/">http://www.mikelcc.gr/</a>	19
4	<a href="http://www.starbucks.com.gr/">http://www.starbucks.com.gr/</a>	12
5	<a href="http://www.flocafe.gr/">http://www.flocafe.gr/</a>	11
6	<a href="http://www.cityofathens.gr/">http://www.cityofathens.gr/</a>	11
7	<a href="http://coffeeisland.gr/">http://coffeeisland.gr/</a>	10
8	<a href="http://www.lapasteria.gr/">http://www.lapasteria.gr/</a>	8
9	<a href="http://everest.com.gr/">http://everest.com.gr/</a>	7
10	<a href="http://www.coffeelab.gr/">http://www.coffeelab.gr/</a>	7
11	<a href="http://www.byzantinemuseum.gr/">http://www.byzantinemuseum.gr/</a>	7
12	<a href="http://www.b-eat.gr/">http://www.b-eat.gr/</a>	6
13	<a href="http://www.benaki.gr/">http://www.benaki.gr/</a>	6
14	<a href="http://www.gregorys.gr/">http://www.gregorys.gr/</a>	6
15	<a href="http://www.palmiebistro.gr/">http://www.palmiebistro.gr/</a>	6
16	<a href="http://www.dominos.gr/">http://www.dominos.gr/</a>	6
17	<a href="http://www.bios.gr/">http://www.bios.gr/</a>	5
18	<a href="http://www.theacropolismuseum.gr/">http://www.theacropolismuseum.gr/</a>	4
19	<a href="http://odysseus.culture.gr/h/1/eh151.jsp?obj_id=3348">http://odysseus.culture.gr/h/1/eh151.jsp?obj_id=3348</a>	4
20	<a href="http://www.coffeedive.com/">http://www.coffeedive.com/</a>	4

**Table 2.** Top 20 PoI websites (last accessed 31 December 2023) by number of occurrences, identified using the Google Maps scraper.

#	Venue's Website	Appearances
1	NULL	4804
2	<a href="http://www.mikelcc.gr/">http://www.mikelcc.gr/</a>	20
3	<a href="https://www.coffeeisland.gr/">https://www.coffeeisland.gr/</a>	20
4	<a href="http://odysseus.culture.gr/h/3/gh351.jsp?obj_id=2384">http://odysseus.culture.gr/h/3/gh351.jsp?obj_id=2384</a>	19
5	<a href="http://www.flocafe.gr/">http://www.flocafe.gr/</a>	13
6	<a href="http://www.starbucks.com.gr/">http://www.starbucks.com.gr/</a>	13
7	<a href="http://www.cityofathens.gr/">http://www.cityofathens.gr/</a>	11
8	<a href="http://coffeeisland.gr/">http://coffeeisland.gr/</a>	10
9	<a href="http://odysseus.culture.gr/h/2/gh251.jsp?obj_id=912">http://odysseus.culture.gr/h/2/gh251.jsp?obj_id=912</a>	9
10	<a href="http://www.coffeelab.gr/">http://www.coffeelab.gr/</a>	9
11	<a href="http://www.lapasteria.gr/">http://www.lapasteria.gr/</a>	8
12	<a href="https://www.goodys.com/">https://www.goodys.com/</a>	8
13	<a href="http://www.byzantinemuseum.gr/">http://www.byzantinemuseum.gr/</a>	8
14	<a href="http://www.emst.gr/">http://www.emst.gr/</a>	7
15	<a href="http://everest.com.gr/">http://everest.com.gr/</a>	7
16	<a href="https://www.cityofathens.gr/episkeptes/aksiotheata/diadmomes-stin-istoria-tis-athinas/o-kipos-tis-athinas">https://www.cityofathens.gr/episkeptes/aksiotheata/diadmomes-stin-istoria-tis-athinas/o-kipos-tis-athinas</a>	7
17	<a href="http://www.dominos.gr/">http://www.dominos.gr/</a>	6
18	<a href="http://www.palmiebistro.gr/">http://www.palmiebistro.gr/</a>	6
19	<a href="http://odysseus.culture.gr/h/2/gh251.jsp?obj_id=12863">http://odysseus.culture.gr/h/2/gh251.jsp?obj_id=12863</a>	6
20	<a href="http://www.panathenaicstadium.gr/%CE%91%CF%81%CF%87%CE%B9%CE%BA%CE%AE/tabid/40/language/el-GR/Default.aspx">http://www.panathenaicstadium.gr/%CE%91%CF%81%CF%87%CE%B9%CE%BA%CE%AE/tabid/40/language/el-GR/Default.aspx</a>	6

Having more place URLs available, the homogenizing of multilingual and heterogeneous venue data may commence. As an example of integration, we can inspect the places referenced by the first URL that is not related to a branch company (No 4 in Table 2). Table 3 presents the 19 database entries bearing the aforementioned URL address. As we can observe, all venue names listed point out the very same place, which is the Athenian Acropolis [128], written in several languages, such as English, Greek, Hebrew, French, Polish, Chinese, or even Greeklish (i.e., Greek language spelled using Latin characters [129]).

Having identified the subsets of equivalent venues, their records can now be merged. When merging a set of records, if a field (e.g., extended description, rating etc.) appears exactly in one of them, the specific value is copied to the merged record. If a field has a value in multiple source records, the value in the result record is determined as follows:

1. For numeric values representing counts, such as the number of check-ins or the number of ratings, the sum of individual values is computed.
2. For numeric averaged values, such as the overall rating score, a new value is computed using the formula  $new\_value = \frac{\sum_{i=1}^k v_i * count_i}{\sum_{i=1}^k count_i}$ , where  $v_i$  is the value of the  $i^{th}$  record
3. For values not falling in the above categories, the following methods can be applied:
  - majority voting, where the most frequently occurring value is maintained,
  - value timestamp, where the value associated with the most recently updated source is selected,
  - manual review, where a human operator reviews values and selects the correct value.

**Table 3.** Database entries (both English and non-English) associated with the URL [http://odysseus.culture.gr/h/3/gh351.jsp?obj\\_id=2384](http://odysseus.culture.gr/h/3/gh351.jsp?obj_id=2384) (last accessed on November 2025).

Id	Name	Category
7408	Acropolis, Athens Greece	Arts and Entertainment
7642	Atina Akropolis	Arts and Entertainment
7753	אקרופוליס אתונה	Arts and Entertainment
13618	Athena, Acropolis	Hotels
14029	Acropole d'Athènes	Landmarks
14037	Vraxakia Acropole	Landmarks
14071	Akropoli	Landmarks
14077	Acropolis (Athens)	Landmarks
14078	The Acropolis, Athens	Landmarks
14307	Grecja, Ateny, Akropol	Landmarks
14308	雅典 城	Landmarks
14328	城 雅典	Landmarks
14379	The Acropolis	Landmarks
14403	The Acropolis	Landmarks
14465	North and east slope of Acropolis	Landmarks
14754	Ακρόπολη Αθήνα	Museums
14765	Acropolis, Athens Greece	Museums
14817	Athenean Acropolis	Museums
15197	Akropolis Athena	Parks and Outdoors

#### 4.2. Closest Neighbors Candidate Matching

For the remaining venue records lacking a URL and thus remaining unclassified, we developed a parameterized module to identify candidate matches based on PoI names stored in the main database. The “TMDedupe” package provides a complete pipeline for comparing a source of PoIs against the “main” venues table (hereafter DB2 and DB1, respectively) to identify potential matches. Supported sources include (i) a supplementary database table of PoIs, which may coincide with the main table (i.e., self-join matches),



or (ii) a custom PoI provided by the user. The matching process proceeds in two stages: (i) blocking, to reduce the number of candidate pairs, and (ii) ranking, through the computation of string similarity for each pair of PoI names, to classify them as true or false matches. The two stages of the matching process are described in the following sections.

#### 4.2.1. Blocking

To identify potential matches, we need to compare DB2 against DB1 and thus check  $n * m$  candidate pairs, where  $n$  and  $m$  are the number of PoI records stored in DB1 and DB2, respectively. This task can prove time-consuming, depending on the magnitudes of  $n$  and  $m$ . In order to reduce computational costs, we use blocking [37,130] by locating the venue pairs that are more likely to match and restricting comparisons to these venue pairs only. The criterion used for the blocking phase in our approach is the physical distance between venues: more specifically, we consider as matching candidates for any given PoI only venues that are located within a radius of 10 m from the target PoI. Then, to achieve quicker nearest-neighbor lookups, we use the KDTree structure [131,132] to index the PoIs in one of the interlinked tables.

#### 4.2.2. Approximate Similarity Measures

After reducing the set of candidate record pairs, we compute string similarity scores for each pair of PoIs to assess whether they refer to the same entity. To do so, we utilize nine similarity measures summarized in Table 4. These measures are selected considering their effectiveness and their complementary strengths in capturing variations in spelling, abbreviations, and typographical differences; a detailed analysis and evaluation of the performance of the utilized string similarity metrics is presented in [33]. Each computed score is compared against its corresponding threshold; pairs exceeding the threshold are classified as matches, while the rest are considered non-matches. The thresholds, adopted from [33], are widely used as general-purpose, domain-independent defaults. However, they can be adjusted by the user via a configuration file to suit specific applications.

**Table 4.** String similarity measures applied to venue name pairs.

#	Measure	Threshold
1	Damerau-Levenshtein	0.55
2	Jaro	0.75
3	Jaro-Winkler	0.70
4	Jaro-Winkler Reversed	0.75
5	Sorted Jaro-Winkler	0.70
6	Cosine N-Grams	0.40
7	Jaccard N-Grams	0.25
8	Jaccard Skip-grams	0.45
9	Dice Bi-Grams	0.50

#### 4.2.3. Homogenization Outcome

The outcome of the above two-step process is stored in a CSV file. Each entry/line in the output file contains the following attributes/columns:

1. `cand_id`, an integer numerical value greater than zero that refers to the primary key index of the candidate PoI in DB1 (the “main” venues table)
2. `cand_name`, a string value that denotes the name of the candidate PoI in DB1
3. `cand_lon`, a float numeric value in the range  $[-180, 180]$  that corresponds to the longitude of the candidate PoI in DB1
4. `cand_lat`, a float numeric value in the range  $[-90, 90]$  that corresponds to the latitude of the candidate PoI in DB1

5. `id`, an integer numeric value that refers to the primary key index of the matching PoI in DB2, linking the specific PoI from the main database to its corresponding potential match in the list of newly harvested PoIs
6. `name`, a string value that denotes the name of the PoI in DB2
7. `lon`, the longitude of PoI in DB2
8. `lat`, the latitude of PoI in DB2
9. `<similarity measures>`, a list of real-valued scores in the interval  $[0, 1]$ , where each element corresponds to the similarity value produced by a specific string similarity measure applied to the candidate pair of entries
10. `<similarity measures>_status`, a list of Boolean values, where each element indicates whether the corresponding similarity score exceeds its threshold (Table 4)
11. `on_translit`, a Boolean flag indicating whether the similarity scores were computed on the transliterated PoI names or on their original forms
12. `status`, a Boolean value representing the final matching decision for the candidate pair, obtained by combining the outcomes of all applied similarity measures

The final decision follows a variation of the hard-voting (i.e., majority voting) rule. The input parameter `voting_size` specifies the minimum number of positive votes (i.e., outcomes from the applied similarity measures favoring the true label), required for a candidate pair to be classified as a match. Table 5 presents an illustrative example of the homogenization outcome.

**Table 5.** Homogenization outcome sample with both English and non-English entries; for presentation purposes, decimal values are reported with two digits of precision.

Attribute	Entry 1	Entry 2	Entry 3	Entry 4
<code>id</code>	17	2507	2959	5355
<code>name</code>	Minnie the Moocher	Mia zwi tinexoume	Pizzoteca Nel Pireo	Tholos Cafe
<code>lon</code>	23.74	23.71	23.64	23.71
<code>lat</code>	37.97	37.98	37.94	37.97
<code>cand_id</code>	2878	3831	5371	4371
<code>cand_name</code>	Minnie The Moocher - Bar	Μια ζωή την έχουμε	Gilly Nel Pireo	Μέντωρ Cafe
<code>cand_lon</code>	23.74	23.71	23.64	23.71
<code>cand_lat</code>	37.97	37.98	37.94	37.97
<code>Damerau_Levenstein</code>	0.82	0.68	0.55	0.54
<code>Jaro</code>	0.94	0.76	0.67	0.75
<code>Jaro_Winkler</code>	0.96	0.85	0.67	0.75
<code>Jaro_Winkler_Rev</code>	0.83	0.84	0.60	0.85
<code>Sorted_Jaro_Winkler</code>	0.85	0.76	0.60	0.85
<code>Cosine_N_Grams</code>	0.92	0.50	0.57	0.39
<code>Jac_N_Grams</code>	0.83	0.33	0.38	0.24
<code>Dice_Bi_Grams</code>	0.88	0.51	0.56	0.4

**Table 5.** *Cont.*

Attribute	Entry 1	Entry 2	Entry 3	Entry 4
Jac_Skip_grams	0.87	0.50	0.54	0.43
Damerau_Levenstein_status	1	1	1	0
Jaro_status	1	1	0	1
Jaro_Winkler_Rev_status	1	1	0	1
reverse_winkler_status	1	1	0	1
Sorted_Jaro_Winkler_status	1	1	0	1
Cosine_N_Grams_status	1	1	1	0
Jac_N_Grams_status	1	1	1	0
Dice_Bi_Grams_status	1	1	1	0
Jac_Skip_grams_status	1	1	1	0
on_translit	0	1	0	1
status	1	1	1	0

Each column in Table 5 corresponds to the comparison of a pair of PoIs whose physical distance is below the threshold. The data of the first PoI within each pair are listed in data rows 1–4 (id-lat), while the data of the second PoI within the pair are listed in data rows 5–8 (cand\_id-cand\_lat). These rows are followed by the similarity measures and the boolean flags described above. According to the computed similarity measures, the first three pairs of PoIs (i.e., entries 1–3) converge, indicating that they refer to the same PoI despite variations in their names. In contrast, the last entry highlights the dissimilarity of the PoI pair. Consequently, only the first three PoI pairs are merged in our dataset.

By applying the proposed process, we reduced the number of PoIs by 57.3% through the homogenization of duplicate information appearing in multiple records of the same venue, resulting in a clean PoI dataset comprising approximately 7.5 K entries that represent landmarks in the Attica region of Greece. The homogenization process effectively identified and merged 9976 duplicate records from the initial dataset, which originally contained about 17 K entries. Table 6 presents the total number of extracted Points of Interest (PoIs) from each source, both before and after the harmonization phase. Moreover, Table 7 provides insights into the overlap of sources, considering only the retained PoI data. As observed, 0.04% of the final dataset content originates from all three sources, while 72.23% is derived from both Facebook and TripAdvisor, and 0.09% is derived from TripAdvisor and Odysseus. The remaining records are sourced exclusively from Facebook (10.64%), TripAdvisor (15.96%), and Odysseus (1.04%), respectively.

**Table 6.** Number of PoIs extracted from each source, before and after the homogenization stage.

Source	# of Venues
Facebook	10,405
TripAdvisor	6869
Odysseus	147
Before/After	# of Venues
Total PoIs before homogenization	17,421
Retained PoIs after homogenization	7445

**Table 7.** Sources overlap considering the retained PoI data.

Source(s)	Overlap
All three sources	0.04%
Facebook + TripAdvisor	72.23%
Facebook + Odysseus	0%
TripAdvisor + Odysseus	0.09%
Facebook (exclusively)	10.64%
TripAdvisor (exclusively)	15.96%
Odysseus (exclusively)	1.04%

To gain further insight into the effectiveness of the homogenization process, visualizations were created illustrating the contribution of different similarity measures in the identification of duplicates. Figure 4 shows a four-dimensional voxel grid plot that identifies which PoI names can be considered true matches to geographically nearest candidate PoIs (based on their coordinates). Specifically, the X and Y axes represent the PoI ID and the candidate PoI ID, respectively, while the Z axis encodes the similarity measures applied to each candidate toponym pair. Each layer of the plot indicates whether a pair of PoIs is identical according to a specific similarity measure. Points colored red (or orange/yellow, depending on the threshold for each string similarity measure) are considered duplicates, whereas points in other colors are classified as distinct places.

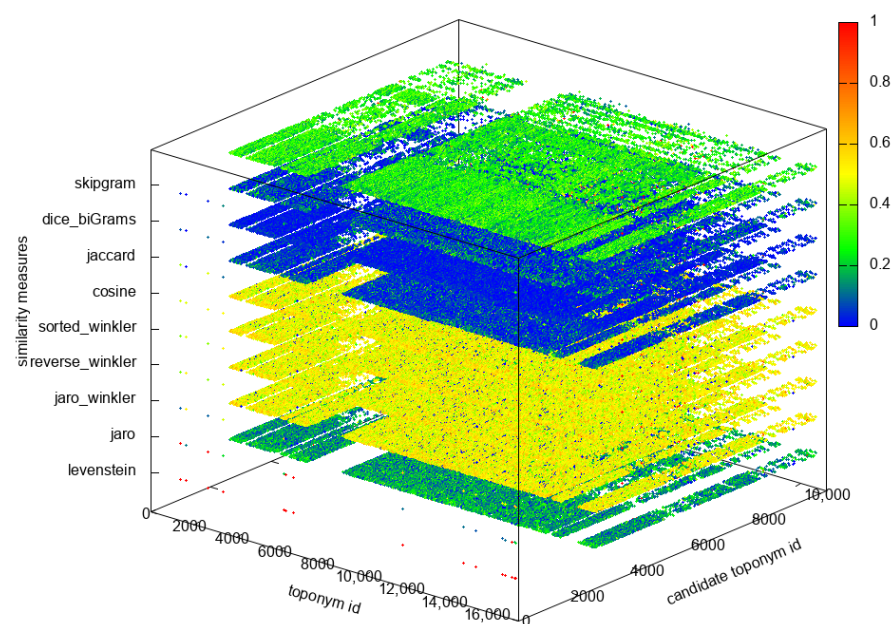
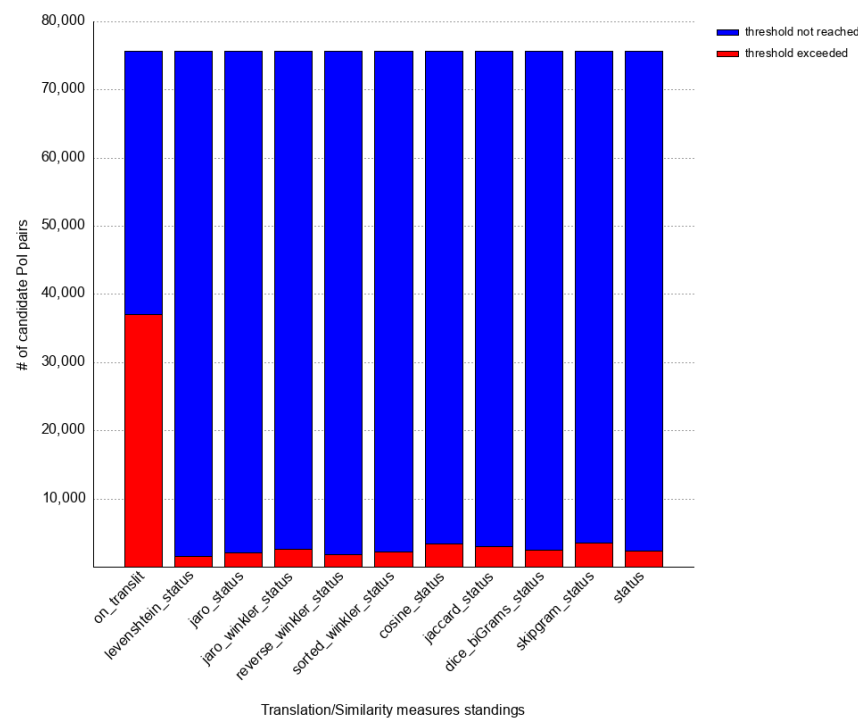
**Figure 4.** Matching PoIs identification.

Figure 5 presents a stacked column histogram summarizing the results of each string similarity measure applied to candidate PoI pairs. The first column shows how many toponyms required translation before comparison, and the last column reflects the final match decision based on the hard-voting rule across all measures. Red bars indicate pairs that exceeded the threshold for a given measure (or required translation in the first column), while blue bars represent pairs that did not meet the threshold and are thus considered distinct. For a detailed description of the thresholds for each measure, see Table 4.



**Figure 5.** Aggregated similarity measure and translation states.

#### 4.2.4. PoI Record Completeness and Correctness

To ensure uniformity and completeness across all PoI records, a record completion procedure is applied. Table 8 summarizes the percentage of record fields available in each data source, illustrating how missing information is supplemented during the homogenization phase. For example, if two identical PoI records from different sources contain complementary details, e.g., one providing the address and the other the geographical coordinates, these are compared and merged to form a single, consistent, and complete PoI record. Regarding the Facebook data, 94.11% of the retrieved records included a non-empty address field; however, in 36.26% of these cases, the address merely replicated the place name and was therefore deemed uninformative. Additionally, 5.94% of the records contained only partial address details (e.g., city or area information). Finally, missing English PoI names were completed through automatic transliteration of Greek names to English, ensuring consistency across the dataset.

**Table 8.** The percentage of PoI record fields available in each data source.

Source	Name_gr	Name_en	Categories	Address	Phone	Site	Coordinates	Days_Hours_Open	E-Mail
Facebook	51.85%	48.15%	100%	94.11% <sup>a</sup>	67.54%	39.96%	92.53%	36.04%	18.31%
TripAdvisor	50.59%	49.41%	93.13%	99.52%	91.63%	-	-	-	-
Odysseus	100%	100%	100%	Sporadic appearances, not structured	32.78%	-	100%	67.26%	28.92%

<sup>a</sup> Note that, 36.26% duplicated the place name and 5.94% included only partial address information.

Finally, by checking a sample of the collected PoI records, we discovered that in some cases the recorded geographical coordinates did not match the actual location. To overcome this issue, a reverse geocoding process was employed, identifying 768 venues (10.32% out of the retained ones) where the harvested coordinates did not correspond to the verified address. The coordinates were updated to reflect the correct ones.



## 5. Data Augmentation

Following the completion of the data homogenization task, we aimed to further enrich the dataset by discovering additional venues, and extracting relevant knowledge from the collected information. To this end, we applied a series of ML techniques to (a) expand the data, (b) generate sightseeing trajectories, (c) perform sentiment analysis on user-generated content from Facebook and TripAdvisor reviews, and (d) recommend routes to the users. In the remainder of this section, we first describe the overall process followed to achieve these goals, and subsequently we analyze the operation of the components that implement the individual tasks.

Since the retrieved data consisted of multilingual content, a translation module was first employed to homogenize the text and ensure consistent processing in subsequent phases. After evaluating several state-of-the-art pre-trained NLP models, we selected the most suitable ones and applied them through the SpaCy open-source library [103] to perform NER on the collected content. To improve entity extraction accuracy, the pre-trained models were enhanced with a custom rule-based pipeline leveraging venue names and categories from both Facebook and TripAdvisor (stored in a global venue table). Through a series of experiments, we enriched our dataset by identifying new places and locations (labeled by the NER module as “GPE” [GeoPolitical Entities, used for entities with a governing body like cities and countries] and/or “LOC” [Location, i.e., a physical location or area that is not a GPE], respectively) that were not previously included in the database. These newly discovered venues were then linked to the texts in which they appeared. Finally, to propose appealing sightseeing routes for tourists, we combined the identified nearby places and locations with sentiment information derived from visitor reviews, using the SpacyTextBlob [133] pipeline for sentiment analysis.

### 5.1. Automatic Translation Component

This component aimed to provide a uniform representation of the harvested data to facilitate their processing. To this end, we implemented a translation module capable of automatically detecting the source language of a given text, including non-standard forms such as Greeklish, and converting it into English. This was achieved using the open-source Python library deep-translator [134] that provides support for several well-known translation services (e.g., Google Translate, DeepL, LibreTranslate). Following extensive manual testing with a sample of more than one hundred tourist reviews related to the Temple of Poseidon at Sounion [135], we found that the most accurate translations were obtained when configuring the deep-translator library to support Google Translate.

### 5.2. Entity Extraction Component

Named Entity Recognition (NER) is a subfield of NLP concerned with identifying and extracting entities from text. An entity is defined as a word or a sequence of words that refers to an object classified under a predefined category. Typical examples of entities include geographic locations, organizations, personal names, historical events, dates, and more. To implement NER, we first evaluated which pre-trained statistical model should be integrated with an NLP library to most accurately identify location names in our dataset, minimizing the risk of misclassification. As described in Section 3, user-generated reviews from TripAdvisor and Facebook were input into the NER processing module.

After a thorough review of existing literature (c.f. Section 2.3) and extensive experimentation with several state-of-the-art NLP frameworks, we concluded that both SpaCy [103] and Hugging Face Transformers [136] are robust, open-source NLP libraries well-suited to our dataset. These libraries offer a range of NLP functionalities, including NER and sentiment analysis, making them appropriate choices for our application.

### 5.2.1. NER Model Evaluation and Selection

Following the selection of the library to be used, there was also the need to select one of the available pre-trained language representation models in order to discover, in an efficient way, places of interest and geographic locations in non-formal texts. The SpaCy library is shipped with various pre-trained models for documents written in many different languages, and for each language different flavors are provided to suit specific NLP tasks, application domains and execution environments. In particular:

1. The *Type* parameter of the model indicates the NLP tasks that can be supported by the model. More specifically, *core*-type models implement a versatile pipeline that offers tagging, parsing, lemmatization and NER, whereas *dep*-type models can be used only for tagging, parsing and lemmatization;
2. The *Genre* parameter of the model indicates the corpus category on which the model is trained, which in turn affects the suitability of the model for application domains. More specifically, the corpora categories are *web* (the corpus comprises generic web content) and *news* (the corpus is limited to material harvested from news agencies, sites etc.);
3. The *Size* parameter of the model indicates either the volume of data of the model, with available versions being small (*sm*), medium (*md*) and large (*lg*), or that the model is a transformer-based one (*trf*).

In the implemented system, the NER module uses the English language model (since all content has been translated to English by the Automatic translation component). For the model type and genre, we have chosen the *core* type, as we needed a pipeline that fully supports NER tasks, and the *web* genre, since the texts that will be processed are generic user input, which are best akin to written online texts (such as blogs, news and comments), instead news-only content (for instance news' articles). Following these decisions, we experimented with variants of the model differing in the *size* parameter to identify the variant that is best-suited for location entity discovery. More specifically, the following variants were evaluated:

- The small-size model (encoded as `en_core_web_sm`),
- The medium-size model (encoded as `en_core_web_md`),
- The large-size model (encoded as `en_core_web_lg`), and
- The transformer-based model (encoded as `en_core_web_trf`).

It is worth noting that the first three models constitute English pipelines optimized for CPU containing the `tok2vec`, `tagger`, `parser`, `senter`, `ner`, `attribute_ruler` and `lemmatizer` components, as well as default word vectors (except the *sm* model), while the transformer-based model introduces RoBERTa base [137], a transformer model pre-trained on a large English corpus in a self-supervised fashion. This model is specially adjusted to fit in SpaCy ecosystem pipelines providing `transformer`, `tagger`, `parser`, `ner`, `attribute_ruler` and `lemmatizer` features; however, it does not support static word vectors.

As an alternative to the SpaCy ecosystem, we also experimented with one of the major Hugging Face Transformer-based NER models known as *bert-large-NER* [138]; this is a fine-tuned BERT model providing high performance in NER tasks, while specializing in location (LOC), organization (ORG), person (PER) and miscellaneous (MISC) entity identification.

To further enhance the performance of the NER task, we augmented the processing pipeline to utilize the DBpedia ontology [139] by employing the DBpedia Spotlight [140] library for SpaCy. Through this module, we submitted appropriate SPARQL queries to the DBpedia ontology API, effectively identifying Dbpedia entities in tourism reviews.

The identified entities were subsequently tagged with the corresponding Dbpedia ontology classes or subclasses.

### 5.2.2. Improving Model Accuracy

After experimenting with the above listed models to identify the best performing one for the purposes of NER task in our corpus, we identified that the models still failed to recognize or match many places, venues and landmarks within the Attica region. To overcome this shortcoming, we exploited the Entity Ruler pipeline component [141], which allows the attachment of named entities according to pattern dictionaries, such as token-based regulations or precise expression matches. More specifically, the Entity Ruler is a pipeline component within the SpaCy ecosystem that enables rule-based NER. It allows the completion of annotated span tuples, returned by the statistical Entity Recognizer, through pattern-based rules that rely on token sequences or exact phrase matches. Integrating the Entity Ruler with SpaCy's statistical Entity Recognizer can enhance the overall accuracy of NER tasks, whereas using it independently results in a purely rule-based approach. The Entity Ruler operates on the basis of user-defined Entity Patterns (see Figure 6), which are dictionaries containing two keys: *label* and *pattern*. The *label* defines the entity annotation assigned when a match occurs, while the *pattern* specifies the matching expression. The Entity Ruler is incorporated into a SpaCy Language object, which represents the text processing pipeline, via the `add_pipe` method [142]. When the language object processes a document, it identifies pattern matches within the Doc object (i.e., a container for linguistic annotations) [143] and embeds them as entities using the corresponding labels. In cases of overlapping matches, the pattern covering the greatest number of tokens is prioritized; if matches are of equal length, the one appearing first in the document is selected. By integrating this component into the NER processing pipeline, we achieved a substantial increase of the percentage of named entities that are successfully recognized.

Integrating Entity Ruler into our system was a two-step process that involved (a) creating a global table of PoIs, by merging both Facebook and TripAdvisor venue collections into a combined view table and (b) transforming the contents of the table into a representation that can be readily utilized by the Entity Ruler component. The latter step was accomplished through a custom Python script, which exported the global PoI table to a pattern-based JSONL (newline-delimited JSON) file that contains one pattern object per line. Each line contains the lexical patterns that refer to the entity (which is searched for in the corpus on which NER is performed), and the label of the entity (i.e., a description). For a given entity, multiple entries can be provided to support the recognition of variants of the named venue, such as alternative names or mixed case writing. Figure 6 presents an excerpt of the file with entity specifications provided to the Entity Ruler module.

```
{“label”: “Monuments and sights Monuments and statues Ancient ruins”,
“pattern”: “Temple of Olympian Zeus”}
{“label”: “Monuments and sights Monuments and statues Ancient ruins”,
“pattern”: [{“LOWER”: “temple”}, {“LOWER”: “of”}, {“LOWER”: “olympian”},
{“LOWER”: “zeus”}]}
```

```
{“label”: “Museums Art museums”, “pattern”: “Benaki Museum”}
{“label”: “Museums Art museums”, “pattern”: [{“LOWER”: “benaki”},
{“LOWER”: “museum”}]}
```

**Figure 6.** An excerpt of the file with entity specifications provided to the Entity Ruler module.

For each new place entry in the database, the NER system updates and rebuilds the Entity Ruler's input file, preparing it for subsequent tasks.

### 5.2.3. Evaluation of the NER Models

To determine which of the above models is best suited for identifying places and venues (GPE and LOC entities) in non-formal corpora, we created a platform to collect expert answers (Figure 7) and conduct a series of experiments evaluating both accuracy and efficiency in named-entity classification and tagging. Specifically, we measured precision, recall, and F-measure at the token level for each model using the same dataset [144]. These evaluation metrics are briefly described below.

#### Evaluation of Named Entities Detected in TripAdvisor & Facebook reviews

Please read the review text, estimate highlighted labels referring to Named-Entities, discover untagged entities (if any) that should have been labeled, and fill in all numeric fields with the appropriate values keeping in mind the following guidelines (focusing mainly on locations, places of interest, monuments and ancient ruins).

NER model: **en\_core\_web\_sm**

Show instructions

The new **Acropolis PERSON** museum is one of the top museums that one can visit. The exhibits need no introduction, centuries of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

NER model: **en\_core\_web\_md**

The new **Acropolis LOC** museum is one of the top museums that one can visit. The exhibits need no introduction, **centuries DATE** of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

NER model: **en\_core\_web\_trf**

The new **Acropolis LOC** museum is **one CARDINAL** of the top museums that one can visit. The exhibits need no introduction, centuries of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

NER model: **en\_core\_web\_lg**

The new **Acropolis LOC** museum is one of the top museums that one can visit. The exhibits need no introduction, **centuries DATE** of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

NER model: **en\_core\_web\_sm\_entityRulerPipeline**

The new **Acropolis museum Museums History museums** is one of the top museums that one can visit. The exhibits need no introduction, centuries of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

NER model: **en\_core\_web\_md\_entityRulerPipeline**

The new **Acropolis museum Museums History museums** is one of the top museums that one can visit. The exhibits need no introduction, **centuries DATE** of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

NER model: **en\_core\_web\_trf\_entityRulerPipeline**

The new **Acropolis museum Museums History museums** is **one CARDINAL** of the top museums that one can visit. The exhibits need no introduction, centuries of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

NER model: **en\_core\_web\_lg\_entityRulerPipeline**

The new **Acropolis museum Museums History museums** is one of the top museums that one can visit. The exhibits need no introduction, **centuries DATE** of ancient **Greek NORP** history simply leave you speechless. The building itself is magnificent, flooded with dazzling **Greek NORP** light and of course the most beautiful view,...

Correct	<input type="text" value="0"/>	Incorrect	<input type="text" value="0"/>
Spurious	<input type="text" value="0"/>	Missing	<input type="text" value="0"/>

Submit

Figure 7. The web-based platform used to evaluate the NER models.

- *Precision* is defined as the fraction of correct named entities identified by the NER model, i.e., those matching the Golden Annotation Standard (GAS), over the total number of entities predicted by the model. The denominator consists of the sum of correct entities, incorrect entities (those not matching the GAS), and spurious entities (those falsely identified as matches). Formally, this can be expressed as:

$$Precision = \frac{Correct}{Correct + Incorrect + Spurious} = \frac{TP}{TP + FP}$$

Note that *TP* stands for True Positives, while *FP* stands for False Positives.

- *Recall* is defined as the fraction of correctly identified named entities relative to the total number of entities that should have been identified according to the GAS. The denominator includes the sum of correct entities, incorrect entities, and missing entities (i.e., entities not detected by the model). Formally, this can be expressed as:

$$Recall = \frac{Correct}{Correct + Incorrect + Missing} = \frac{TP}{TP + FN}$$

Note that *FN* stands for False Negatives.

- *F-Measure* is the weighted average of Precision and Recall, that takes into account both *FP* and *FN*. Formally, it can be expressed as:

$$F\_Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

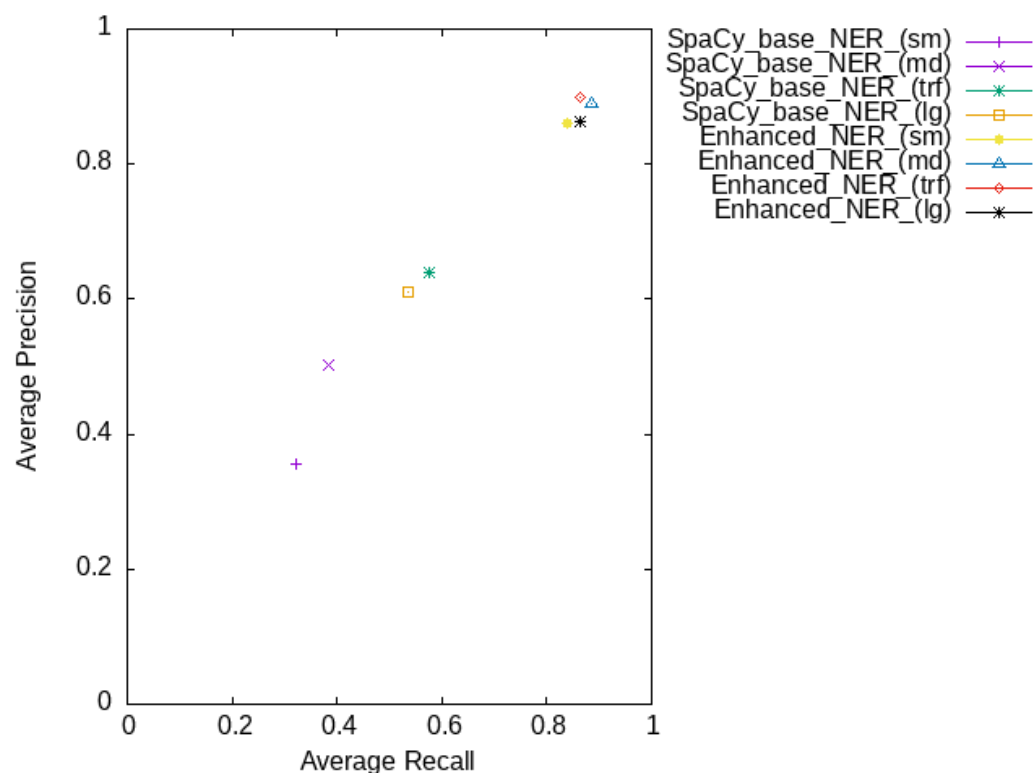
To determine which labeled entities were correct (i.e., matched the GAS exemplar), incorrect, spurious, or missing, we relied on a globally distributed group of 30 experts, who were assigned annotated data and asked to provide their assessments. The group of experts consisted of 10 professionals in the field of CI, three stakeholders from cultural and touristic venues, and 17 postgraduate students working/studying in the field of CI, capable of distinguishing which of the terms presented to them refer to cultural entities or have been falsely annotated by the corresponding NER model. To support this process, we developed a web-based evaluation platform using the Python micro-framework Flask [145]. The application retrieves random venue review texts from a MongoDB database, each annotated by four SpaCy language models and their enhanced versions. For every highlighted text, the interface displays four numerical fields corresponding to the counts of correct, incorrect, spurious, and missing entities, which are to be filled in by the evaluators. Once a submission is made, the respective text is marked as “reviewed” and the system fetches a new random review. Previously revised texts are excluded from subsequent iterations.

Over a period of nearly two months, we collected more than 3.5% of the annotated venue reviews dataset (approximately 10 K records) that provided a sufficient sample. As discussed previously, to determine which of the language models is best suited for identifying location entities, we conducted experiments based on the assessment of the domain experts. To start with, we compared the average precision of each language model against its average recall; next we compared the average F-measure of the models. Some results, where NER models were applied on TripAdvisor’s user reviews, are shown in Figures 8 and 9, whereas Table 9 summarizes the evaluation for all NER models.

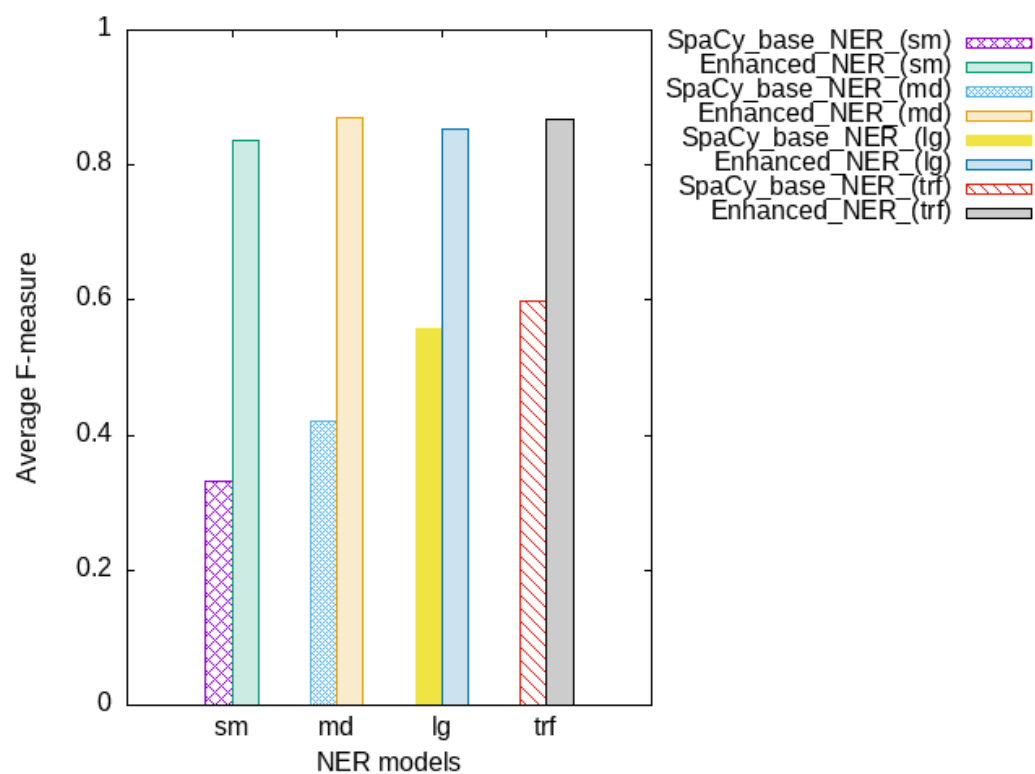
As shown in Figures 8 and 9 and Table 9, among the non-enhanced versions, the transformer-based model achieves the highest performance in all measures (precision, recall and F-measure), followed by the large model, while the medium and small models exhibit significantly lower performance. The enhanced versions (where the entity ruler pipeline is included) outperform their non-enhanced counterparts, with performance



gains ranging from 41% (for the precision of the transformer-based pipeline) to 161% (for the recall of the small-sized model).



**Figure 8.** Average precision versus average recall; NER models were applied on TripAdvisor's user reviews.



**Figure 9.** Average F-measure; NER models were applied on TripAdvisor's user reviews.

**Table 9.** NER models' evaluation scores.

NER Model	Precision	Recall	F-Measure
SpaCy_base_NER_(SM)	0.36	0.32	0.33
SpaCy_base_NER_(MD)	0.50	0.38	0.42
SpaCy_base_NER_(LG)	0.61	0.54	0.56
SpaCy_base_NER_(TRF)	0.64	0.58	0.60
Enhanced_NER_(SM)	0.86	0.84	0.84
Enhanced_NER_(MD)	0.89	0.88	0.87
Enhanced_NER_(LG)	0.86	0.86	0.85
Enhanced_NER_(TRF)	0.90	0.88	0.89

Among the enhanced versions of the language models, the enhanced transformer-based model exhibits the highest precision (approximately 0.90) and high recall (about 0.88), while the medium-sized enhanced model is the evaluation runner-up considering performance, by achieving precision equal to 0.89. The same ranking holds for the recall and F-measure metrics; more specifically, the transformer-based pipeline achieves recall and F-measure equal to 0.88 and 0.89, respectively, while the corresponding figures for the medium-sized model are 0.88 and 0.87.

To further evaluate the improved NER models relative to the native SpaCy models, we have generated the following accuracy metrics based on data collected from the web-based evaluation platform mentioned earlier (c.f. Figure 7).

- *Correctly identified* is defined as the fraction of correct named entities identified by the NER model (i.e., those matching the GAS), over the total number of entities that appear in the document and are either correctly detected, incorrectly detected, or not detected at all. The denominator consists of the sum of correct entities, incorrect entities (those not matching the GAS), and missing entities (those not detected by the model). Formally, this can be expressed as:

$$\text{Correctly identified} = \frac{\text{Correct}}{\text{All GAS entities (Correct + Incorrect + Missing)}}$$

- *Falsely identified* is defined as the fraction of the total named entities falsely recognized by the NER model (i.e., those not matching the GAS), over the total number of entities that appear in the document and are either correctly detected, incorrectly detected, or not detected at all. Typically, this can be formed as:

$$\text{Falsely identified} = \frac{\text{Incorrect + Spurious}}{\text{All GAS entities (Correct + Incorrect + Missing)}}$$

- *Not identified* is defined as the fraction of named entities that were not detected by the NER model, over the total number of entities that appear in the document and are either correctly detected, incorrectly detected, or not detected at all. Formally, this can be expressed as:

$$\text{Not identified} = \frac{\text{Missing}}{\text{All GAS entities (Correct + Incorrect + Missing)}}$$

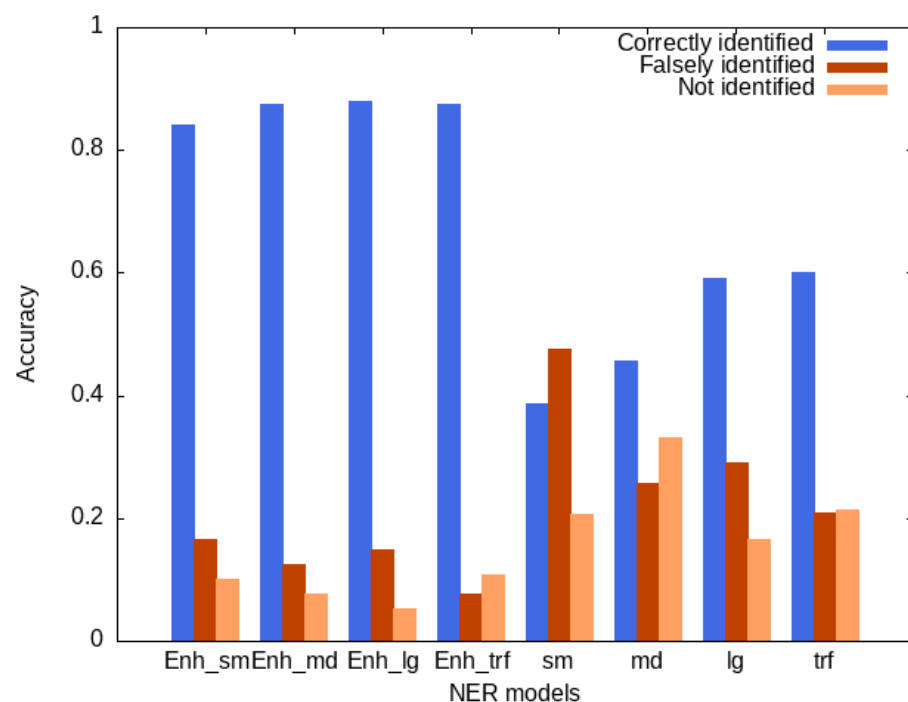
As shown in Table 10 and the corresponding accuracy plot in Figure 10, the enhanced NER models incorporating the Entity Ruler pipeline consistently outperform their baseline counterparts across all evaluated accuracy metrics (please note that in each column, the best performance value is marked in boldface). Specifically, the Enhanced\_NER\_(LG) model achieves the highest accuracy (88%) in correctly identifying named entities, demonstrating

its strong capacity for capturing a wide range of entity patterns. The Enhanced\_NER\_(MD) and Enhanced\_NER\_(TRF) follow closely, both attaining approximately 88% accuracy. In contrast, the Enhanced\_NER\_(SM) model shows greater precision in avoiding false positives, though at the expense of slightly reduced overall detection rates. It is also worth noting that the Enhanced\_NER\_(LG) model tends to overlook a small number of entities that appear infrequently within the text. Overall, the Enhanced\_NER\_(TRF) model exhibits the most balanced and consistent performance across all evaluation metrics, including Precision, Recall, and F1-score.

Observing the evaluation scores of each NER model, the results were as expected. The transformer-based model outperforms the other built-in SpaCy baselines in terms of accuracy, as transformers tend to be more robust than the simpler tok2vec-based components in the SM, MD, and LG pipelines that adopt the configuration of CNN models. Additionally, the SpaCy transformer model excels at capturing complex semantic relationships, surpassing the other models, particularly in tasks that require a deep contextual understanding, such as NER. It is worth noting that the entity ruler pipeline brought clear improvements to all models it was fitted to, still providing the best accuracy to the TRF model in terms of Precision, Recall, and F-measure.

**Table 10.** NER models' accuracy.

NER Model	Correctly Identified	Falsely Identified	Not Identified
SpaCy_base_NER_(SM)	38.82%	47.65%	20.59%
SpaCy_base_NER_(MD)	45.63%	25.63%	33.13%
SpaCy_base_NER_(LG)	59.17%	28.99%	16.57%
SpaCy_base_NER_(TRF)	60.12%	20.83%	21.43%
Enhanced_NER_(SM)	84.12%	16.47%	10.00%
Enhanced_NER_(MD)	87.5%	12.5%	7.71%
Enhanced_NER_(LG)	<b>88.02%</b>	14.97%	<b>5.39%</b>
Enhanced_NER_(TRF)	87.43%	<b>7.78%</b>	10.78%



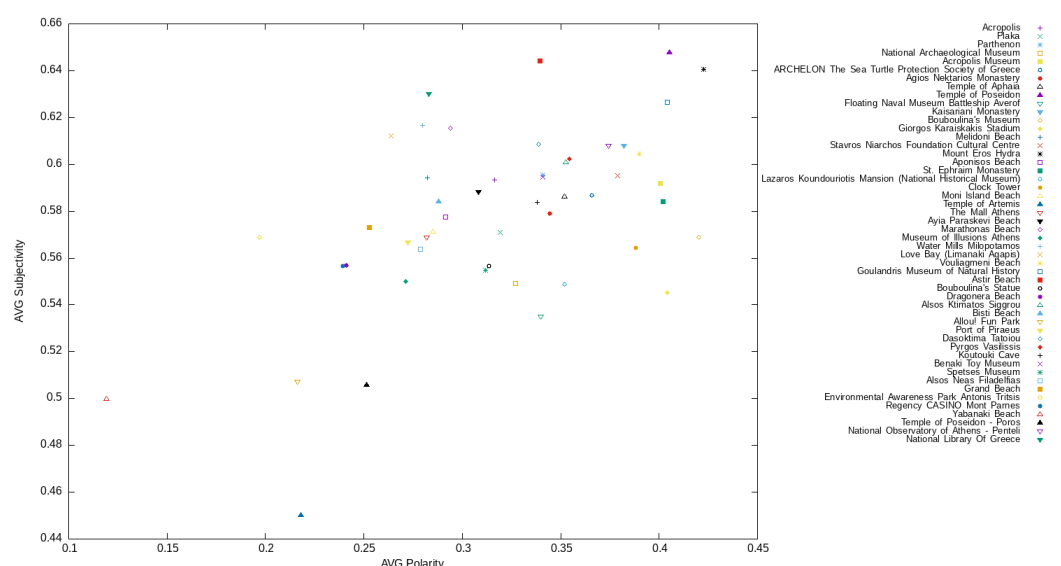
**Figure 10.** NER models' named entity identification accuracy.

### 5.3. Sentiment Analysis of User Reviews

Tourists often share their experiences about places they have visited through reviews or comments on social media platforms, where user sentiments are frequently implicit rather than explicitly stated [83,84]. Sentiment analysis is a subfield of NLP that helps analysts infer an author's mood and emotions from written text, thus extracting contextually meaningful insights. To uncover hidden emotions, we employed the SpacyTextBlob (version 4.0.0) [133] pipeline that leverages the TextBlob Python library (version 0.15.3). TextBlob provides a consistent API supporting tasks such as part-of-speech tagging, noun phrase extraction, and sentiment analysis.

In this framework, sentiment is determined by the semantic orientation of words and their contextual strength within a phrase. This requires a predefined dictionary, in which each term is classified as positive or negative, while sentences are treated as bags of words. The overall sentiment of a sentence is then derived by aggregating and averaging the individual scores assigned to each word. TextBlob outputs two metrics: polarity and subjectivity. Polarity measures sentiment orientation on a scale from  $-1$  to  $+1$ , where  $-1$  indicates strong negative emotions (e.g., anger, sadness, distrust) and  $+1$  indicates strong positive emotions (e.g., admiration, trust, love). Subjectivity measures the extent of personal opinion or experience versus the actual content of a sentence, ranging from 0 (objective) to 1 (highly subjective) [81].

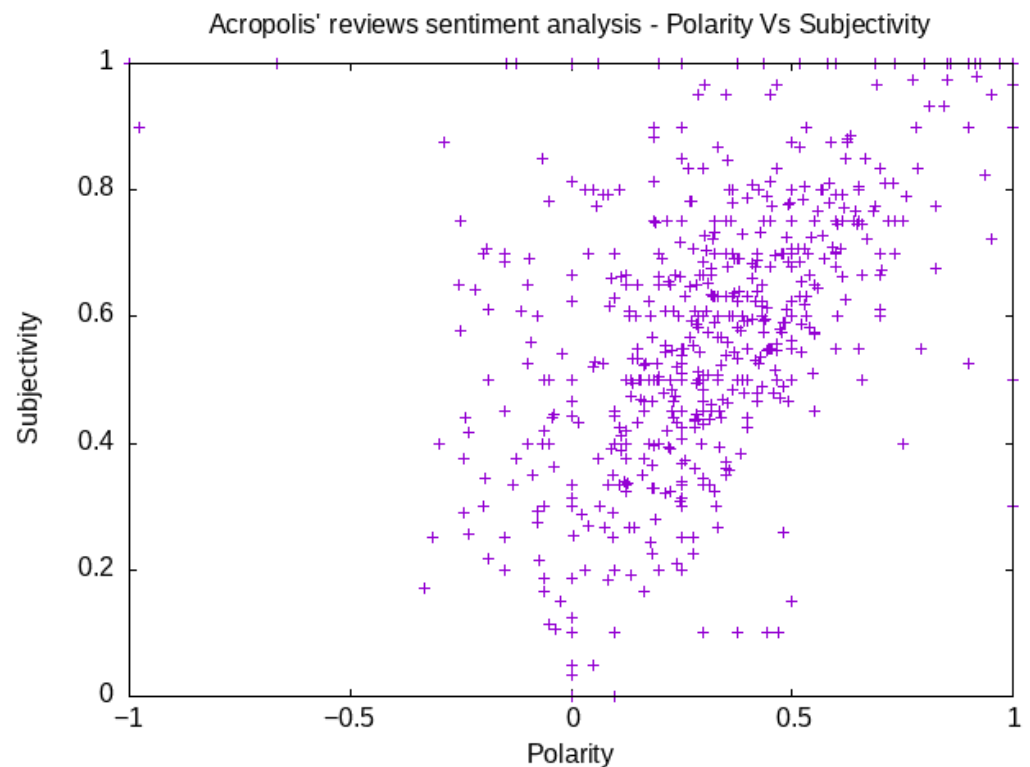
Through the application of sentiment analysis to our dataset of user comments and reviews extracted from social media platforms, we identified texts expressing both negative and positive emotions about PoIs. However, the majority of entries were classified as subjective by the sentiment analyzer, which is expected since social media users typically describe experiences from a personal perspective. Figure 11 illustrates the results of sentiment analysis on a sample of TripAdvisor user reviews. The plot shows the average polarity (X-axis) and average subjectivity (Y-axis) scores for reviews of 50 PoIs in the Attica region of Greece. Please note that high values of polarity combined with high values of subjectivity indicates personal enthusiasm rather than objective evaluation. Overall, most places were evaluated with mainly positive sentiments.



**Figure 11.** Average sentiment analysis derived from TripAdvisor user reviews for 50 PoIs located in the Attica region, Greece.

Figure 12 focuses on the Acropolis landmark in Athens and illustrates the distribution of user reviews according to their emotional orientation. The plot depicts polarity scores

on the X-axis and subjectivity scores on the Y-axis for 700 TripAdvisor reviews referring to Acropolis. As observed, most reviews express positive emotions toward this historic landmark and, as expected, the majority are characterized by a high degree of subjectivity.



**Figure 12.** TripAdvisor user reviews referencing the Acropolis landmark in Athens; points toward the upper-right corner indicate reviews that are both highly positive and strongly subjective.

#### 5.4. Recommendation of Sightseeing Trajectories

An important contribution of this research lies in its potential to enhance the tourism sector by generating attraction trajectories for tourists, using a crowdsourcing approach that derives recommendations from the collective experiences of past visitors. Specifically, we propose a procedure for extracting and recommending sightseeing routes to tourists in the Attica region of Greece, structured in four distinct phases.

##### 5.4.1. Phase 1: Automatic Translation of Multilingual Review Content

As mentioned in Section 5.1, in order to create a homogeneous dataset, we reinforced our method with the deep-translator component [134], which automatically detects the language of a given review text and translates it into English in real time. Homogenizing multilingual content renders the NER process easier, more efficient and accurate.

##### 5.4.2. Phase 2: Location Entity Extraction

Next, we applied the NER process to each review text using the SpaCy framework (Section 5.2). To optimize location entity identification, we extended SpaCy's pre-trained transformer language model with a custom entity ruler pipeline. For a sightseeing trajectory to be generated, at least three distinct location entities must be identified within a single review. Accordingly, reviews containing three or more location entities were flagged in the database.

#### 5.4.3. Phase 3: Sentiment Analysis of User Reviews Containing Routes

Once we have identified and flagged the review texts that entail trajectories, we apply sentiment analysis (Section 5.3). In this step, we keep the user reviews expressing positive sentiments, while excluding reviews with negative mood. Figure 13 provides an example of a review text expressing positive sentiments and containing more than three location entities (which are highlighted).

#### 5.4.4. Phase 4: Route Generation and Recommendation

Finally, after identifying the required location entities in emotionally positive texts, we used the Google Maps platform to generate sightseeing routes. According to the Google Maps documentation [146] and our analysis of the behavior of Google Maps, a route can be constructed by concatenating the base URL prefix <https://www.google.gr/maps/dir/> with a sequence of parameters. These parameters typically consist of location names separated by slashes (/), with multi-word location names being encoded according to the URL specification (notably, plus signs (+) are used in place of spaces). Additional information can also be included to refine the route, such as geographic coordinates (latitude and longitude), postal codes, or the region of a given route element. Figure 14 illustrates an itinerary generated from the review shown in Figure 13, where the extracted location entities were passed to the Google Maps Directions API.

The road to **Sounio GPE** is a unique idyllic experience and not only...! Picturesque landscape along the way with the sea never leaving the passenger seat for a moment... and challenging you every now and then to stop to admire the scenery! After **Varkiza GPE** , **Agia Marina GPE** , **Lagonisi GPE** , **Saronida GPE** and **Anavyssos GPE** , as you move away from civilization and the landscape becomes more and more desolate, you have the feeling that you are alone on earth... and when you see the columns of the **Temple of Poseidon Monuments and sights Ancient ruins** from afar, you feel like to call you... as if he has been waiting for you for **years DATE** ! And when you reach him, your soul is filled with awe and happiness...peace and fulfillment! You feel like you've arrived home!

**Figure 13.** A positive user review, containing more than three highlighted location entities.

However, the preliminary evaluation of our approach unveiled that some trajectories included location points that were geographically distant from each other, making them unsuitable for visitors to the Attica region of Greece. For instance, we encountered user reviews such as: *“An impeccable place, one thinks that he is not in Piraeus, but in a cafe in Paris or New York. Even the layout of the space gives this feeling. I especially liked that there is no fixed ceiling, as this space was a former courtyard.”*. To avoid creating routes spanning distant locations, limiting each route to a more confined geographical area, we enhanced our method with the Python Nominatim geocoder from the geopy module [147], which leverages the OpenStreetMap API [148]. By employing this library, we were able to retrieve the geographic coordinates of locations detected in user reviews, compute the distances between them, and filter out points exceeding a predefined radius threshold. The remaining locations are then used to generate coherent and practical tourist routes.

We evaluated the effect of different radius thresholds on the extracted tourist routes: (i) 0.05 (about 5.5 km distance), (ii) 0.1 (11 km), (iii) 0.15 (about 16.5 km), and (iv) 0.2 (22 km). Table 11 summarizes the resulting routes from the TripAdvisor review dataset after applying these distance-based thresholds. Out of 57,527 documents in the original corpus, 53,693 were classified as semantically positive, and 803 contained at least three location entities. Smaller radius thresholds resulted in more compact, locally coherent routes, while larger thresholds allowed inclusion of more distant points, increasing route



coverage. In this context, the selection of the radius threshold is inherently flexible and can be adjusted according to user preferences and travel conditions. For example, travelers with access to private transportation may opt for larger radius values, allowing them to explore points of interest located beyond the urban core, such as natural caves, coastal or freshwater ecosystems, and mountainous areas. Conversely, visitors with a stronger cultural focus, or those relying primarily on public transportation, may prefer smaller radius thresholds, enabling them to discover historical sites, ancient landmarks, and museums concentrated within the city center.



**Figure 14.** A tourist trajectory generated from the review in Figure 13.

**Table 11.** Number of routes extracted in respect to the radius threshold.

Geocoding Radius Threshold	# of Routes Extracted
0.05	263
0.1	287
0.15	297
0.2	304

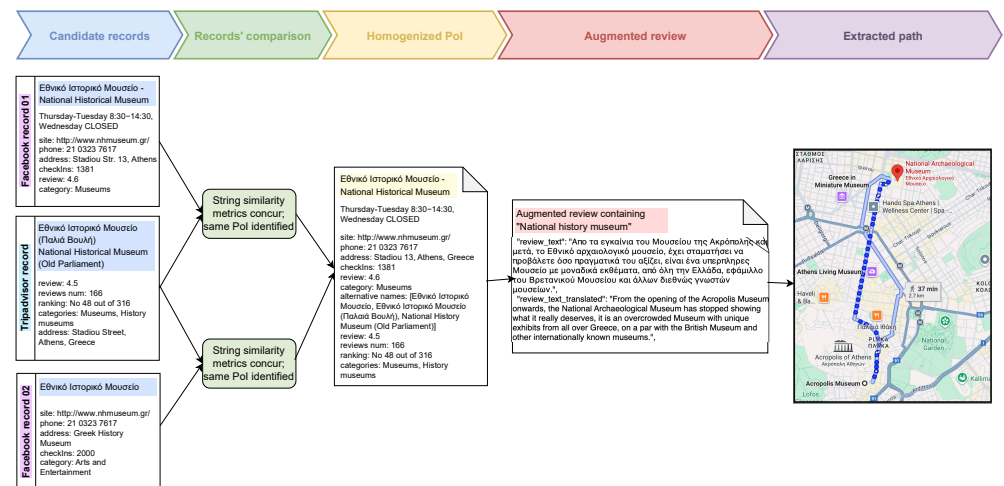
## 6. Discussion

In this section we provide an overview of the proposed methodology targeted to its reproducibility, comparison against other approaches, and framework applicability, while also commenting on the limitations of the proposed framework.

### 6.1. Data Workflow

The current research is extended in several phases, combining various cutting-edge technological solutions to eventually provide an innovative methodology, along with essential aspects that could prove valuable for the CH and Tourism communities. Overall, the proposed methodology consists of (i) harvesting cultural data from heterogeneous social media sources, (ii) harmonizing overlapping records referring to the same PoI, (iii) augmenting the collected data by leveraging NLP techniques for NER and sentiment

analysis tasks, and (iv) extracting cultural tourism trajectories from crowdsourced PoI reviews. To further clarify the suggested methodology, Figure 15 illustrates the process from raw sources to homogenized results and the eventual route extraction. Note that, for presentation purposes, Figure 15 depicts simplified JSON entries; non-essential fields have been omitted and only the key elements are shown to facilitate discussion and explanation of the process.



**Figure 15.** A running example illustrating sample data from all phases of the proposed framework; for clarity of presentation, JSON entries are simplified, while for readability purposes elements that originally appeared in Greek, such as addresses, have been translated into English.

As shown in Figure 15, from left to right, the first part presents the initial PoI data in JSON format. All three records correspond to the same entity (i.e., the National Historical Museum of Greece) and have passed the blocking phase (see Section 4.2.1) due to the proximity of their geographical coordinates. The first and third records were obtained from Facebook, while the second record was collected from TripAdvisor. The subsequent section illustrates the pairwise comparison of these records using string similarity metrics (see Section 4.2.2); in both cases, the comparisons confirm that the records refer to the same PoI based on their overall status field value. Next, the records are integrated into a single homogenized PoI record, with the `integrated_entities` field linking all three sources. The following phase depicts an augmented review for the Acropolis Museum, where the reviewer compares visits to two museums (i.e., the Acropolis and the National Historical Museum). The review is translated into English, annotated for entity recognition, and sentimentally analyzed, yielding a positive polarity score. The `location_list` field identifies all locations mentioned, while the `gmaps_route_link` field stores a Google Maps directions URL. Finally, the last phase illustrates a Google Maps trajectory generated from the locations in the review. This route allows travelers to visit both museums while experiencing the unique cultural landmarks along this compact yet culturally rich itinerary in central Athens.

## 6.2. Methodology Validation

Comparing our results with those reported in the literature, several observations can be drawn. Evaluating the effectiveness of our web scraping procedures against literature-based approaches is inherently challenging due to the contextual differences between studies. Literature reviews typically assess methods within a well-defined academic framework, whereas web scraping projects must contend with the dynamic and heterogeneous nature

of the web. Each project is defined by its specific combination of target websites and data points, making direct performance comparisons or generalization difficult.

Regarding the data homogenization methods reviewed in Section 2.2, we found considerable variation across existing strategies, primarily driven by the unique characteristics of each dataset. Similarly, the distinct nature of our PoI dataset necessitated a tailored approach. To the best of our knowledge, none of the related approaches follows a similar strategy. The method most closely related to ours is presented in [33]; however, our approach extends beyond this, by first reducing candidate pairs based on geographical coordinates before applying string similarity measures, resulting in a more computationally efficient process.

With respect to the NER task, our selection of SpaCy was guided by its open-source availability, robust models, and successful application across multiple domains. Nonetheless, when applied to our dataset, SpaCy's baseline models exhibited only moderate performance, likely due to their lack of specialization for cultural heritage and touristic content. By integrating rule-based entity recognition via the Entity Ruler pipeline with statistical or transformer-based Entity Recognizers from the baseline models, we developed enhanced NER pipelines that consistently outperformed SpaCy's defaults across all examined metrics (Precision, Recall, and F1-score, as well as Correctly, Falsely and Not identified), as shown in Figures 8–10, and summarized in Tables 9 and 10.

### 6.3. Limitations

The data collected and utilized in this study are limited to the Cultural Heritage and Tourism sectors, which are closely interconnected. This restricts the generalizability of the proposed framework to other interdisciplinary domains. Additionally, this work focuses on the acquisition of social media data for cultural and touristic points of interest within specific geographic areas (in this case, the Attica region in Greece). Consequently, the process involves a semi-supervised selection mechanism for the target areas and PoIs, as well as manual curation and inspection of the crawled data. Additionally, although the framework supports integrated data ingestion, allowing users to load custom datasets of any type, still the processing and extraction mechanisms remain primarily optimized for cultural and tourism-related content. These characteristics inherently limit the applicability of the proposed solution to the aforementioned domains. Despite these constraints, the architecture can be readily extended to related PoI-based applications, such as mobility, transport, or combined tourism-mobility analyses.

## 7. Conclusions

In this work, we presented a novel framework for collecting and analyzing Points of Interest (PoI) data. Our approach integrates automated social media web-scrapers and topic-focused crawlers, data homogenization, geolocation-based filtering to reduce duplicate candidates, and multiple toponym similarity measures to refine the dataset. Furthermore, we extract navigation trajectories by leveraging Machine Learning (ML) algorithms, NER language models, and sentiment analysis pipelines applied to crowdsourced user reviews and comments. These interconnected components offer a comprehensive toolkit that can serve as a valuable resource for organizations in the Cultural Heritage (CH) and tourism sectors, providing actionable insights from large volumes of user-generated content. To the best of our knowledge, this is the first framework in the cultural informatics domain to combine these cutting-edge technologies in a unified pipeline, producing results that exceed the capabilities of any single method alone.

Furthermore, we presented the architectural design supporting the proposed system, detailing the distinct yet interconnected modular services and the technological compo-

nents that enable their functionality. We described the organization of each unit within the framework and reported extensive, multi-faceted evaluations demonstrating the effectiveness of data homogenization. Additionally, we validated the robustness of the enhanced NER language models used to augment the dataset by uncovering additional location entities across user reviews, and we illustrated examples of sentiment analysis applied to reviews of PoIs in the Attica region of Greece.

It is worth emphasizing that the creation of personalized tourist navigation trajectories relies critically on the integration of all the aforementioned services, as each stage of data acquisition and processing is interdependent. Specifically, assembling information from multiple digital sources using automated tools is essential to generate a comprehensive PoI dataset. However, increasing the number of sources also raises the likelihood of duplicate records, particularly on social media platforms where crowdsourced content often results in multiple entries for the same location. Data homogenization plays a key role at this stage by reducing overlapping entries, incorporating complementary information, and providing a clean dataset for the subsequent NER annotation process. In addition to data purification, the dataset includes multilingual content, which must be standardized for consistent processing. The built-in translation component automatically detects the source language of each text and translates it into English. Finally, after applying the selected NER model to the homogenized, monolingual dataset, it is crucial to distinguish reviews expressing positive sentiments from others, enabling the recommendation of satisfying and meaningful tourist routes.

As for future research directions, we aim to explore alternative approaches for retrieving cultural information. One potential direction is to extend the Social Media Harvesting unit to incorporate additional social media platforms, either via web scraping or through available APIs, thereby enriching the dataset with more structured or unstructured content, such as images depicting the current state of cultural landmarks. Another strategy to enhance the performance, availability, and partition tolerance of the data lake, particularly under heavy workloads, is horizontal scaling. In this approach, we plan to leverage MongoDB native sharding methodology [149], distributing the system load across a cluster of servers, with each shard efficiently hosting a replica set of data. Furthermore, inspired by the recent advances in language models, we aim to fine-tune and integrate an LLM to provide advanced, prompt-driven support for data lake users. This integration is expected to enable tasks such as generating insights, personalized recommendations, and summaries across large datasets, thereby mitigating the challenges associated with navigating vast and heterogeneous data.

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

**Funding:** This work was supported in part by projects ENIRISST+ (grant agreement No. MIS 5047041) and TripMentor (grant agreement T1EDK-03874) from the General Secretariat for ERDF & CF, under Operational Programme Competitiveness, Entrepreneurship and Innovation 2014–2020 (EPAnEK) of the Greek Ministry of Economy and Development (co-financed by Greece and the EU through the European Regional Development Fund).

**Institutional Review Board Statement:** All the data used in this work are publicly accessible and anonymous. The web scraping of publicly accessible data does not violate the Computer Fraud and Abuse Act (CFAA).

**Informed Consent Statement:** All the data used in this work are publicly accessible and anonymous. Informed consent for participation is not required as per GDPR (recital 50; article 14 par. 5b).

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

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

## References

1. Ardito, L.; Cerchione, R.; Del Vecchio, P.; Raguseo, E. Big data in smart tourism: Challenges, issues and opportunities. *Curr. Issues Tour.* **2019**, *22*, 1805–1809. [CrossRef]
2. Xiang, Z.; Fesenmaier, D.R. Big data analytics, tourism design and smart tourism. In *Analytics in Smart Tourism Design: Concepts and Methods*; Springer: Cham, Switzerland, 2017; pp. 299–307.
3. Özemre, M.; Kabadurmus, O. A big data analytics based methodology for strategic decision making. *J. Enterp. Inf. Manag.* **2020**, *33*, 1467–1490. [CrossRef]
4. 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]
5. Vajjhala, N.R.; Strang, K.D. Measuring organizational-fit through socio-cultural big data. *New Math. Nat. Comput.* **2017**, *13*, 145–158. [CrossRef]
6. Lomborg, S.; Bechmann, A. Using APIs for Data Collection on Social Media. *Inf. Soc.* **2014**, *30*, 256–265. [CrossRef]
7. Murray-Rust, P. Open data in science. *Nat. Preced.* **2008**, *34*, 52–64 [CrossRef]
8. Couper, M.P. New developments in survey data collection. *Annu. Rev. Sociol.* **2017**, *43*, 121–145. [CrossRef]
9. Fuhr, N.; Tsakonas, G.; Aalberg, T.; Agosti, M.; Hansen, P.; Kapidakis, S.; Klas, C.; Kovács, L.; Landoni, M.; Micsik, A.; et al. Evaluation of digital libraries. *Int. J. Digit. Libr.* **2007**, *8*, 21–38. [CrossRef]
10. Khder, M.A. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *Int. J. Adv. Soft Comput. Its Appl.* **2021**, *13*, 145–168. [CrossRef]
11. Deligiannis, K.; Tryfonopoulos, C.; Raftopoulou, P.; Platis, N.; Vassilakis, C. EnQuest: A Cloud Datalake Infrastructure for Heterogeneous Analytics in Maritime and Tourism Domains. In Proceedings of the DATA 2023 (demo), Rome, Italy, 11–13 July 2023.
12. Barbera, G.; Araujo, L.; Fernandes, S.C. The Value of Web Data Scraping: An Application to TripAdvisor. *Big Data Cogn. Comput.* **2023**, *7*, 121. [CrossRef]
13. Deligiannis, K.; Raftopoulou, P.; Tryfonopoulos, C.; Platis, N.; Vassilakis, C. Hydria: An Online Data Lake for Multi-Faceted Analytics in the Cultural Heritage Domain. *Big Data Cogn. Comput.* **2020**, *4*, 7. [CrossRef]
14. Vonitsanos, G.; Kanavos, A.; Mohasseb, A.; Tsoilis, D. A NoSQL Approach for Aspect Mining of Cultural Heritage Streaming Data. In Proceedings of the 10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019, Patras, Greece, 15–17 July 2019; Bourbakis, N.G., Tsihrintzis, G.A., Virvou, M., Eds.; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4. [CrossRef]
15. Freire, N.; Silva, M.J. Domain-Focused Linked Data Crawling Driven by a Semantically Defined Frontier—A Cultural Heritage Case Study in Europeana. In Proceedings of the Digital Libraries at Times of Massive Societal Transition—22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, Kyoto, Japan, 30 November–1 December 2020; Ishita, E., Pang, N.L., Zhou, L., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12504, pp. 340–348. [CrossRef]
16. Wang, W.; Yu, L. UCrawler: A learning-based web crawler using a URL knowledge base. *J. Comput. Methods Sci. Eng.* **2021**, *21*, 461–474. [CrossRef]
17. Wang, K.; Jalal, M.; Jefferson, S.; Zheng, Y.; Nsoesie, E.O.; Betke, M. Scraping Social Media Photos Posted in Kenya and Elsewhere to Detect and Analyze Food Types. *arXiv* **2019**, arXiv:1909.00134. Available online: <http://arxiv.org/abs/1909.00134> (accessed on 13 November 2025). [CrossRef]
18. Autelitano, A.; Pernici, B.; Scalia, G. Spatio-temporal mining of keywords for social media cross-social crawling of emergency events. *GeoInformatica* **2019**, *23*, 425–447. [CrossRef]
19. AlZu'bi, S.; Aqel, D.; Mughaid, A.; Jararweh, Y. A Multi-Levels Geo-Location Based Crawling Method for Social Media Platforms. In Proceedings of the Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS 2019, Granada, Spain, 22–25 October 2019; Alsmirat, M.A., Jararweh, Y., Eds.; IEEE: Piscataway, NJ, USA, 2019; pp. 494–498. [CrossRef]
20. Xu, K.; Gao, K.Y.; Callan, J. A Structure-Oriented Unsupervised Crawling Strategy for Social Media Sites. *arXiv* **2018**, arXiv:1804.02734. Available online: <http://arxiv.org/abs/1804.02734> (accessed on 13 November 2025). [CrossRef]
21. Erlandsson, F.; Bródka, P.; Boldt, M.; Johnson, H. Do We Really Need to Catch Them All? A New User-Guided Social Media Crawling Method. *Entropy* **2017**, *19*, 686. [CrossRef]
22. Wani, M.A.; Agarwal, N.; Jabin, S.; Hussai, S.Z. Design of iMacros-Based Data Crawler and the Behavioral Analysis of Facebook Users. *arXiv* **2018**, arXiv:1802.09566.



23. Seo, M.; Kim, J.; Yang, H. Frequent Interaction and Fast Feedback Predict Perceived Social Support: Using Crawled and Self-Reported Data of Facebook Users. *J. Comput. Mediat. Commun.* **2016**, *21*, 282–297. [\[CrossRef\]](#)
24. Hernandez-Suarez, A.; Sanchez-Perez, G.; Toscano-Medina, K.; Martinez-Hernandez, V.; Sanchez, V.; Pérez-Meana, H. A Web Scraping Methodology for Bypassing Twitter API Restrictions. *arXiv* **2018**, arXiv:1803.09875. Available online: <http://arxiv.org/abs/1803.09875> (accessed on 13 November 2025). [\[CrossRef\]](#)
25. El Akbar, R.R.; Shofa, R.N.; Paripurna, M.I.; Supratman. The implementation of Naïve Bayes algorithm for classifying tweets containing hate speech with political motive. In Proceedings of the 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC), Bandung, Indonesia, 20–22 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 144–148.
26. Xiang, Z.; Du, Q.; Ma, Y.; Fan, W. Assessing reliability of social media data: Lessons from mining TripAdvisor hotel reviews. *J. Inf. Technol. Tour.* **2018**, *18*, 43–59. [\[CrossRef\]](#)
27. Li, J.; Yang, L. A Rule-Based Chinese Sentiment Mining System with Self-Expanding Dictionary—Taking TripAdvisor as an Example. In Proceedings of the 14th IEEE International Conference on e-Business Engineering, ICEBE 2017, Shanghai, China, 4–6 November 2017; Hussain, O., Jiang, L., Fei, X., Lan, C., Chao, K., Eds.; IEEE Computer Society: Piscataway, NJ, USA, 2017; pp. 238–242. [\[CrossRef\]](#)
28. Xhumari, E.; Xhumari, I. A review of web crawling approaches. In Proceedings of the 4th International Conference on Recent Trends and Applications in Computer Science and Information Technology, Tirana, Albania, 21–22 May 2021; Volume 2872, pp. 158–163.
29. Gupta, A.; Anand, P. Focused web crawlers and its approaches. In Proceedings of the 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 25–27 February 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 619–622.
30. Saini, C.; Arora, V. Information retrieval in web crawling: A survey. In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, Jaipur, India, 21–24 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2635–2643. [\[CrossRef\]](#)
31. Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 1–16. [\[CrossRef\]](#)
32. Martins, B. A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records. In Proceedings of the GeoSpatial Semantics—4th International Conference, GeoS 2011, Brest, France, 12–13 May 2011; Claramunt, C., Levashkin, S., Bertolotto, M., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6631, pp. 34–51. [\[CrossRef\]](#)
33. Santos, R.; Murrieta-Flores, P.; Martins, B. Learning to combine multiple string similarity metrics for effective toponym matching. *Int. J. Digit. Earth* **2018**, *11*, 913–938. [\[CrossRef\]](#)
34. Long, Y.; Li, H.; Wan, Z.; Tian, P. Data Redundancy Detection Algorithm based on Multidimensional Similarity. In Proceedings of the 2023 International Conference on Frontiers of Robotics and Software Engineering (FRSE), Changsha, China, 16–18 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 180–187.
35. Omar, Z.A.; Abu Bakar, M.A.; Zamzuri, Z.H.; Ariff, N.M. Duplicate Detection Using Unsupervised Random Forests: A Preliminary Analysis. In Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), Ipoh, Malaysia, 7–8 September 2022; pp. 66–71. [\[CrossRef\]](#)
36. Lin, Y.; Wang, H.; Li, J.; Gao, H. Efficient Entity Resolution on Heterogeneous Records. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 912–926. [\[CrossRef\]](#)
37. Zhang, D.; Li, D.; Guo, L.; Tan, K.L. Unsupervised Entity Resolution with Blocking and Graph Algorithms. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 1501–1515. [\[CrossRef\]](#)
38. Cao, K.; Liu, H. Entity Resolution Algorithm for Heterogeneous Data Sources. In Proceedings of the 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI), Kunming, China, 17–19 September 2021; pp. 553–557. [\[CrossRef\]](#)
39. Karapiperis, D.; Gkoulalas-Divanis, A.; Verykios, V.S. MultiBlock: A Scalable Iterative Approach for Progressive Entity Resolution. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021; pp. 219–228. [\[CrossRef\]](#)
40. Park, S.; Lee, S.; Woo, S.S. BertLoc: Duplicate location record detection in a large-scale location dataset. In Proceedings of the 36th Annual ACM Symposium on Applied Computing, SAC '21, New York, NY, USA, 22–26 March 2021; pp. 942–951. [\[CrossRef\]](#)
41. Gu, Q.; Dong, Y.; Hu, Y.; Liu, Y. A Method for Duplicate Record Detection Using Deep Learning. In Proceedings of the Web Information Systems and Applications—16th International Conference, WISA 2019, Qingdao, China, 20–22 September 2019; Ni, W., Wang, X., Song, W., Li, Y., Eds.; Springer: Cham, Switzerland, 2019; pp. 85–91. [\[CrossRef\]](#)
42. Lattar, H.; Ben Salem, A.; Ben Ghezala, H.H. Duplicate record detection approach based on sentence embeddings. In Proceedings of the 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Virtual, 10–13 September 2020; pp. 269–274. [\[CrossRef\]](#)



43. Ziv, R.; Gronau, I.; Fire, M. CompanyName2Vec: Company Entity Matching based on Job Ads. In Proceedings of the 9th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2022, Shenzhen, China, 13–16 October 2022; Huang, J.Z., Pan, Y.; Hammer, B., Khan, M.K., Xie, X., Cui, L., He, Y., Eds.; IEEE: Piscataway, NJ, USA, 2022; pp. 1–10. [\[CrossRef\]](#)
44. Zhang, P. Similar Duplicate Record Detection of Big Data Based on Entropy Grouping Clustering. In Proceedings of the 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 22–24 April 2022; pp. 646–650. [\[CrossRef\]](#)
45. Abbes, H.; Gargouri, F. MongoDB-Based Modular Ontology Building for Big Data Integration. *J. Data Semant.* **2018**, *7*, 1–27. [\[CrossRef\]](#)
46. Song, R.; Yu, T.; Chen, Y.; Chen, Y.; Xia, B. An Approximately Duplicate Records Detection Method for Electric Power Big Data Based on Spark and IPOP-Simhash. *J. Inf. Hiding Multim. Signal Process.* **2018**, *9*, 410–422.
47. Qi, Y.; Ren, W.; Shi, M.; Liu, Q. A Combinatorial Method based on Machine Learning Algorithms for Enhancing Cultural Economic Value. *Int. J. Perform. Eng.* **2020**, *16*, 1105–1117. [\[CrossRef\]](#)
48. Penagos-Londoño, G.I.; Rodriguez-Sanchez, C.; Ruiz-Moreno, F.; Torres, E. A machine learning approach to segmentation of tourists based on perceived destination sustainability and trustworthiness. *J. Destin. Mark. Manag.* **2021**, *19*, 100532. [\[CrossRef\]](#)
49. Kar, A.K.; Choudhary, S.K.; Ilavarasan, P.V. How can we improve tourism service experiences: Insights from multi-stakeholders' interaction. *Decision* **2023**, *50*, 73–89. [\[CrossRef\]](#)
50. Tae, K.H.; Roh, Y.; Oh, Y.H.; Kim, H.; Whang, S.E. Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning, Amsterdam, The Netherlands, 30 June 2019; pp. 1–4.
51. Fu, Y.; Shi, K.; Xi, L. Artificial intelligence and machine learning in the preservation and innovation of intangible cultural heritage: Ethical considerations and design frameworks. *Digit. Scholarsh. Humanit.* **2025**, *40*, 487–508. [\[CrossRef\]](#)
52. de la Rosa, J. Machine learning at the National Library of Norway. In *Navigating Artificial Intelligence for Cultural Heritage Organisations*; Jaillant, L., Warwick, C., Gooding, P., Aske, K., Layne-Worthey, G., Downie, J.S., Eds.; UCL Press: London, UK, 2025; pp. 61–90. [\[CrossRef\]](#)
53. Sousa, J.J.; Lin, J.; Wang, Q.; Liu, G.; Fan, J.; Bai, S.; Zhao, H.; Pan, H.; Wei, W.; Rittlinger, V.; et al. Using machine learning and satellite data from multiple sources to analyze mining, water management, and preservation of cultural heritage. *Geo Spat. Inf. Sci.* **2024**, *27*, 552–571. [\[CrossRef\]](#)
54. Nadkarni, P.M.; Ohno-Machado, L.; Chapman, W.W. Natural language processing: An introduction. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 544–551. [\[CrossRef\]](#)
55. Schmitt, X.; Kubler, S.; Robert, J.; Papadakis, M.; LeTraon, Y. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 338–343.
56. Miao, Y.; Jin, Z.; Zhang, Y.; Chen, Y.; Lai, J. Compare Machine Learning Models in Text Classification Using Steam User Reviews. In Proceedings of the 2021 3rd International Conference on Software Engineering and Development (ICSSED), Xiamen, China, 19–21 November 2021; pp. 40–45.
57. Kadam, K.; Godbole, S.; Joiode, D.; Karoshi, S.; Jadhav, P.; Shilaskar, S. Multilingual Information Retrieval Chatbot. In *Modern Approaches in Machine Learning & Cognitive Science: A Walkthrough*; Studies in Computational Intelligence; Springer: Cham, Switzerland, 2022; Volume 1027, pp. 107–121. [\[CrossRef\]](#)
58. Ferreira, J.; Gonalo Oliveira, H.; Rodrigues, R. Improving NLTK for processing Portuguese. In Proceedings of the 8th Symposium on Languages, Applications and Technologies (SLATE 2019), Coimbra, Portugal, 27–28 June 2019; Schloss Dagstuhl-Leibniz-Zentrum für Informatik: Wadern, Germany, 2019. [\[CrossRef\]](#)
59. Kapan, A.; Kirmizialtin, S.; Kukreja, R.; Wrisley, D.J. Fine Tuning NER with spaCy for Transliterated Entities Found in Digital Collections from the Multilingual Arabian/Persian Gulf. In Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, 15–18 March 2022.
60. Rouhou, A.C.; Dhiaf, M.; Kessentini, Y.; Salem, S.B. Transformer-based approach for joint handwriting and named entity recognition in historical document. *Pattern Recognit. Lett.* **2022**, *155*, 128–134. [\[CrossRef\]](#)
61. Hanks, C.; Maiden, M.; Ranade, P.; Finin, T.; Joshi, A. CyberEnt: Extracting Domain Specific Entities from Cybersecurity Text. In Proceedings of the Mid-Atlantic Student Colloquium on Speech, Language and Learning, Philadelphia, PA, USA, 30 April 2022.
62. Alam, T.; Bhusal, D.; Park, Y.; Rastogi, N. Cyner: A python library for cybersecurity named entity recognition. *arXiv* **2022**, arXiv:2204.05754; [\[CrossRef\]](#)
63. Koloveas, P.; Chantzios, T.; Alevizopoulou, S.; Skiadopoulos, S.; Tryfonopoulos, C. Intime: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence. *Electronics* **2021**, *10*, 818. [\[CrossRef\]](#)
64. De Magistris, G.; Russo, S.; Roma, P.; Starczewski, J.T.; Napoli, C. An explainable fake news detector based on named entity recognition and stance classification applied to COVID-19. *Information* **2022**, *13*, 137. [\[CrossRef\]](#)

65. ElDin, H.G.; AbdulRazek, M.; Abdelshafi, M.; Sahlol, A.T. Med-Flair: Medical named entity recognition for diseases and medications based on Flair embedding. *Procedia Comput. Sci.* **2021**, *189*, 67–75. [CrossRef]
66. Patel, H. Bionerflair: Biomedical named entity recognition using flair embedding and sequence tagger. *arXiv* **2020**, arXiv:2011.01504.
67. Kim, H.; Kang, J. How Do Your Biomedical Named Entity Recognition Models Generalize to Novel Entities? *IEEE Access* **2022**, *10*, 31513–31523. [CrossRef] [PubMed]
68. Chai, Z.; Jin, H.; Shi, S.; Zhan, S.; Zhuo, L.; Yang, Y. Hierarchical shared transfer learning for biomedical named entity recognition. *BMC Bioinform.* **2022**, *23*, 8. [CrossRef] [PubMed]
69. Uronen, L.; Salanterä, S.; Hakala, K.; Hartiala, J.; Moen, H. Combining supervised and unsupervised named entity recognition to detect psychosocial risk factors in occupational health checks. *Int. J. Med Inform.* **2022**, *160*, 104695. [CrossRef]
70. Tambuscio, M.; Andrews, T.L. Geolocation and Named Entity Recognition in Ancient Texts: A Case Study about Ghewond's Armenian History. In Proceedings of the CHR, Online, 17–19 November 2021; pp. 136–148.
71. Milanova, I.; Silc, J.; Serucnik, M.; Eftimov, T.; Gjoreski, H. LOCALE: A Rule-based Location Named-entity Recognition Method for Latin Text. In Proceedings of the 5th International Workshop on Computational History, HistoInformatics@TPDL 2019, Oslo, Norway, 12 September 2019; Volume 2461, pp. 13–20.
72. Molina-Villegas, A.; Muñoz-Sanchez, V.; Arreola-Trapala, J.; Alcántara, F. Geographic Named Entity Recognition and Disambiguation in Mexican News using word embeddings. *Expert Syst. Appl.* **2021**, *176*, 114855. [CrossRef]
73. Koirala, S. Event Discovery from Social Media Feeds. 2021. Available online: <https://urn.fi/URN:NBN:fi:aalto-2021121910935> (accessed on 13 November 2025).
74. Finin, T.; Murnane, W.; Karandikar, A.; Keller, N.; Martineau, J.; Dredze, M. Annotating named entities in Twitter data with crowdsourcing. In Proceedings of the NAACL Workshop on Creating Speech and Text Language Data with Amazon's Mechanical Turk, Los Angeles, CA, USA, 6 June 2010.
75. Liu, L.; Wang, M.; Zhang, M.; Qing, L.; He, X. UAMNer: Uncertainty-aware multimodal named entity recognition in social media posts. *Appl. Intell.* **2022**, *52*, 4109–4125. [CrossRef]
76. Asgari-Chenaghlu, M.; Feizi-Derakhshi, M.R.; Farzinvash, L.; Balafar, M.; Motamed, C. CWI: A multimodal deep learning approach for named entity recognition from social media using character, word and image features. *Neural Comput. Appl.* **2022**, *34*, 1905–1922. [CrossRef]
77. Eligüz, N.; Çetinkaya, C.; Dereli, T. Application of named entity recognition on tweets during earthquake disaster: A deep learning-based approach. *Soft Comput.* **2022**, *26*, 395–421. [CrossRef]
78. Egger, R.; Gokce, E. Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*; Springer International Publishing: Cham, Switzerland, 2022; pp. 307–334.
79. Bouabdallaoui, I.; Guerouate, F.; Bouhaddour, S.; Saadi, C.; Sbihi, M. Named Entity Recognition applied on Moroccan tourism corpus. *Procedia Comput. Sci.* **2022**, *198*, 373–378. [CrossRef]
80. Chantrapornchai, C.; Tunsakul, A. Information extraction on tourism domain using SpaCy and BERT. *ECTI Trans. Comput. Inform. Technol.* **2021**, *15*, 108–122.
81. Montoyo, A.; Martínez-Barco, P.; Balahur, A. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decis. Support Syst.* **2012**, *53*, 675–679. [CrossRef]
82. Kauffmann, E.; Peral, J.; Gil, D.; Ferrández, A.; Sellers, R.; Mora, H. Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. *Sustainability* **2019**, *11*, 4235. [CrossRef]
83. Ahuja, S.; Dubey, G. Clustering and sentiment analysis on Twitter data. In Proceedings of the 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), Noida, India, 10–11 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
84. Bravo-Marquez, F.; Mendoza, M.; Poblete, B. Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013, Chicago, IL, USA, 11 August 2013*; Cambria, E., Liu, B., Zhang, Y., Xia, Y., Eds.; ACM: New York, NY, USA, 2013; pp. 1–9. [CrossRef]
85. Subbaiah, B.; Murugesan, K.; Saravanan, P.; Marudhamuthu, K. An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. *Artif. Intell. Rev.* **2024**, *57*, 34. [CrossRef]
86. Thareja, R. Multimodal Sentiment Analysis of Social Media Content and Its Impact on Mental Wellbeing: An Investigation of Extreme Sentiments. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, Bangalore, India, 4–7 January 2024; Natarajan, S., Bhattacharya, I., Singh, R., Kumar, A., Ranu, S., Bali, K., K, A., Eds.; ACM: New York, NY, USA, 2024; pp. 469–473. [CrossRef]
87. Rodríguez-Ibáñez, M.; Casañez-Ventura, A.; Castejón-Mateos, F.; Cuenca-Jiménez, P.M. A review on sentiment analysis from social media platforms. *Expert Syst. Appl.* **2023**, *223*, 119862. [CrossRef]

88. Cuzzocrea, A. Big data lakes: Models, frameworks, and techniques. In Proceedings of the 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju Island, Republic of Korea, 17–20 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4.
89. Facebook. Available online: <https://www.facebook.com/> (accessed on 18 October 2025).
90. TripAdvisor. Available online: <https://www.tripadvisor.com/> (accessed on 18 October 2025).
91. Google Maps. Available online: <https://www.google.com/maps> (accessed on 18 October 2025).
92. Odysseus Portal—Ministry of Culture and Sports. Available online: [http://odysseus.culture.gr/index\\_en.html](http://odysseus.culture.gr/index_en.html) (accessed on 18 October 2025).
93. Search Culture—Culture in the Digital Public Space. Available online: <https://www.searchculture.gr/aggregator/portal/?language=en> (accessed on 18 October 2025).
94. Portal of Cultural Stakeholders. Available online: <https://portal.culture.gov.gr/> (accessed on 18 October 2025).
95. Athens Culture Net. Available online: <https://www.cityofathens.gr/who/athens-culture-net/> (accessed on 13 October 2025).
96. X Microblogging Platform. Available online: <https://x.com/> (accessed on 18 October 2025).
97. Hridoy, M.T.A.; Saha, S.R.; Islam, M.M.; Uddin, M.A.; Mahmud, M.Z. Leveraging web scraping and stacking ensemble machine learning techniques to enhance detection of major depressive disorder from social media posts. *Soc. Netw. Anal. Min.* **2024**, *14*, 239. [\[CrossRef\]](#)
98. Reddit Platform. Available online: <https://www.reddit.com/> (accessed on 18 October 2025).
99. Bouabdelli, L.F.; Abdelhedi, F.; Hammoudi, S.; Hadjali, A. An Advanced Entity Resolution in Data Lakes: First Steps. In Proceedings of the 14th International Conference on Data Science, Technology and Applications, DATA 2025, Bilbao, Spain, 10–12 June 2025; pp. 661–668. [\[CrossRef\]](#)
100. Liu, C.; Rong, X. Automated Graph Attention Network for Heterogeneous Entity Resolution. In Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, 6–11 April 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5. [\[CrossRef\]](#)
101. Levchenko, M. Evaluating Named Entity Recognition Models for Russian Cultural News Texts: From BERT to LLM. *arXiv* **2025**, arXiv:2506.02589. [\[CrossRef\]](#)
102. Li, Y.; Yan, H.; Yang, Y.; Wang, X. A Method for Cultural Relics Named Entity Recognition Based on Enhanced Lexical Features. In Proceedings of the International Joint Conference on Neural Networks, IJCNN 2024, Yokohama, Japan, 30 June–5 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8. [\[CrossRef\]](#)
103. SpaCy, Industrial-Strength Natural Language Processing. Available online: <https://spacy.io/> (accessed on 18 October 2025).
104. Kumar, D.; Pandey, S.; Patel, P.; Choudhari, K.; Hajare, A.; Jante, S. Generalized Named Entity Recognition Framework. In Proceedings of the 2021 Asian Conference on Innovation in Technology (ASIANCON), Pune, India, 27–29 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–4.
105. Ferro, S.; Giovanelli, R.; Leeson, M.; Bernardin, M.D.; Traviglia, A. A novel NLP-driven approach for enriching artefact descriptions, provenance, and entities in cultural heritage. *Neural Comput. Appl.* **2025**, *37*, 21275–21296. [\[CrossRef\]](#)
106. Zheng, Y.; Li, F.; Li, C.; Zhang, Z.; Cao, R.; Noman, S.M. A Natural Language Processing Model for Automated Organization and Analysis of Intangible Cultural Heritage. *J. Organ. End User Comput.* **2024**, *36*, 1–27. [\[CrossRef\]](#)
107. Loper, E.; Bird, S. Nltk: The natural language toolkit. *arXiv* **2002**, arXiv:cs/0205028. [\[CrossRef\]](#)
108. Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; Vollgraf, R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MN, USA, 2–7 June 2019; pp. 54–59.
109. Čević, H.; Kurdija, A.S.; Delač, G.; Šilić, M. Named Entity Recognition for Addresses: An Empirical Study. *IEEE Access* **2022**, *10*, 42108–42120. [\[CrossRef\]](#)
110. Yu, J.; Ji, B.; Li, S.; Ma, J.; Liu, H.; Xu, H. S-NER: A Concise and Efficient Span-Based Model for Named Entity Recognition. *Sensors* **2022**, *22*, 2852. [\[CrossRef\]](#)
111. Zhao, X.; Greenberg, J.; An, Y.; Hu, X.T. Fine-Tuning BERT Model for Materials Named Entity Recognition. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December; IEEE: Piscataway, NJ, USA, 2021; pp. 3717–3720.
112. Krovetz, R.; Deane, P.; Madnani, N. The Web is not a PERSON, Berners-Lee is not an ORGANIZATION, and African-Americans are not LOCATIONS: An Analysis of the Performance of Named-Entity Recognition. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World, MWE@ACL 2011, Portland, OR, USA, 23 June 2011*; Kordoni, V., Ramisch, C., Villavicencio, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 57–64.
113. Mehrabi, N.; Gowda, T.; Morstatter, F.; Peng, N.; Galstyan, A. Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition. In *Proceedings of the HT '20: 31st ACM Conference on Hypertext and Social Media, Virtual Event, 13–15 July 2020*; Gadiraju, U., Ed.; ACM: New York, NY, USA, 2020; pp. 231–232. [\[CrossRef\]](#)
114. Myers, D.; McGuffee, J.W. Choosing scrapy. *J. Comput. Sci. Coll.* **2015**, *31*, 83–89.

115. Hoffman, P.; Grana, D.; Olveyra, M.; Garcia, G.; Cetrulo, M.; Bogomyagkov, A.; Canabal, D.; Moreira, A.; Carnales, I.; Aguirre, M.; et al. Scrapy at a Glance. Available online: <https://docs.scrapy.org/en/latest/intro/overview.html> (accessed on 18 October 2025).
116. Vassilakis, C.; Pouloupoulos, V.; Wallace, M.; Antoniou, A.; Lepouras, G. TripMentor Project: Scope and Challenges. In Proceedings of the Workshop on Cultural Informatics Co-Located with the 14th International Workshop on Semantic and Social Media Adaptation and Personalization, CI@SMAP 2019, Larnaca, Cyprus, 9 June 2019; Volume 2412.
117. Mehta, H.; Kanani, P.; Lande, P. Google maps. *Int. J. Comput. Appl* **2019**, *178*, 41–46. [CrossRef]
118. Selenium Web Driver. Available online: <https://www.selenium.dev/documentation/webdriver/> (accessed on 18 October 2025).
119. Peng, J.; Ma, Y.; Zhou, F.-r.; Wang, S.-l.; Zheng, Z.-z.; Li, J. Web Crawler of Power Grid Based on Selenium. In Proceedings of the 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 13–15 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 114–118.
120. USA: Web Scraping Held to Be Legal in Lawsuit Brought by LinkedIn over Privacy Concerns. Available online: <https://www.business-humanrights.org/en/latest-news/usa-web-scraping-held-to-be-legal-in-lawsuit-brought-by-linkedin-over-privacy-concerns/> (accessed on 18 October 2025).
121. Google Catches Bing Copying; Microsoft Says ‘So What?’. Available online: <https://www.wired.com/2011/02/bing-copies-google/> (accessed on 18 October 2025).
122. Parliament, E.; The Council of the European Union. REGULATION (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679> (accessed on 18 October 2025).
123. Yelp Open Dataset. Available online: <https://business.yelp.com/data/resources/open-dataset/> (accessed on 18 October 2025).
124. European Data—Open Data and Tourism. Available online: <https://data.europa.eu/en/publications/datastories/open-data-and-tourism> (accessed on 18 October 2025).
125. ACHE Crawler Documentation. Available online: <https://ache.readthedocs.io/en/latest/> (accessed on 18 October 2025).
126. Vieira, K.; Barbosa, L.; da Silva, A.S.; Freire, J.; de Moura, E.S. Finding seeds to bootstrap focused crawlers. *World Wide Web* **2016**, *19*, 449–474. [CrossRef]
127. Deligiannis, K.; Raftopoulou, P.; Tryfonopoulos, C.; Vassilakis, C. A System for Collecting, Managing, Analyzing and Sharing Diverse, Multi-Faceted Cultural Heritage and Tourism Data. In Proceedings of the 16th International Workshop on Semantic and Social Media Adaptation & Personalization, SMAP 2021, Corfu, Greece, 4–5 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8. [CrossRef]
128. Hurwit, J.M. *The Athenian Acropolis: History, Mythology, and Archaeology from the Neolithic Era to the Present*; Cambridge University Press: Cambridge, UK, 2000.
129. Androutsopoulos, J. ‘Greeklish’: Transliteration practice and discourse in the context of computer-mediated digraphia. In *Standard Languages and Language Standards—Greek, Past and Present*; Routledge: London, UK, 2016; pp. 249–278.
130. Papadakis, G.; Skoutas, D.; Thanos, E.; Palpanas, T. A Survey of Blocking and Filtering Techniques for Entity Resolution. *arXiv* **2020**, arXiv:1905.06167. [CrossRef]
131. Greenspan, M.; Yurick, M. Approximate kd tree search for efficient ICP. In Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling, 2003, 3DIM 2003, Proceedings, Banff, AB, Canada, 6–10 October 2003; IEEE: Piscataway, NJ, USA, 2003; pp. 442–448.
132. Foley, T.; Sugerman, J. KD-tree acceleration structures for a GPU raytracer. In Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware, Los Angeles, CA, USA, 30–31 July 2005; pp. 15–22.
133. SpaCyTextBlob, A TextBlob Sentiment Analysis Pipeline Component for spaCy. Available online: <https://spacy.io/universe/project/spacy-textblob> (accessed on 18 October 2025).
134. Deep-Translator, A Flexible Free and Unlimited Python Tool to Translate Between Different Languages in a Simple Way Using Multiple Translators. Available online: <https://deep-translator.readthedocs.io/en/latest/index.html> (accessed on 18 October 2025).
135. Paga, J.; Miles, M.M. The archaic temple of Poseidon at Sounion. *Hesperia J. Am. Sch. Class. Stud. Athens* **2016**, *85*, 657–710.
136. Hugging Face Transformers, State-of-the-Art Machine Learning for PyTorch, TensorFlow, and JAX. Available online: <https://huggingface.co/docs/transformers/index> (accessed on 18 October 2025).
137. 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. [CrossRef]
138. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
139. Dbpedia Ontology. Available online: <https://dbpedia.org/ontology/> (accessed on 18 October 2025).
140. Mendes, P.N.; Jakob, M.; García-Silva, A.; Bizer, C. DBpedia Spotlight: Shedding Light on the Web of Documents. In Proceedings of the 7th International Conference on Semantic Systems, I-Semantics ’11, Graz, Austria, 7–9 September 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 1–8. [CrossRef]



141. SpaCy Entity Ruler, Pipeline Component for Rule-Based Named Entity Recognition. Available online: <https://spacy.io/api/entityruler> (accessed on 18 October 2025).
142. Language Class, a Text-Processing Pipeline for SpaCy. Available online: <https://spacy.io/api/language> (accessed on 18 October 2025).
143. Doc Class, a Container for Accessing Linguistic Annotations. Available online: <https://spacy.io/api/doc> (accessed on 18 October 2025).
144. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
145. Grinberg, M. *Flask Web Development: Developing Web Applications with Python*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2018.
146. Google Maps Directions. Available online: <https://developers.google.com/maps/documentation/urls/get-started> (accessed on 18 October 2025).
147. Nominatim, Open-Source Geocoding with OpenStreetMap Data. Available online: <https://nominatim.org/> (accessed on 18 October 2025).
148. OpenStreetMap API v0.6. Available online: [https://wiki.openstreetmap.org/wiki/API\\_v0.6](https://wiki.openstreetmap.org/wiki/API_v0.6) (accessed on 18 October 2025).
149. MongoDB—Sharding. Available online: <https://www.mongodb.com/docs/manual/sharding/> (accessed on 18 October 2025).

**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.