Journal of *Personalized Medicine*

Review

Artificial Intelligence-Based Medical Data Mining

Amjad Zia [1], Muzzamil Aziz [2], Ioana Popa [1], Sabih Ahmed Khan [2], Amirreza Fazely Hamedani [2] and Abdul R. Asif [1,*]

[1] Department for Clinical Chemistry/Interdisciplinary UMG Laboratories, University Medical Center, 37075 Göttingen, Germany
[2] Future Networks, eScience Group, Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), 37077 Göttingen, Germany
* Correspondence: asif@med.uni-goettingen.de

Abstract: Understanding published unstructured textual data using traditional text mining approaches and tools is becoming a challenging issue due to the rapid increase in electronic open-source publications. The application of data mining techniques in the medical sciences is an emerging trend; however, traditional text-mining approaches are insufficient to cope with the current upsurge in the volume of published data. Therefore, artificial intelligence-based text mining tools are being developed and used to process large volumes of data and to explore the hidden features and correlations in the data. This review provides a clear-cut and insightful understanding of how artificial intelligence-based data-mining technology is being used to analyze medical data. We also describe a standard process of data mining based on CRISP-DM (Cross-Industry Standard Process for Data Mining) and the most common tools/libraries available for each step of medical data mining.

Keywords: text mining; artificial intelligence; machine learning; medical data; healthcare information

Citation: Zia, A.; Aziz, M.; Popa, I.; Khan, S.A.; Hamedani, A.F.; Asif, A.R. Artificial Intelligence-Based Medical Data Mining. *J. Pers. Med.* **2022**, *12*, 1359. https://doi.org/10.3390/jpm12091359

Academic Editor: Yu-Feng Hu

Received: 5 July 2022
Accepted: 17 August 2022
Published: 24 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

With the rapid growth in online available medical literature, it is almost hard for readers to obtain the desired information without an extensive time investment. For example, in the ongoing COVID-19 pandemic, the number of publications talking about COVID-19 increased very rapidly. In the first 2 years of the pandemic, there were 228,640 articles in PubMed, 282,883 articles in PMC, and 7551 COVID-19 clinical trials listed in ClinicalTrials.gov databases (Data accessed on 16 February 2022), and this is increasing at an amazing speed. Because of the high degree of dimensional heterogeneity, irregularity, and timeliness, these data are often underutilized. This exponential growth in the scientific literature has made it difficult for the researchers to (i) obtain relevant information from the literature, (ii) present information in a concise and structured manner from an unstructured literature pile, and (iii) fully comprehend the current state and the direction of development in a research field.

The rapidly increasing literature cannot be managed and/or processed using traditional technologies and methods within an acceptable period. This massive volume of data makes it rather difficult for researchers to explore, analyze, visualize, and obtain a concise outcome. The process of extracting hidden, meaningful, and engrossing patterns from unstructured text literature is known as text mining [1]. Traditional text mining techniques are not sufficient to cope with the current large volumes of published literature. Therefore, a rapid increase in the development of new data mining techniques based on artificial intelligence can be seen on the horizon for the benefit of patients and physicians. The inclusion of artificial intelligence (also machine learning (ML), deep learning (DL), and natural language processing (NLP) as the subsets) empowers the data mining process with multifold benefits: Gaining new insights into the decision-making process, processing large dataset with increased accuracy and efficiency, and the ability to learn and improve continuously from the new data.

The current review sheds light on the role of different AI-based methods, i.e., NLP and neural network (NN) in medical text mining, the current data mining processes, different database sources, and various AI-based tools used in the text mining process along with various algorithms. We reviewed the latest text mining approaches, highlighted the key differences between medical and non-medical data mining, and presented a set of tools and techniques currently being used for each step of medical literature text mining. Additionally, we described the role of artificial intelligence and machine learning in medical data mining and pointed out challenges, difficulties, and opportunities along the road.

1.1. Medical vs. Non-Medical Literature Text Mining

Human medical data are unique and may be difficult when it comes to mining and analysis. First, due to the fact that humans are the most advanced and the most observed (in-depth) species on the globe, their observation is enriched because humans may provide their sensory input easily compared to the other species on the earth [2]. However, medical data mining faces numerous key challenges, mainly due to the heterogeneity and verbosity of data coming from various non-standardized patient records. Similarly, the insufficient quality of data is also a known issue in medical science that needs to be handled with care for data mining. Such challenges can be met by standardization of the process of selection of patients, collection, storage, annotation, and management of data [3]. However, sometimes this means that existing data and data acquired at multiple centers without good coordination and standard operating procedures (SOPs) could not be used. The major divergence between medical data and non-medical data mining is expected in ethical and legal aspects. The use of information that can be traced back to individuals involves privacy risks, which could result in legal issues. More than fifteen Federal US departments with the US Department of Health and Human Services have issued final revisions to the Federal Policy for the Protection of Human Subjects "the Common Rule, 45 CFR 46, Subpart A" (Protection of Human Subjects, 45 CFR 46 (2018). The federal framework for privacy and security does not apply to the information, which is de-identified or anonymized [4].

The ownership of medical data is another critical issue, as the data are acquired by different entities where the individuals may have been during their treatment or for diagnostic purposes. These entities can gather and store the data as per the authorization of the individual at the time of data acquisition. However, this permission on consent can be withdrawn by the patient at any time, and/or the consent is only valid for a limited period and data must be erased after this time [5]. Most of the clinical text is produced in a telegraphic way and the information is highly enriched. Additionally, it is written for the clinical staff and colleagues, therefore is full of incomplete sentences and abbreviations. Special tools are required to read, understand, and process this text [6]. Electronic patient records, also known as clinical text, have a unique problem in that they are written in a highly specialized language that can only be processed with a few available tools. Secondly, patient records are sometimes written in a telegraphic and information-dense style for clinician-to-clinician communication, and there exists no developed dictionary for such communications to check grammar and spelling mistakes. In addition, doctors and medical staff frequently use rudimentary sentences and frequently fail to mention the object, such as the patient, because the patient is implied in the text. "Arrived with 38.3 fever and a pulse of 132", for example, could be written or simply mentioned.

1.2. Use of Artificial Intelligence and Machine Learning in Medical Literature Data Mining

The digital era has shown immense trust and growing confidence in machine learning techniques to increase the quality of life in almost every field of life. This is the case in health care and precision medicine, where a continuous feed of medical data from heterogeneous sources becomes a key enabler for AI/ML-assisted treatments and diagnosis. For instance, AI today can help doctors to bring better patient outcomes with early diagnosis and treatment plans as well as increased quality of life. Similarly, health organizations and authorities also aim for the timely execution of AI routines for the prognosis of outbreaks and pandemics at the national and international levels. Healthcare today is also witnessing the use of AI-aided procedures for operational management in the form of automated documentation, appointment scheduling, and virtual assistance for patients. In this section, we will see some real-life references of AI\ML tools and technologies currently used in various areas of medical sciences (Table 1).

Table 1. AI\ML products and research prototypes from some leading organizations in healthcare.

Products/Research Prototypes	Treatment/Field of Study	Company/Institution	Reference
MergePACS™	Clinical Radiology Imaging	IBM Watson	Merge PACS—Overview | IBM
BiometryAssist™	Diagnostic Ultrasound	Samsung Medison	https://www.intel.com/content/www/us/en/developer/tools/oneapi/application-catalog/full-catalog/diagnostic-ultrasound.html (accessed on 17 February 2022)
LaborAssist™	Diagnostic Ultrasound	Samsung Medison	
Breast Cancer Detection Solution	Ultrasound, mammography, MRI	Huiying's solution	https://builders.intel.com/ai/solutionscatalog/breast-cancer-detection-solution-657 (accessed on 17 February 2022)
CT solution	Early detection of COVID-19	Huiying's solution	https://builders.intel.com/ai/solutionscatalog/ct-solution-for-early-detection-of-covid-19-704 (accessed on 17 February 2022)
Dr. Pecker CT Pneumonia CAD System	Classification and quantification of COVID-19	Jianpei Technology	https://www.intel.com/content/www/us/en/developer/tools/oneapi/application-catalog/full-catalog/dr-pecker-ct-pneumonia-cad-system.html (accessed on 17 February 2022)

Before going into further detail, it is worth mentioning that data mining and machine learning concepts go hand in hand and overlap each other to an extent but with a clear distinction of the overall outcome of both technologies. Data mining is the process of discovering correlations, anomalies, and new patterns in a large set of data from an experiment or event to forecast results [7]. The basis of data mining is statistical modeling techniques to represent data in some well-defined mathematical model and then use this model to create relationships and patterns among the data variables. Machine learning, on the other hand, is a one-step-ahead approach to data mining, where machine learning algorithms let the computer machine understand the data (with the help of statistical models) and make predictions of its own. That said, data mining techniques always require human interaction to find interesting patterns from a given dataset, whereas machine learning is a relatively modernized technique that enables computer programs to learn from the data automatically and provide predictions without any human interaction.

Natural Language Processing

Natural Language Processing (NLP) is an artificial intelligence (AI) discipline that converts human language into machine language. With the increased usage of computer technology over the last 20 years, this sector has grown significantly [8]. Clinical documentation, speech recognition, computer-assisted coding, data mining research, automated registry reporting, clinical decision support, clinical trial matching, prior authorization, AI chatbots and virtual scribes, risk adjustment models, computational phenotyping, review management and sentiment analysis, dictation and EMR implementations, and root cause analysis are some of the most popular applications of NLP in healthcare [9]. In the literature, a wide range of applications of NLP have been illustrated.

Liu et al. [10] used clinical text for entity recognition using word embedding (WE)-skipgram and long short-term memory (LSTM) techniques and achieved an accuracy of 94.37 percent, 92.29 percent, and 85.81 percent for de-identification, event detection, and concept extraction, respectively, based on the micro-average F1-score. Deng et al. [11] used concept embedding (CE)–continuous bag of words (CBOW), skip-gram, and random projection to generate code and semantic representations from clinical text. Afzal et al. [12]

have developed a pipeline for question generation, evidence quality recognition, ranking, and summarization of evidence from biomedical literature and presented an accuracy of 90.97 percent. Besides these examples, Pandey et al. [13] listed 57 papers published between 2017 and 2019 that used NLP techniques and various text sources, such as clinical text, EHR inputs, Chinese medical text, cancer pathology reports, biomedical text, randomized controlled trial (RCT) articles, clinical notes, and EMR text-radiology reports, among others.

2. Standard Process for Data Mining

In response to the demand for a standard data mining method, industry leaders collaborated with a diverse group of practitioners (service providers, management consultants, data mining users, data warehouse vendors) and data mining experts to develop a free, well-documented, and non-proprietary data mining model [14]. Numerous methods are available for data mining, such as ASUM (Analytics Solutions Unified Method), CRISP-DM (Cross-Industry Standard Process for Data Mining), KDD (Knowledge discovery in databases), SEMMA (Sampling, Exploring, Modifying, Modelling, and Assessing), and POST-DS (Process Organization and Scheduling electing Tools for Data Science) [15]. In this study, we employ the CRISP-DM model for data mining because it is a complete and comprehensive data mining approach. In 1997, the CRISP-DM consortium developed a generic process model for data mining to establish guidelines for data mining beginners, the community, and experts, which can be modified for any particular need [14]. For example, to deal with the problem of multidimensional time-series data in a neonatal intensive care unit (NICU), the CRISP-DM model was modified to support and accommodate temporal data mining (TDM), which is named CRISP-TDM [16]. In the lifecycle of a data mining process, the CRISP-DM reference model has six phases (Figure 1): Business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The details of the available tools and technologies for each phase are described in the rest of this article.

2.1. Business Understanding

The first and most critical part of data mining is business understanding, which includes setting project objectives, and targets, assessing the situation, execution plans, and risk assessments [14]. Setting project objectives requires a complete grasp of the project's genuine goal to define the associated variables. The steps in the data understanding phase according to CRISP-DM are to (1) determine the business objectives (to fully comprehend the project's goal, identify the key players, and establish business success criteria), (2) assess the situation (to identify resource availability (especially data), identify project risks and potential solutions to those risks, and calculate the cost–benefit ratio), (3) clarify the data mining goals (to establish project goals and success criteria), (4) produce a project plan (to develop detailed plans for each project segment, including a timeline and technology and tool selection).

Martins et al. [18] used a data mining approach to predict cardiovascular diseases (while using RapidMiner and Weka software). The main question addressed by the project is how to detect cardiovascular disease at an early stage in a person who is at a high risk of developing the disease and thus avoid premature death. As a result, the primary set of goals is to create a solution for predicting cardiovascular diseases in patients using patient data, to shorten the time required for disease diagnosis, and to provide the patients with immediate and adequate treatment.

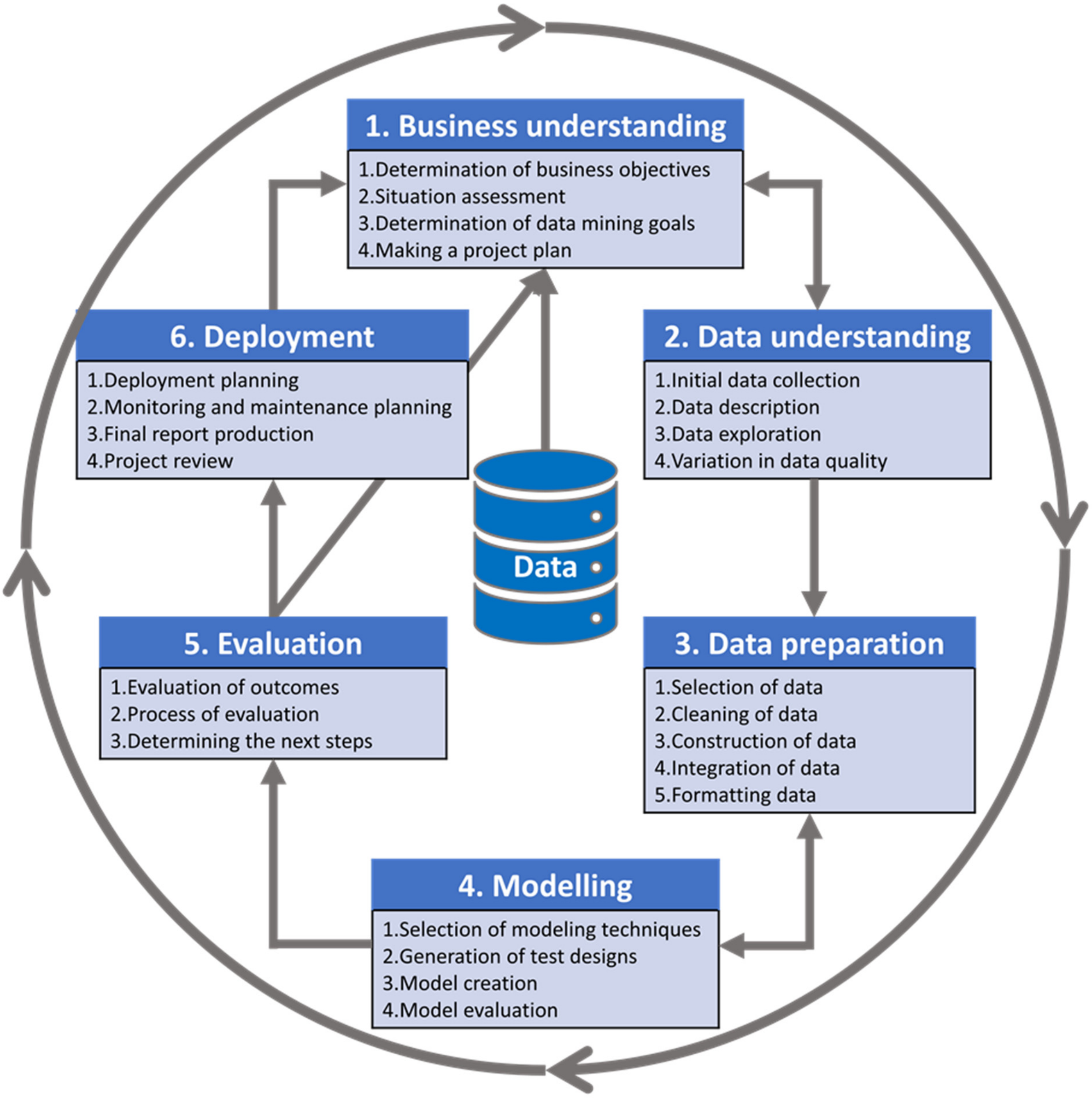

Figure 1. Cross-Industry Standard Process for Data Mining (CRISP-DM)—adapted from the webpage of the Data Science Process Alliance [17] (www.datascience-pm.com/crisp-dm-2/, accessed on 16 April 2022). The circular nature of the data mining process is symbolized by the outer circle, while the arrows that connect the phases show the most essential and common dependencies.

2.2. Data Understanding

The emphasis in this phase (second phase), according to CRISP-DM, is on data source identification, data acquisition, initial data collection, familiarization with the data, and identifying problems in the acquired data. The steps in the data understanding phase are (1) acquire the initial data (to gather the data from various sources, insert it into the analysis program, and integrate it), (2) explain the data (to study and report on the acquired data's surface properties such as field identities, data format, data quantity, and the number of records, etc.), (3) explore the data (to delve deeper into the data by querying, visualizing, and identifying relationships between data points, as well as to generate an exploration report), and (4) verify data quality (to inspect and document the data quality and any quality-related issues) [14]. In this phase, one focuses on identifying data sources for various types of data, the process of acquisition of the data, and handling access restrictions in data acquisition. A tremendous amount of data is generated by the health care industry and medical institutions every day from medical imaging, patient monitoring, and medical records [7]. Some of the most common types of medical data are experimental data, medical literature, clinical textual data, medical records, images/videos (e.g., MRI), and omics data (e.g., genomics, proteomics). For example, Martins et al. [18] used a data mining approach to predict cardiovascular diseases. For data understanding, the dataset for cardiovascular disease prediction came from the Kaggle data repository and focused on detecting cases of cardiovascular disease. The dataset included 70,000 registered patients with 12 disease-related attributes collected during the patients' medical examinations.

2.2.1. Literature Extraction/Data Gathering

The first task in the data understanding phase is to identify data sources, acquire data from these sources, identify problems during data acquisition, such as data restrictions and data privacy policies, and document the solutions [14]. Text/data mining frequently uses public Internet-based sources such as the World Wide Web. The retrieval of content from public sources is referred to as "web scraping" or "web crawling". Web scraping can be performed manually, but it can also be performed automatically with the help of a web crawler. Manual scraping a large database such as PubMed, which contains millions of peer-reviewed publications, requires a lot of time and effort. Only automated processing can provide the necessary quality, response time, and homogeneity for their analysis with such a large database. As a result, there is always a high demand for web scraping techniques and tools tailored to customer requirements. PubMed, for example, is a massive database of biomedical literature that contains 34 million citations (as of 11 May 2022) collected from online books, life science journals, and MEDLINE, and a massive number of new publications are added every year [19]. Web crawlers are used to search for and harvest the necessary data from it. Guo et al. [20], for example, collected COVID-19 data published by local health authorities using a web crawler (developed using the Python language and connected with a MySQL database).

Although web scraping and web crawling may seem to be identical, they have several distinctions (Figure 2). While the terms "web scraping" and "web crawling" are sometimes interchanged, they refer to two distinct processes [21,22]. Web crawling is a broad term that refers to the process of downloading information from a website, extracting the hyperlinks included within, and following them (Figure 2). Typically, downloaded information is saved in a database or indexed to enable searching. Essentially, search engines are crawlers. All that is required is to see a page in its entirety and indexing it. When a bot crawls a website, it scans each page and link, all the way to the website's last line, looking for any information. Web crawlers are primarily used by major search engines such as Google, Bing, and Yahoo, as well as statistics organizations and online aggregators. Typically, a web crawler collects general information, while scrapers collect particular datasets [23,24]. On the other hand, web scraping is the process of obtaining data from a web page and extracting specific information that can be saved almost anywhere (database, file, etc.) as shown in Figure 2. An online scraper, also known as a web data extractor, is similar to a web crawler in that it detects and locates website content. In contrast to a web crawler, which uses pseudo-random IDs, web scraping uses specific identifiers, such as the HTML structure of the web pages from which data must be collected. Web scraping refers to the use of robots to extract specific datasets from the internet. The obtained data can be compared, checked, and analyzed in accordance with the demands and objectives of a specific organization [25].

Several text mining tools are now available. Kaur and Chopra [26] compared 55 popular text mining tools and their features and discovered three categories: (1) Proprietary (company-owned—39 tools); (2) open source (free—13 tools); and (3) online text mining tools (run directly from a website—3 tools). Four tools that were not examined in the prior review but are now on the list of well-liked text mining tools are contrasted in Table 2. All of these Python-based tools serve the same purpose, but with different goals and objectives. ''Requests" has an advantage over other tools in that it is easy to use, making it an excellent choice for any simple web scraping task. Scrapy is best suited for large-scale web scraping projects, as opposed to the other three tools (requests, beautiful soup, and selenium), which are best suited for small-scale scraping tasks. The "Beautiful Soup" tool has advantages such as being simple to understand, learn, and use, and it can extract information from a disorganized website. Selenium has a significant advantage over the other scraping tools described because it can scrape websites with heavy JavaScript. Table 2 provides descriptions of more hierarchical comparisons.

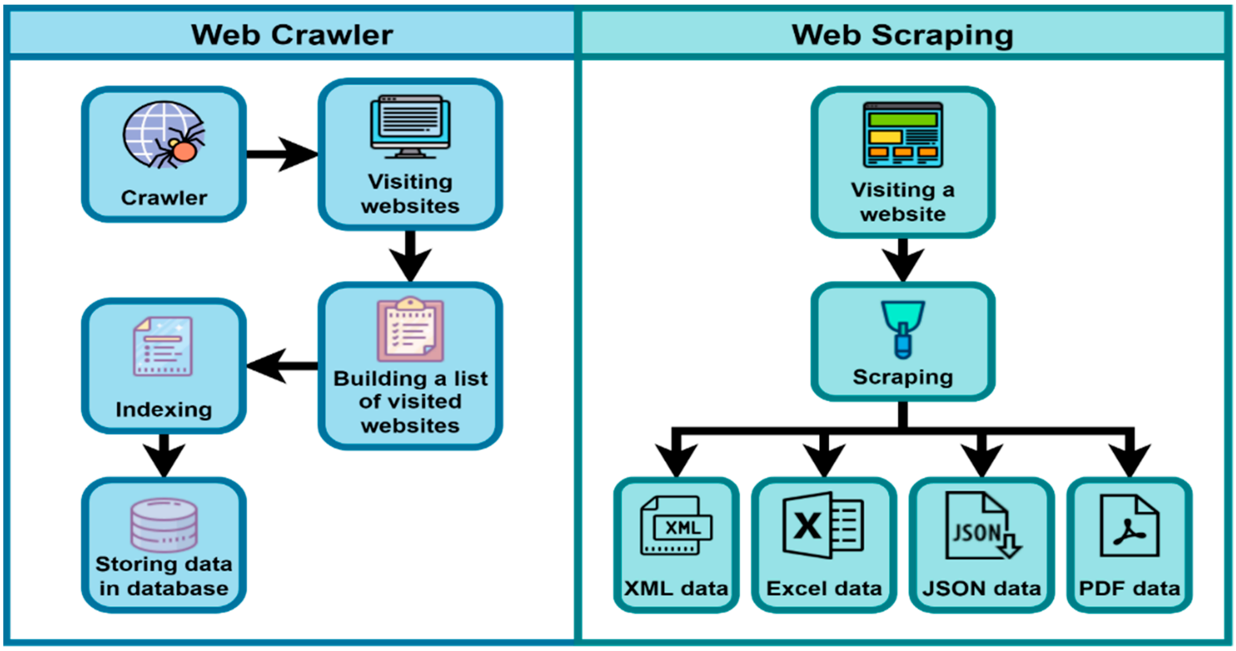

Figure 2. Comparison between web crawling and web scraping.

Table 2. Comparison between four text mining tools.

	Requests	Scrapy	Beautiful Soup	Selenium
What is it?	HTTP library for Python	Open-source web framework written in Python	python library	Open-source application framework tool and python library
Goal	Sending HTTP/1.1 requests using Python	• Can crawl or scrape websites and extract the structured data and saves it • Can also be used for a wide range of tasks, monitoring, and automated testing	• Can parse the data and scrape the web pages • Extract information from XML and HTML documents	• Useful for web scraping websites that are JavaScript heavy
Ideal usage	Used for simple and low-level complex web scraping tasks	• Framework used for complex web scraping or web crawling tasks. • Used for large-scale projects	• Used for smaller web scraping tasks • Toolkit for searching through a document (XML or HTML) and extracting important information	• Developed for web testing • Used for test automation of web applications • Scraping JavaScript-heavy websites • Used for small-scale and low-level complex projects

Table 2. *Cont.*

	Requests	Scrapy	Beautiful Soup	Selenium
Advantage	• A simple way to retrieve data from URL • Scraping data from web • Allows to read, write, post, delete, and update the data for the given URL • Extremely easy to deal with cookies and sessions	• Portable library • Runs on Linux, Windows, and Mac • One of the faster scraping libraries • Can extract websites much faster than other tools • Consumes less memory and CPU usage • Building a robust, and flexible application with different functions	• Learning and mastering it is easy • Community support is readily available to resolve issues.	• Deals with the Core JavaScript-heavy website • Can handle AJAX and PJAX requests
Selectors	None	JCSS and XPath	CSS	CSS and Xpath
Documentation	Detailed and simple to understand	Detailed and simple to understand	Detailed and simple to understand	Detailed and very complex
GitHub stars	46.8 k	42.7 k	-	22.7 k
Reference	Chandra and Varanasi [27]	Kouzis-Loukas [28]	Richardson [29]	Sharma [30]

Access Restriction

When a web crawler visits a website, some pages or the entire website possess access restrictions. These restrictions are implemented mainly by the site owners due to data confidentiality, data integrity, and data quality, as well as legal concerns. A crawler usually performs multiple requests per second and downloads large files to obtain the data in a short time, which can cause a website server to crash. To tackle this problem, numerous methods are available. Canonical tag, robots.txt, x-robots-tag, the metarobots tag, and others are files provided by the website owners to follow the instructions for scraping the website without creating any problem. For example, "robots.txt" files are frequently used by websites to convey their scraping and crawling intents. Robots.txt files enable scraping bots to crawl specific sites, while malevolent bots, on the other hand, are uninterested in robots.txt files (which act as a "do not enter" sign) as explained below in Figure 3.

Data Collection from Different Sources

The pace at which medical data are being generated is increasing day by day during the massive information explosion year, and global information is being produced in massive quantities in every field, including healthcare [31,32]. Administrative records, biometric data, clinical registration, diagnostics, X-rays, electronic health records, patient report data, treatments, results, and other types of medical data are all included in medical data. These massive and complex characteristics make data difficult to deal with for a meaningful and unknown outcome. Healthcare centers and medical institutions around the world have proposed a variety of medical information systems to deal with rapidly growing data and provide the best possible services and care to patients [32]. The most common way to collect and store the data is by management software, which can store all electronic and non-electronic records. Several software products are available, e.g., eHospital Systems (adroitinfosystems.com/products/ehospital-systems, accessed on 11 April 2022) and the DocPulse Clinic/Hospital Information Management System (docpulse.com, accessed on 11 April 2022).

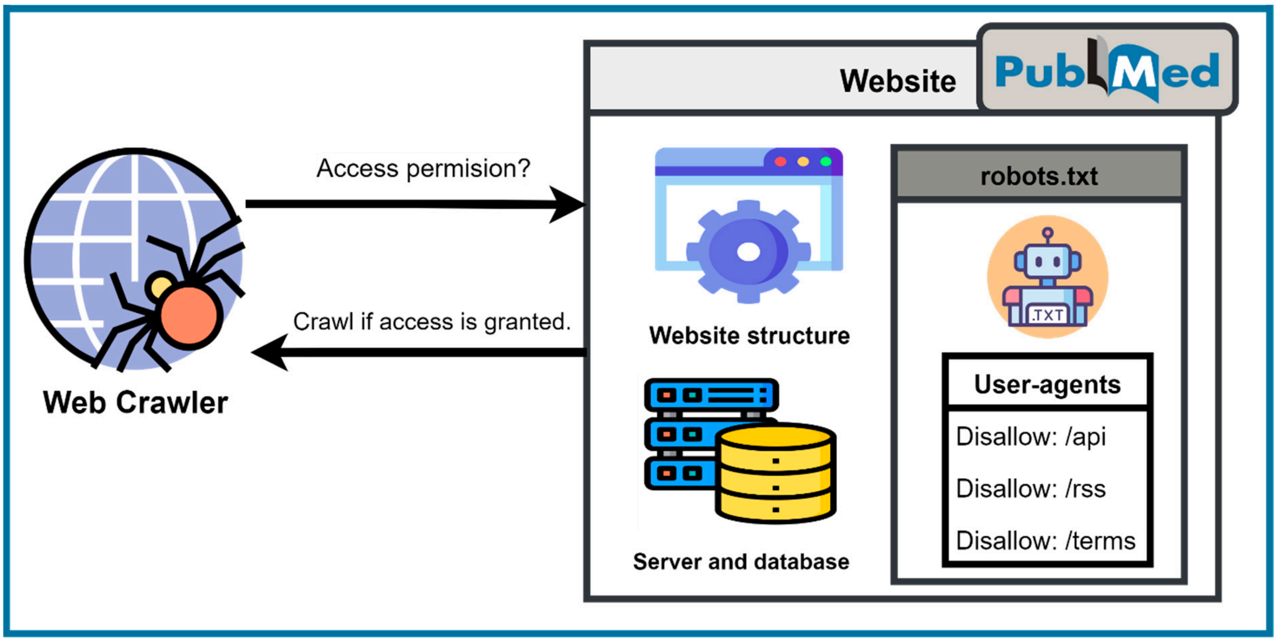

Figure 3. Layout for access restrictions.

For text mining, data collection from data sources is the key step. In medical science, various types of medical data, as well as trends, are generated at a rapid pace, which can be differentiated into five categories, as follows:

1. Hospital management software (Patient data/Clinical narratives).
2. Clinical trials.
3. Research data in Medicine.
4. Publication platforms for Medicine (PubMed, for instance).
5. Pharmaceuticals and regulatory data.

Tables 3–5 provide further details about the different types of data sources. Patient data generated by clinical trials is available from various sources, as shown in Table 3. Medical researchers benefit from open-access databases because they have enormous volumes of data, rich data content, broad data coverage, and a cost-effective study strategy. There exist several datasets and databases publicly available related to various medical fields that contain many medical record variables (Table 4). Textual information is growing rapidly, and it is difficult to grab concise information fast and structured manner. The published literature is the most abundant and primary source of textual information in the health care field (Table 5).

Table 3. Databases and registries for clinical trials.

Databases/Registries	**Trial Numbers**	**Provided by**	**Location**	**Founded Year**	**URL**
ClinicalTrials.gov	405,612	U.S. National Library of Medicine	Bethesda, MD, USA	1997	https://clinicaltrials.gov/ (accessed on 11 April 2022)
Cochrane Central Register of Controlled Trials (CENTRAL)	1,854,672	a component of Cochrane Library	London, UK	1996	https://www.cochranelibrary.com/central (accessed on 11 April 2022)
WHO International Clinical Trials Registry Platform (ICTRP)	353,502	World Health Organization	Geneva, Switzerland	-	https://trialsearch.who.int/ (accessed on 11 April 2022)
The European Union Clinical Trials Database	60,321	European Medicines Agency	Amsterdam, The Netherlands	2004	https://www.clinicaltrialsregister.eu/ctr-search/search (accessed on 11 April 2022)

Table 3. *Cont.*

Databases/Registries	Trial Numbers	Provided by	Location	Founded Year	URL
CenterWatch	50,112	-	Boston, MA, USA	1994	http://www.centerwatch.com/clinical-trials/listings/ (accessed on 11 April 2022)
German Clinical Trials Register (Deutsches Register Klinischer Studien—DRKS)	>13,000	Federal Institute for Drugs and Medical Devices	Cologne, Germany		https://www.bfarm.de/EN/BfArM/Tasks/German-Clinical-Trials-Register/_node.html (accessed on 11 April 2022)

Table 4. Research data in Medicine.

Databases	No. of Datasets	Owned by	Domains	Available Resources	URL	Ref
Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC)	262	National Institute of Health, Calverton, MD, USA	Cardiovascular, pulmonary, and hematological	Specimens and Study Datasets	https://biolincc.nhlbi.nih.gov/studies/ (accessed on 4 April 2022)	[33]
Biomedical Translational Research Information System (BTRIS)	Five billion rows of data	Bethesda, MD, USA	Multiple subjects	Study Datasets	https://btris.nih.gov/ (accessed on 4 April 2022)	[34]
Clinical Data Study Request	3135	The consortium of clinical study Sponsors	Multiple subjects	Study Datasets	https://www.clinicalstudydatarequest.com/ (accessed on 4 April 2022)	[35]
Surveillance, Epidemiology, and End Results (SEER)	-	National Cancer Institute, Bethesda, MD, USA	Cancer (All types)—Stage and histological details	Study Datasets	https://seer.cancer.gov/ (accessed on 4 April 2022)	[36]
Medical Information Mart for Intensive Care (MIMIC) MIMIC-III	53,423 patients	MIT Laboratory for Computational Physiology, Cambridge, MA, USA	Intensive Care	Patient data (vital signs, medications, laboratory measurements, observations and notes charted by care providers, survival data, hospital length of stay, imaging reports, diagnostic codes, procedure codes, and fluid balance)	https://mimic.mit.edu/ (accessed on 4 April 2022)	[37,38]
MIMIC-CXR	65,379 patients (377,110 images of chest radiographs)					[39]
National Health and Nutrition Examination Survey (NHANES)	-	Centers for disease control and prevention, Hyattsville, MD, USA	Dietary assessment and other nutrition surveillance	data nutritional status, dietary intake, anthropometric measurements, laboratory tests, biospecimens, and clinical findings.	https://www.cdc.gov/nchs/nhanes/index.htm (accessed on 4 April 2022)	[40]
Global Burden of Disease (GBDx)	-	Institute for Health Metrics and Evaluation, Seattle, WA, USA	Epidemic patterns and disease burden	Surveys, censuses, vital statistics, and other health-related data	https://ghdx.healthdata.org/ (accessed on 4 April 2022)	[41]
UK Biobank (UKB)	0.5 million	Stockport, UK	In-depth genetic and health information	Genetic, biospecimens, and health data	https://www.ukbiobank.ac.uk/ (accessed on 4 April 2022)	[42]
The Cancer Genome Atlas (TCGA)	molecularly characterized over 20,000 cancer samples spanning 33 cancer types	National Cancer Institute, NIH, Bethesda, MD, USA	Cancer genomics	over 2.5 petabytes of epigenomic, proteomic, transcriptomic, and genomic data	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (accessed on 4 April 2022)	[43]
Gene Expression Omnibus (GEO)	4,981,280 samples	National Center for Bioinformatics (NCBI), NIH, Bethesda, MD, USA	Sequencing and gene expression	4348 datasets available	https://www.ncbi.nlm.nih.gov/geo/ (accessed on 4 April 2022)	[44]

Table 5. Biomedical literature sources.

Source	Articles (Million)	Launched by	Publication Type	Topic	Online	Link
PubMed	33	National Center for Biotechnology Information (NCBI)	Abstracts	Biomedical and life sciences	1996	https://www.ncbi.nlm.nih.gov/pubmed/ (accessed on 4 April 2022)
PubMed Central (PMC)	7.6	National Center for Biotechnology Information (NCBI)	Full text	Biomedical and life sciences	2000	https://www.ncbi.nlm.nih.gov/pmc/ (accessed on 4 April 2022)
Cochrane Library	-	Cochrane	Abstracts and full text	Healthcare	-	https://www.cochranelibrary.com/search (accessed on 4 April 2022)
bioRxiv	-	Cold Spring Harbor Laboratory (CSHL)	Unpublished preprints	Biological sciences	2013	https://www.biorxiv.org/ (accessed on 4 April 2022)
medRxiv	-	Cold Spring Harbor Laboratory (CSHL)	Unpublished manuscripts	Health sciences	2019	https://www.medrxiv.org/ (accessed on 4 April 2022)
arXiv	2.05	Cornell Tech	Non-peer-reviewed	Multidisciplinary	1991	https://arxiv.org/ (accessed on 4 April 2022)
Google Scholar	100 (in 2014)	Google	full text or metadata	Multidisciplinary	2004	https://scholar.google.com/ (accessed on 4 April 2022)
Semantic Scholar	205.25	Allen Institute for Artificial Intelligence	Abstracts and full text	Multidisciplinary	2015	https://www.semanticscholar.org/ (accessed on 4 April 2022)
Elsevier	17 (as of 2018)	Elsevier	Abstracts and full text	Multidisciplinary	1880	https://www.elsevier.com/ (accessed on 4 April 2022)
Springer Nature	-	Springer Nature Group	Abstracts and full text	Multidisciplinary	2015	https://www.springernature.com/ (accessed on 4 April 2022)
Springer	-	Springer Nature	Abstracts and full text	Multidisciplinary	1842	https://link.springer.com/ (accessed on 4 April 2022)

2.3. Data Preparation

In the third phase (data preparation) of CRISP-DM, a final dataset is created from the raw data, which will be used in the modeling tool. This phase is the major part (ca. 80%) of a text/data mining project. The steps in the data preparation phase are (1) data selection (to choose the dataset along with its attributes that will be used for the analysis based on the project goals, quality, data type, and volume.), (2) data cleaning (to estimate missing data and improve the dataset by correcting, imputing, or removing incorrect values), (3) data construction (to create derived attributes or entirely new records, as well as to transform data as needed), (4) data integration (to create new datasets and aggregate new values by combining data from multiple sources), (5) data formation (to remove inappropriate characters from the data and change the data's format or design so that it fits into the model) [14].

2.3.1. Data Cleaning/Data Transformation

The primary goal of data cleaning is to detect and remove duplicate data and errors from a dataset to create a reliable dataset. Cleaning data entails identifying and removing entries from a dataset that are corrupt, incorrect, duplicated, incomplete, or improperly formatted (see Figure 4). Data cleaning is required to analyze information from multiple sources [45–47].

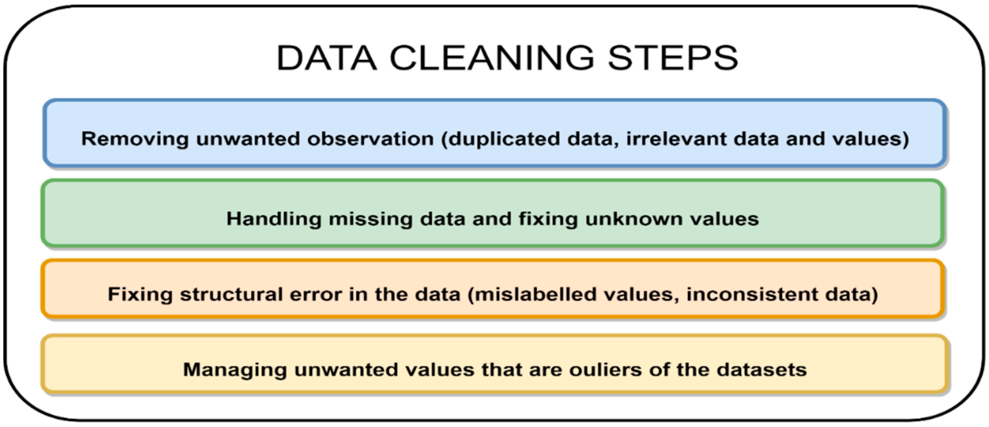
DATA CLEANING STEPS
Removing unwanted observation (duplicated data, irrelevant data and values)
Handling missing data and fixing unknown values
Fixing structural error in the data (mislabelled values, inconsistent data)
Managing unwanted values that are ouliers of the datasets

Figure 4. Steps for data cleaning.

Various related tools and python libraries are discussed in the following sections.

Python Libraries for Data Cleaning include the following:

1. NumPy is a quick and easy-to-use open-source Python library for data processing. Because many of the most well-known Python libraries, including Pandas and Matplotlib, are based on NumPy, it is a fundamentally crucial library for the data science environment. The primary purpose of the NumPy library is the straightforward manipulation of large multidimensional arrays, vectors, and matrices. For numerical calculations, NumPy also offers effectively implemented functions [48].
2. Data processing tasks such as data cleaning, data manipulation, and data analysis are performed using the well-known Python library Pandas. The Python Data Analysis Library is referred to as "Pandas". Multiple modules for reading, processing, and writing CSV, JSON, and Excel files are available in the library. Although there are many data cleaning tools available, managing and exploring data with the Pandas library is incredibly quick and effective [49].
3. An open-source Python library for automating data cleaning procedures is called DataCleaner. Pandas Dataframe and scikit-learn data preprocessing features comprise its two separate modules [50].

The data are then transformed into the proper format after being cleaned (Excel, JSON, or XML). Data transformation makes it simpler to preprocess data and/or text. Depending on the modifications that must be made, the data transformation may be straightforward or complicated. The data are easier to use for both humans and computers after transformation because it is more structured and organized. Additionally, it becomes simpler to integrate into various programs and systems [46].

Various related tools are discussed in the following sections.

1. Generation of Bibliographic Data is known as GROBID. It is a machine-learning library that has developed into a state-of-the-art open-source library for removing metadata from PDF-formatted technical and scientific documents. The library plans to reconstruct the logical structure of its original document in addition to simple bibliographic extraction in order to support large-scale advanced digital library processes and text analysis.

GROBID develops fully automated solutions based on machine learning models for that reason. ResearchGate, Mendeley, CERN Inspire, and HAL, France's national publication repository, are just a few of the commercial and open-access scientific services that the library is connected to.

The result is to extract and transform PDF documents into XML TEI format, supplement the extracted information with other online services, and illustrate the findings gathered in PDF documents of scientific papers [51,52].

2. BioC is a straightforward and straightforward format for exchanging text data and annotations, as well as for simple text processing. Its primary goal is to provide an abundance of research data and articles for text mining and information retrieval. They are available in a variety of file formats, including BioC XML, BioC JSON, Unicode, and ASCII. These formats are available through a Web API or FTP [53].

To summarize, data cleansing improves a dataset's consistency, while transformation simplifies data processing. Both processes improve the training dataset's quality for model construction.

2.3.2. Feature Engineering

Choosing, modifying, and converting raw data into features that may be utilized in supervised learning is a process of feature engineering, often referred to as feature extraction. This machine learning technique, feature engineering, uses data to generate new variables that are not present in the training set. To streamline and accelerate data transformations while also improving model accuracy, it can generate new features for both supervised and unsupervised learning. With machine learning models, feature engineering is necessary. Regardless of the architecture or the data, a bad feature will directly affect your model. Numerous tools are available to automate the entire feature engineering process and to generate a large pool of features in a short period for both classification and regression tasks. Some feature engineering tools are FeatureTools, AutoFeat, TsFresh, Turi, Azure Machine Learning Studio, ZOMBIE, FeatureFu, and OneBM [54,55].

Vijithananda et al. [56] extracted features from MRI ADC images of a brain tumor. The following features were extracted from labeled MRI brain ADC image slices from 195 patients: Skewness, cluster shade, pixel values (he demographics), prominence, Grey Level Co-occurrence Matrix (GLCM) features, energy, contrast, entropy, variance, mean, correlation, homogeneity, and kurtosis. Both GLCM homogeneity and skewness were excluded because they scored the lowest in the ANOVA f-test feature selection process. The Random Forest classifier outperformed Decision Trees, Nave Bayes, Linear Discriminant Analysis, K-Nearest Neighbors (KNN), and Logistic Regression and was chosen for further model development. The final model had an accuracy of 90.41 percent in predicting malignant and benign neoplasms.

2.3.3. Searching for Keywords

The extraction of keywords or key phrases from text documents is known as keyword extraction. They are chosen from among the phrases in the text document and describe the topic of the document. Several popular methods are available for automatically extracting keywords. Those are used in processes that automatically extract keywords from documents to select the most frequently used and significant words or phrases from the text document. This classifies keyword extraction methods as part of the natural language processing field, which is important in machine learning and artificial intelligence. [57]. Keyword extractors are used to extract words (keywords) or groups of two or more words that form a phrase (key phrases).

FlashText, for example, is a free and open-source Python package that enables keyword search and replacement and is one of the recently described keyword extraction tools [58]. It performs a full analysis using an Aho-Corasick algorithm and a Trie Dictionary. As a general rule, keyword matching entails scanning the corpus (human-created documents comprise a large, structured set of texts) for each term. Consider the following scenario: Someone has 100 keywords and needs to search through 2000 papers; a single term is selected at a time and a search of the 2k corpus is performed; the search is continued for 100 × 2000 is 200,000 iterations. In addition to this keyword search tool, four Python-based tools are selected from the various keyword and phrase extraction tools that are available, and their features, benefits, and NLP tasks are contrasted in Table 6.

Table 6. Searching for relevant content.

	Natural Language Toolkit	**SpaCy**	**Scikit-Learn NLP Toolkit**	**Gensim**
What is it?	open-source python platform for handling human language data	open-source python library for advanced natural language processing	machine learning software library for the Python programming language	fastest python library for the training of vector embedding
Features			• Based on NumPy, SciPy, and Matplotlib • An easy and efficient way to analyze predictive data • Easily accessible and reusable in different contexts	
Advantage	• Most well-known and comprehensive NLP libraries with many extensions • offers support in the largest number of languages	• easy to use • fully integrated with Python • compatible with other deep learning frameworks • many already trained statistical models available • applicable to many different languages • high speed and performance • freely available • able to process long texts • platform-independent usable	• simple and efficient tools for machine learning, data mining, and data analysis • freely available for everyone • applicable to different application areas, like natural language processing	• Provides ready-to-use models and corpora • Models pre-trained for specific areas such as health care • Processes large amounts of data using streaming data
NLP Tasks	• Classification • Tokenization • Stemming • Tagging • Parsing	• Classification • Tokenization • Stemming • Tagging • Parsing • Named Entity recognition • Sentiment Analysis	• Classification • Topic Modeling • Sentiment Analysis	• Text similarity • Text summarization • Topic Modeling
GitHub stars	10.4 k	22.4 k	49 k	12.9 k
Website	nltk.org (accessed on 16 March 2022)	spacy.io (accessed on 16 March 2022)	scikit-learn.org (accessed on 16 March 2022)	radimrehurek.com/gensim/ (accessed on 16 March 2022)
Reference	Bird et al. [59]	Honnibal [60]	Pedregosa et al. [61], Pinto et al. [62]	Rehurek and Sojka [63]

2.4. Modeling

In the fourth phase (known as modeling) of CRISP-DM, various modeling techniques are tested and calibrated by adjusting the model parameters to achieve the best results [14]. The steps in the modeling process are (1) choosing a modeling technique to select one or more task-specific models/algorithms/assumptions, (2) the creation of test designs to determine the model's strength by evaluating the model's quality and validity, (3) the building of models (to use the modeling tool for building models from the prepared dataset, adjust the model parameter, and describe the model), and (4) the evaluation of models to explain the model outcome based on subject knowledge, the predetermined success norms, and the test design, rank the multiple generated models, and readjust the parameter settings—if required.

From several available models for organizing and analyzing the data, the selection of a model depends on the purpose (e.g., forecast) and the type of data used (unstructured or structured). A model is a set of data, patterns, and statistics. The available data-mining models are divided into two categories: Predictive and descriptive. Descriptive models are frequently used to determine patterns in data that can be explained by humans. Predictive models use known results from various datasets to forecast unidentified or future values of other variables of interest. Predictive models are usually based on the previously provided data and their results. Classification, prediction, regression, and time series analysis are tasks in the predictive models. Descriptive model data mining tasks comprise clustering, associating rules, sequence discovery, and summarization (Figure 5). A number of algorithms/methods are available for the prediction and analysis of patterns in the data. However, the selection of the algorithm is mainly depending on the dependent variables whether labeled or unlabeled. If the dependent variable/s in the dataset are labeled, a supervised learning algorithm is used. Decision trees, the random forest (RF), support vector machines (SVMs), and competitive risk model are commonly used algorithms. In contrast, if the dependent variables in the data are not labeled, an unsupervised learning method is used. Clustering analysis, partition clustering, hierarchical clustering, principal component analysis (PCA), and association analysis are some of the unsupervised learning algorithms [64,65].

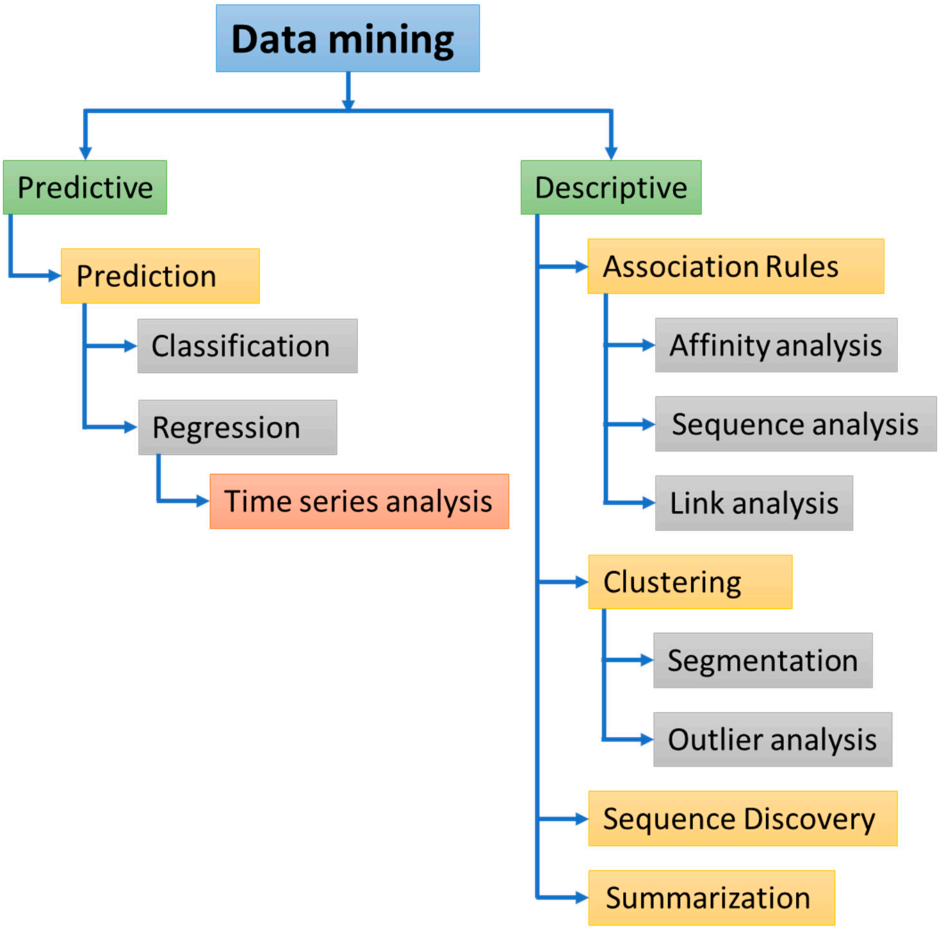

Figure 5. Predictive and descriptive data mining tasks.

The dataset is the primary distinction between supervised and unsupervised machine learning. It is referred to as supervised learning if the dataset employs a labeled dataset for input and output, whereas unsupervised learning techniques use unlabeled data. As the name suggests, supervised learning entails the external supervision of a model's training. Unsupervised learning, on the other hand, does not involve any supervision. Additionally, in the case of supervised learning, the goal is to predict the outcome of new data. In the case of unsupervised learning, the goal is to find hidden patterns and gain insight from enormous amounts of new data. In contrast to supervised learning models, which are straightforward, unsupervised learning models require a large training set to produce the desired results, making them computationally complex. Some of the applications of supervised learning models include diagnosis, identity fraud detection, image classification, price predictions, sentiment analysis, spam detection, market forecasting, and weather forecasting. Unsupervised learning models are used in the pipelines for anomaly detection, big data visualization, customer personas, feature elicitation, recommended systems, structure discovery, and targeted marketing [64,66].

As an instance of the modeling example, the suitability of a WebCrawler (StormCrawler) for the acquisition of all health-related web content on the German Health Web (Germany, Austria, and Switzerland) was investigated by Zowalla et al. [67]. For this purpose, a support vector machine classifier model was trained to distinguish between health-related and non-health-related web pages using the dataset created from the German health web. This model was tested for accuracy and precision on an 80/20 training/test split and against a crowd-validated dataset. For predicting cardiovascular diseases, the best-suited technique was the 'Decision Tree' compared with eight other techniques, i.e., Deep Learning, Nearest Neighbor (k-NN), Gradient Boosted Tree, Generalized Linear Model, Logistic Regression, Naïve Bayes, Random Forest, and Rule Induction [18]. Furthermore, some parameters were optimized using the optimized parameters operator to achieve better results when using the 'Decision Tree'.

2.5. Data Model Validation and Testing

This step's primary goal is to validate and test the selected model for the data in the model development process. The validation procedure is used to ensure that the developed model is accurate enough for the intended use [68]. The first half of this step, model validation, is important because the used/newly developed model cannot be relied on solely because it was designed to fit the training data and demonstrates that the training data fits the model well. To validate a model, output predictions are made in scenarios unrelated to the training set, and the same statistical measures of fit are computed. The second half of this step involves testing the model with test data and comparing its accuracy with the results of the validation step. Only when a model is compared to test data and statistical calculations show a satisfactory match is it considered "ready". For the classification of tumor and non-tumor samples, Dong et al. [69] employed a training dataset (which consists of mass spectrometry (MS) raw data obtained from 194 paired tumor and non-tumor samples) to train different models and used a similar type of dataset (which consists of MS raw data obtained from 58 paired tumor and non-tumor samples) as a test dataset. The convolutional neural network (CNN), gradient boosting decision tree (GBDT), support-vector machine (SVM), principal component analysis (PCA) plus SVM, logistic regression (LR), and random forest (RF) were compared, and the CNN model showed the highest accuracy. Some of the ML model validation testing tools include Apache Spark, Excel, Hadoop, KNIME, Python, R, RapidMiner, SAS, SQL, and Tableau.

2.6. Evaluation

In the fifth phase (known as evaluation) of CRISP-DM, a more thorough evaluation and review of the model's construction is conducted to ensure that the model properly achieves the business objectives. The steps in the evaluation phase are (1) the assessment of outcomes to assess how well the model achieves the project's goals, discover additional constraints, information, or clues about future paths, and present the project's final statement, (2) the review process to conduct a more in-depth review of the project and address quality assurance concerns, and (3) the decision for further steps to determine whether or not to proceed with the deployment or to make changes for the improvement [14].

After the analysis of text data, the next step is to visualize the data meaningfully for interpretation and communication purposes. Text visualization is primarily accomplished through the use of charts, graphs, maps, timelines, networks, word clouds, and so on. These visualized results allow humans to read the most important aspects of a large amount of information. There are several tools available to display the analyzed data. These tools make it easy to identify and discover patterns, outliers, trends, and insights in data straightforwardly and understandably. Effective data visualization has benefits and advantages such as easy understanding of the outcome, effortless and prompt decision-making, and a higher degree of engagement for a diverse audience over other communication methods (e.g., verbal communication). For successful data visualization, there are three main principles: (1) Depending on the purpose, select the appropriate visualization style, (2) the selected visualization style should be appropriate for the targeted audience, and (3) the chosen visualization style should be accompanied by an effective graphic design [70]. The most important aspects of selecting the appropriate visualization style are considering the selected data and the aim of the visualization. For example, line and bar charts are suitable for comparing data points across a dataset. Diverse visualization styles are available for creating attractive and effective visual information, i.e., typographic visualization (e.g., word cloud), graph visualization (e.g., tree), chart visualization (e.g., bar/line chat), 3D visualization, etc. Below, in Table 7, we provide a list of various visualization styles along with a few of the available tools in each category.

Table 7. Data visualization style with exemplary tools.

Visualization Style	Tool [Reference]
Text marking/highlighting	cite2vec [71], TopicLens [72], SurVis [73], Poemage [74], Overview [75]
Tags or word cloud	SentenTree [76], InfoVis [77], VisOHC [78], IncreSTS [79], Word storms [80]
Bar charts	TextTile [81], SentiCompass [82], NewsViews [83], WeiboEvents [84], CatStream [85]
Scatterplot	PhenoLines [86], SocialBrands [87], TopicPanorama [88], #FluxFlow [89], PEARL [90]
Line chart	Vispubdata.org [91], GameFlow [92], MultiConVis [93], Contextifier [94], Google+Ripples [95]
Node-link	NEREx [96], iForum [97], NameClarifier [98], DIA2 [99], Information Cartography [100]
Tree	OpinionFlow [101], Rule-based Visual Mappings [102], HierarchicalTopics [103], Whisper [104], The World's Languages Explorer [105]
Matrix	Interactive Ambiguity Resolution [106], Fingerprint Matrices [107], Conceptual recurrence plots [108], The Deshredder [109], Termite [110]
Stream graph timeline	VAiRoma [111], CiteRivers [112], ThemeDelta [113], EvoRiver [114], LeadLine [115]
Flow timeline	TimeLineCurator [116], Interactive visual profiling [117]
Radial visualization	ConToVi [118], ConVis [119]
3D visualization	Two-stage Framework [120]
Maps/Geo chart	Can Twitter save lives? [121], Visualizing Dynamic Data with Maps [122], Spatiotemporal Anomaly Detection [123]

Besides these tools, there are software available with gigantic capabilities to visualize the data, such as, Microsoft Excel's PivotTables, R, Tableau, Power-BI, datawrapper, and Google Charts. These tools are easy to use and very helpful in creating a clear and dynamic display of data because of their interactive graphical interface. Furthermore, different libraries written in different programming languages are also available for data visualization, which are easy to use for programmers, such as JavaScript libraries (e.g., D3.js, Chart.js, and Highcharts), python libraries (e.g., Matplotlib, Seaborn, and Plotly), and R libraries (e.g., ggplot2, Leaflet, and Esquisse). The major challenges of data visualization are the massive amount of data, the complexity of data, and missing/duplicate entries [124].

2.7. Deployment

In the deployment phase (sixth and final phase of CRISP-DM, Shearer [14]), the knowledge gained from the project is organized and presented (e.g., live demonstrations) in a way that is useful for the project, the company, and the customer. This phase's complexity varies greatly. The steps in the deployment phase are as follows: (1) Create a deployment plan to formulate and note a deployment strategy for the model, (2) plan the monitoring and maintenance to create well-thought-out planning of maintenance and monitoring to shun problems during the operational phase of a model, (3) produce a final report to prepare and present a final report of the project in the form of a written document and verbal meeting, and (4) review the project to evaluate successes and failures, as well as potential areas for improvement in future projects.

3. Conclusions and Future Outlook

The amount of medical text data is rapidly increasing. From medical text data, data mining can be used to extract new and useful information or knowledge. The CRISP-DM system presented in this study focuses on each step of data mining while using medical examples to explain each step. The authors plan to develop an artificial intelligence-based web crawling system with 4D visualization of the data in a summarized and easy-to-understand manner and use these data as a source of information for researchers, as well as for the education of patients and medical staff in future work.

Author Contributions: Conceptualization, A.Z., M.A., I.P., S.A.K., A.F.H. and A.R.A.; writing—original draft preparation, A.Z., M.A., I.P. and A.R.A.; writing—review and editing, A.Z., M.A. and A.R.A.; Artificial Intelligence and Machine Learning in Medical Literature Data Mining, M.A., S.A.K. and A.F.H.; Natural Language Processing, A.Z., S.A.K. and A.F.H.; funding acquisition, M.A. and A.R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was Funded by the German Bundesministerium für Bildung und Forschung (BMBF) under the CELTIC-NEXT AI-NET-PROTECT project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors acknowledge the support by the German Research Foundation (DFG) and the Open Access Publication Fund of the University of Göttingen.

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

References

1. Sumathy, K.L.; Chidambaram, M. Text Mining: Concepts, Applications, Tools and Issues—An Overview. *Int. J. Comput. Appl.* **2013**, *80*, 29–32. [CrossRef]
2. Cios, K.J.; Moore, G.W. Uniqueness of medical data mining. *Artif. Intell. Med.* **2002**, *26*, 1–24. [CrossRef]
3. Yang, Y.; Li, R.; Xiang, Y.; Lin, D.; Yan, A.; Chen, W.; Li, Z.; Lai, W.; Wu, X.; Wan, C.; et al. Standardization of Collection, Storage, Annotation, and Management of Data Related to Medical Artificial Intelligence. *Intell. Med.* **2021**. [CrossRef]
4. Thorpe, J.H.; Gray, E.A. Big data and public health: Navigating privacy laws to maximize potential. *Public Health Rep.* **2015**, *130*, 171–175. [CrossRef]
5. McGuire, A.L.; Beskow, L.M. Informed consent in genomics and genetic research. *Annu. Rev. Genom. Hum. Genet.* **2010**, *11*, 361–381. [CrossRef]
6. Tayefi, M.; Ngo, P.; Chomutare, T.; Dalianis, H.; Salvi, E.; Budrionis, A.; Godtliebsen, F. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comp. Stat.* **2021**, *13*, 1–19. [CrossRef]
7. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann Publisher: Waltham, MA, USA, 2011; ISBN 978-0-12-381479-1.
8. Locke, S.; Bashall, A.; Al-Adely, S.; Moore, J.; Wilson, A.; Kitchen, G.B. Natural language processing in medicine: A review. *Trends Anaesth. Crit. Care* **2021**, *38*, 4–9. [CrossRef]
9. Vyas, A. Top 14 Use Cases of Natural Language Processing in Healthcare. 6 July 2019. Available online: https://marutitech.com/use-cases-of-natural-language-processing-in-healthcare/ (accessed on 29 July 2022).

10. Liu, Z.; Yang, M.; Wang, X.; Chen, Q.; Tang, B.; Wang, Z.; Xu, H. Entity recognition from clinical texts via recurrent neural network. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 67. [CrossRef]
11. Deng, Y.; Faulstich, L.; Denecke, K. Concept Embedding for Relevance Detection of Search Queries Regarding CHOP. *Stud. Health Technol. Inform.* **2017**, *245*, 1260.
12. Afzal, M.; Hussain, M.; Malik, K.M.; Lee, S. Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence from Biomedical Literature: Empirical Study. *JMIR Med. Inform.* **2019**, *7*, e13430. [CrossRef]
13. Pandey, B.; Kumar Pandey, D.; Pratap Mishra, B.; Rhmann, W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *34*, 5083–5099. [CrossRef]
14. Shearer, C. The CRISP-DM Model: The New Blueprint for Data Mining. *Int. J. Data Warehous.* **2000**, *5*, 13–22.
15. Costa, C.J.; Aparicio, J.T. POST-DS: A Methodology to Boost Data Science. In Proceedings of the 15th Iberian Conference on Information Systems and Technologies (CISTI), Sevilla, Spain, 24–27 June 2020; pp. 1–6, ISBN 978-989-54659-0-3.
16. Catley, C.; Smith, K.; McGregor, C.; Tracy, M. Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. In Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems, Albuquerque, NM, USA, 2–5 August 2009; pp. 1–5, ISBN 978-1-4244-4878-4.
17. Data Science Process Alliance. What Is CRISP DM? Available online: https://www.datascience-pm.com/crisp-dm-2/ (accessed on 16 April 2022).
18. Martins, B.; Ferreira, D.; Neto, C.; Abelha, A.; Machado, J. Data Mining for Cardiovascular Disease Prediction. *J. Med. Syst.* **2021**, *45*, 6. [CrossRef]
19. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2013**, *41*, D8–D20. [CrossRef]
20. Guo, C.X.; He, L.; Yin, J.Y.; Meng, X.G.; Tan, W.; Yang, G.P.; Bo, T.; Liu, J.P.; Lin, X.J.; Chen, X. Epidemiological and clinical features of pediatric COVID-19. *BMC Med.* **2020**, *18*, 250. [CrossRef] [PubMed]
21. Miuțescu, A. Web Scraping vs. Web Crawling: Understand the Difference. WebScrapingAPI [Online]. 7 January 2021. Available online: https://www.webscrapingapi.com/web-scraping-vs-web-crawling/ (accessed on 19 April 2022).
22. Octoparse. What Is Web Scraping—Basics & Practical Uses—DataDrivenInvestor. DataDrivenInvestor [Online]. 25 January 2022. Available online: https://medium.datadriveninvestor.com/what-is-web-scraping-basics-practical-uses-66e1063cfa74 (accessed on 19 April 2022).
23. Batsakis, S.; Petrakis, E.G.; Milios, E. Improving the performance of focused web crawlers. *Data Knowl. Eng.* **2009**, *68*, 1001–1013. [CrossRef]
24. Yuan, X.; MacGregor, M.H.; Harms, J. An efficient scheme to remove crawler traffic from the Internet. In Proceedings of the Eleventh International Conference on Computer Communications and Networks. Eleventh International Conference on Computer Communications and Networks, Miami, FL, USA, 14–16 October 2002; pp. 90–95, ISBN 0-7803-7553-X.
25. DeVito, N.J.; Richards, G.C.; Inglesby, P. How we learnt to stop worrying and love web scraping. *Nature* **2020**, *585*, 621–622. [CrossRef]
26. Kaur, A.; Chopra, D. Comparison of text mining tools. In Proceedings of the 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 9 July–9 September 2016; pp. 186–192, ISBN 978-1-5090-1489-7.
27. Chandra, R.V.; Varanasi, B.S. *Python Requests Essentials: Learn How to Integrate Your Applications Seamlessly with Web Services Using Python Requests*; Packt: Birmingham, UK; Mumbai, India, 2015; ISBN 9781784395414.
28. Kouzis-Loukas, D. *Learning Scrapy: Learn the Art of Efficient Web Scraping and Crawling with Python*; Packt: Birmingham, UK, 2016; ISBN 9781784390914.
29. Richardson, L. Beautiful Soup Documentation. Available online: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed on 16 April 2022).
30. Sharma, P.R. *Selenium with Python: A Beginner's Guide*; BPB: Delhi, India, 2019; ISBN 9789389328820.
31. Gu, D.; Li, J.; Li, X.; Liang, C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int. J. Med. Inform.* **2017**, *98*, 22–32. [CrossRef]
32. Ristevski, B.; Chen, M. Big Data Analytics in Medicine and Healthcare. *J. Integr. Bioinform.* **2018**, *15*, 1–5. [CrossRef]
33. Giffen, C.A.; Carroll, L.E.; Adams, J.T.; Brennan, S.P.; Coady, S.A.; Wagner, E.L. Providing Contemporary Access to Historical Biospecimen Collections: Development of the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). *Biopreserv. Biobank.* **2015**, *13*, 271–279. [CrossRef]
34. Cimino, J.J.; Ayres, E.J.; Remennik, L.; Rath, S.; Freedman, R.; Beri, A.; Chen, Y.; Huser, V. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): Design, contents, functionality and experience to date. *J. Biomed. Inform.* **2014**, *52*, 11–27. [CrossRef] [PubMed]
35. Mayo-Wilson, E.; Doshi, P.; Dickersin, K. Are manufacturers sharing data as promised? *BMJ* **2015**, *351*, h4169. [CrossRef] [PubMed]
36. Doll, K.M.; Rademaker, A.; Sosa, J.A. Practical Guide to Surgical Data Sets: Surveillance, Epidemiology, and End Results (SEER) Database. *JAMA Surg.* **2018**, *153*, 588–589. [CrossRef] [PubMed]

37. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Lehman, L.W.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef]
38. Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L.A.; Mark, R. *MIMIC-IV*; Version 1.0; PhysioNet: Cambridge, MA, USA, 2021. [CrossRef]
39. Johnson, A.E.W.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.Y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef]
40. Ahluwalia, N.; Dwyer, J.; Terry, A.; Moshfegh, A.; Johnson, C. Update on NHANES Dietary Data: Focus on Collection, Release, Analytical Considerations, and Uses to Inform Public Policy. *Adv. Nutr.* **2016**, *7*, 121–134. [CrossRef]
41. Vos, T.; Lim, S.S.; Abbafati, C.; Abbas, K.M.; Abbasi, M.; Abbasifard, M.; Abbasi-Kangevari, M.; Abbastabar, H.; Abd-Allah, F.; Abdelalim, A.; et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **2020**, *396*, 1204–1222. [CrossRef]
42. Palmer, L.J. UK Biobank: Bank on it. *Lancet* **2007**, *369*, 1980–1982. [CrossRef]
43. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef]
44. Davis, S.; Meltzer, P.S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847. [CrossRef]
45. Woolley, C.S.C.; Handel, I.G.; Bronsvoort, B.M.; Schoenebeck, J.J.; Clements, D.N. Is it time to stop sweeping data cleaning under the carpet? A novel algorithm for outlier management in growth data. *PLoS ONE* **2020**, *15*, e0228154. [CrossRef] [PubMed]
46. Coupler.io Blog. Data Cleansing vs. Data Transformation | Coupler.io Blog. Available online: https://blog.coupler.io/data-cleansing-vs-data-transformation/#What_is_data_transformation (accessed on 24 June 2022).
47. Elgabry, O. The Ultimate Guide to Data Cleaning—Towards Data Science. 28 February 2019. Available online: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4 (accessed on 24 June 2022).
48. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
49. McKinney, W. Data structures for statistical computing in python. In Proceedings of the Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 51–56.
50. Gordon, B.; Fennessy, C.; Varma, S.; Barrett, J.; McCondochie, E.; Heritage, T.; Duroe, O.; Jeffery, R.; Rajamani, V.; Earlam, K.; et al. Evaluation of freely available data profiling tools for health data research application: A functional evaluation review. *BMJ Open* **2022**, *12*, e054186. [CrossRef] [PubMed]
51. Lopez, P. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries*; Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 473–474, ISBN 978-3-642-04345-1.
52. Lo, K.; Wang, L.L.; Neumann, M.; Kinney, R.; Weld, D. S2ORC: The Semantic Scholar Open Research Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4969–4983.
53. Comeau, D.C.; Wei, C.H.; Doğan, R.I.; Lu, Z. PMC text mining subset in BioC: 2.3 million full text articles and growing. *arXiv* **2018**, arXiv:1804.05957. [CrossRef]
54. Rawat, T.; Khemchandani, V. Feature engineering (FE) tools and techniques for better classification performance. *Int. J. Innov. Eng. Technol.* **2017**, *8*, 169–179. [CrossRef]
55. Heaton, J. An empirical analysis of feature engineering for predictive modeling. In Proceedings of the SoutheastCon 2016, Norfolk, VA, USA, 30 March–3 April 2016; IEEE: Manhattan, NY, USA, 2016; pp. 1–6, ISBN 978-1-5090-2246-5.
56. Vijithananda, S.M.; Jayatilake, M.L.; Hewavithana, B.; Gonçalves, T.; Rato, L.M.; Weerakoon, B.S.; Kalupahana, T.D.; Silva, A.D.; Dissanayake, K.D. Feature extraction from MRI ADC images for brain tumor classification using machine learning techniques. *BioMed Eng. OnLine* **2022**, *21*, 52. [CrossRef]
57. Rus, A. Keyword-Recherche: Die richtigen Keywords Finden Leicht Gemacht. Evergreen Media AR GmbH. 7 September 2021. Available online: https://www.evergreenmedia.at/ratgeber/keyword-recherche/ (accessed on 20 April 2022).
58. Singh, V. Replace or Retrieve Keywords in Documents at Scale. 2017. Available online: https://arxiv.org/pdf/1711.00046 (accessed on 19 April 2022).
59. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python*; O'Reilly: Beijing, China; Farnham, UK, 2009; ISBN 9780596516499.
60. Honnibal, M. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. Sentometrics Research. 1 January 2017. Available online: https://sentometrics-research.com/publication/72/ (accessed on 19 April 2022).
61. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-learn: Machine Learning in Python. *JMLR* **2011**, *12*, 2825–2830.
62. Pinto, A.; Oliveira, H.G.; Alves, A.O. Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text. In Proceedings of the 5th Symposium on Languages, Applications and Technologies (SLATE'16), Maribor, Slovenia, 20–21 June 2016; Mernik, M., Leal, J.P., Oliveira, H.G., Eds.; Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik GmbH: Wadern/Saarbruecken, Germany, 2016; ISBN 978-3-95977-006-4.

63. Rehurek, R.; Sojka, P. Gensim-python framework for vector space modelling. *NLP Cent. Fac. Inform. Masaryk. Univ. Brno Czech Repub.* **2011**, *3*, 2.
64. Nadif, M.; Role, F. Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Brief. Bioinform.* **2021**, *22*, 1592–1603. [CrossRef]
65. Wu, W.T.; Li, Y.J.; Feng, A.Z.; Li, L.; Huang, T.; Xu, A.D.; Lyu, J. Data mining in clinical big data: The frequently used databases, steps, and methodological models. *Mil. Med. Res.* **2021**, *8*, 44. [CrossRef]
66. Berry, M.W. *Supervised and Unsupervised Learning for Data Science*; Springer: Berlin/Heidelberg, Germany, 2020; ISBN 978-3-030-22474-5.
67. Zowalla, R.; Wetter, T.; Pfeifer, D. Crawling the German Health Web: Exploratory Study and Graph Analysis. *J. Med. Internet Res.* **2020**, *22*, e17853. [CrossRef] [PubMed]
68. Tsioptsias, N.; Tako, A.; Robinson, S. (Eds.) *Model Validation and Testing in Simulation: A Literature Review*; Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik GmbH: Wadern/Saarbruecken, Germany, 2016; p. 11.
69. Dong, H.; Liu, Y.; Zeng, W.-F.; Shu, K.; Zhu, Y.; Chang, C. A Deep Learning-Based Tumor Classifier Directly Using MS Raw Data. *Proteomics* **2020**, *20*, e1900344. [CrossRef] [PubMed]
70. OWOX. What Is Data Visualization: Definition, Examples, Principles, Tools. Available online: https://www.owox.com/blog/articles/data-visualization/ (accessed on 12 April 2022).
71. Berger, M.; McDonough, K.; Seversky, L.M. cite2vec: Citation-Driven Document Exploration via Word Embeddings. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 691–700. [CrossRef] [PubMed]
72. Kim, M.; Kang, K.; Park, D.; Choo, J.; Elmqvist, N. TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 151–160. [CrossRef]
73. Beck, F.; Koch, S.; Weiskopf, D. Visual Analysis and Dissemination of Scientific Literature Collections with SurVis. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 180–189. [CrossRef]
74. McCurdy, N.; Lein, J.; Coles, K.; Meyer, M. Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 439–448. [CrossRef]
75. Brehmer, M.; Ingram, S.; Stray, J.; Munzner, T. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 2271–2280. [CrossRef]
76. Hu, M.; Wongsuphasawat, K.; Stasko, J. Visualizing Social Media Content with SentenTree. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 621–630. [CrossRef]
77. Hinrichs, U.; Forlini, S.; Moynihan, B. Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 429–438. [CrossRef]
78. Kwon, B.C.; Kim, S.-H.; Lee, S.; Choo, J.; Huh, J.; Yi, J.S. VisOHC: Designing Visual Analytics for Online Health Communities. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 71–80. [CrossRef]
79. Liu, C.-Y.; Chen, M.-S.; Tseng, C.-Y. IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 2986–3000. [CrossRef]
80. Castellà, Q.; Sutton, C. Word storms: Multiples of word clouds for visual comparison of documents. In Proceedings of the 23rd International Conference on World Wide Web—WWW '14, Seoul, Korea, 7–11 April 2014; Chung, C.-W., Broder, A., Shim, K., Suel, T., Eds.; ACM Press: New York, NY, USA, 2014; pp. 665–676, ISBN 9781450327442.
81. Felix, C.; Pandey, A.V.; Bertini, E. TextTile: An Interactive Visualization Tool for Seamless Exploratory Analysis of Structured Data and Unstructured Text. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 161–170. [CrossRef] [PubMed]
82. Wang, F.Y.; Sallaberry, A.; Klein, K.; Takatsuka, M.; Roche, M. SentiCompass: Interactive visualization for exploring and comparing the sentiments of time-varying twitter data. In Proceedings of the 2015 IEEE Pacific Visualization Symposium (PacificVis), Hangzhou, China, 14–17 April 2015; pp. 129–133, ISBN 978-1-4673-6879-7.
83. Gao, T.; Hullman, J.R.; Adar, E.; Hecht, B.; Diakopoulos, N. NewsViews: An automated pipeline for creating custom geovisualizations for news. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 26 April–1 May 2014; Jones, M., Palanque, P., Schmidt, A., Grossman, T., Eds.; ACM: New York, NY, USA, 2014; pp. 3005–3014, ISBN 9781450324731.
84. Ren, D.; Zhang, X.; Wang, Z.; Li, J.; Yuan, X. WeiboEvents: A Crowd Sourcing Weibo Visual Analytic System. In Proceedings of the 2014 IEEE Pacific Visualization Symposium (PacificVis), Yokohama, Japan, 4–7 March 2014; pp. 330–334, ISBN 978-1-4799-2874-3.
85. Garcia Esparza, S.; O'Mahony, M.P.; Smyth, B. CatStream: Categorising tweets for user profiling and stream filtering. In Proceedings of the 2013 International Conference on Intelligent User Interfaces—IUI '13, Santa Monica, CL, USA, 19–22 March 2013; Kim, J., Nichols, J., Szekely, P., Eds.; ACM Press: New York, NY, USA, 2013; p. 25, ISBN 9781450319652.
86. Glueck, M.; Naeini, M.P.; Doshi-Velez, F.; Chevalier, F.; Khan, A.; Wigdor, D.; Brudno, M. PhenoLines: Phenotype Comparison Visualizations for Disease Subtyping via Topic Models. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 371–381. [CrossRef] [PubMed]
87. Liu, X.; Xu, A.; Gou, L.; Liu, H.; Akkiraju, R.; Shen, H.-W. SocialBrands: Visual analysis of public perceptions of brands on social media. In Proceedings of the 2016 IEEE Conference on Visual Analytics Science and Technology (VAST), Baltimore, MD, USA, 23–28 October 2016; pp. 71–80, ISBN 978-1-5090-5661-3.
88. Wang, X.; Liu, S.; Liu, J.; Chen, J.; Zhu, J.; Guo, B. TopicPanorama: A Full Picture of Relevant Topics. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 2508–2521. [CrossRef]
89. Zhao, J.; Cao, N.; Wen, Z.; Song, Y.; Lin, Y.-R.; Collins, C. #FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1773–1782. [CrossRef]

90. Zhao, J.; Gou, L.; Wang, F.; Zhou, M. PEARL: An interactive visual analytic tool for understanding personal emotion style derived from social media. In Proceedings of the 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), Paris, France, 25–31 October 2014; pp. 203–212, ISBN 978-1-4799-6227-3.
91. Isenberg, P.; Heimerl, F.; Koch, S.; Isenberg, T.; Xu, P.; Stolper, C.D.; Sedlmair, M.; Chen, J.; Moller, T.; Stasko, J. Vispubdata.org: A Metadata Collection About IEEE Visualization (VIS) Publications. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 2199–2206. [CrossRef]
92. Chen, W.; Lao, T.; Xia, J.; Huang, X.; Zhu, B.; Hu, W.; Guan, H. GameFlow: Narrative Visualization of NBA Basketball Games. *IEEE Trans. Multimed.* **2016**, *18*, 2247–2256. [CrossRef]
93. Hoque, E.; Carenini, G. MultiConVis: A Visual Text Analytics System for Exploring a Collection of Online Conversations. In Proceedings of the 21st International Conference on Intelligent User Interfaces, Sonoma, CL, USA, 7–10 March 2016; Nichols, J., Mahmud, J., O'Donovan, J., Conati, C., Zancanaro, M., Eds.; ACM: New York, NY, USA, 2016; pp. 96–107, ISBN 9781450341370.
94. Hullman, J.; Diakopoulos, N.; Adar, E. Contextifier: Automatic Generation of Annotated Stock Visualizations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; Mackay, W.E., Brewster, S., Bødker, S., Eds.; ACM: New York, NY, USA, 2013; pp. 2707–2716, ISBN 9781450318990.
95. Viégas, F.; Wattenberg, M.; Hebert, J.; Borggaard, G.; Cichowlas, A.; Feinberg, J.; Orwant, J.; Wren, C. Google + Ripples: A Native Visualization of Information Flow. In Proceedings of the 22nd International Conference on World Wide Web—WWW '13, Rio de Janeiro, Brazil, 13–17 May 2013; Schwabe, D., Almeida, V., Glaser, H., Baeza-Yates, R., Moon, S., Eds.; ACM: New York, NY, USA, 2013; pp. 1389–1398, ISBN 9781450320351.
96. El-Assady, M.; Sevastjanova, R.; Gipp, B.; Keim, D.; Collins, C. NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations. *Comput. Graph. Forum* **2017**, *36*, 213–225. [CrossRef]
97. Fu, S.; Zhao, J.; Cui, W.; Qu, H. Visual Analysis of MOOC Forums with iForum. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 201–210. [CrossRef]
98. Shen, Q.; Wu, T.; Yang, H.; Wu, Y.; Qu, H.; Cui, W. NameClarifier: A Visual Analytics System for Author Name Disambiguation. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 141–150. [CrossRef]
99. Madhavan, K.; Elmqvist, N.; Vorvoreanu, M.; Chen, X.; Wong, Y.; Xian, H.; Dong, Z.; Johri, A. DIA2: Web-based Cyberinfrastructure for Visual Analysis of Funding Portfolios. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1823–1832. [CrossRef] [PubMed]
100. Shahaf, D.; Yang, J.; Suen, C.; Jacobs, J.; Wang, H.; Leskovec, J. Information cartography: Creating Zoomable, Large-Scale Maps of Information. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; Ghani, R., Senator, T.E., Bradley, P., Parekh, R., He, J., Grossman, R.L., Uthurusamy, R., Dhillon, I.S., Koren, Y., Eds.; ACM: New York, NY, USA, 2013; pp. 1097–1105, ISBN 9781450321747.
101. Wu, Y.; Liu, S.; Yan, K.; Liu, M.; Wu, F. OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1763–1772. [CrossRef] [PubMed]
102. Abdul-Rahman, A.; Lein, J.; Coles, K.; Maguire, E.; Meyer, M.; Wynne, M.; Johnson, C.R.; Trefethen, A.; Chen, M. Rule-based Visual Mappings—with a Case Study on Poetry Visualization. *Comput. Graph. Forum* **2013**, *32*, 381–390. [CrossRef]
103. Dou, W.; Yu, L.; Wang, X.; Ma, Z.; Ribarsky, W. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2002–2011. [CrossRef]
104. Cao, N.; Lin, Y.R.; Sun, X.; Lazer, D.; Liu, S.; Qu, H. Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time. *IEEE Trans. Vis. Comput. Graph.* **2012**, *18*, 2649–2658. [CrossRef]
105. Rohrdantz, C.; Hund, M.; Mayer, T.; Wälchli, B.; Keim, D.A. The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts. *Comput. Graph. Forum* **2012**, *31*, 935–944. [CrossRef]
106. Stoffel, F.; Jentner, W.; Behrisch, M.; Fuchs, J.; Keim, D. Interactive Ambiguity Resolution of Named Entities in Fictional Literature. *Comput. Graph. Forum* **2017**, *36*, 189–200. [CrossRef]
107. Oelke, D.; Kokkinakis, D.; Keim, D.A. Fingerprint Matrices: Uncovering the dynamics of social networks in prose literature. *Comput. Graph. Forum* **2013**, *32*, 371–380. [CrossRef]
108. Angus, D.; Smith, A.; Wiles, J. Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Trans. Vis. Comput. Graph.* **2012**, *18*, 988–997. [CrossRef]
109. Butler, P.; Chakraborty, P.; Ramakrishan, N. The Deshredder: A visual analytic approach to reconstructing shredded documents. In Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), Seattle, WA, USA, 14–19 October 2012; pp. 113–122, ISBN 978-1-4673-4753-2.
110. Chuang, J.; Manning, C.D.; Heer, J. Termite: Visualization techniques for assessing textual topic models. In Proceedings of the International Working Conference on Advanced Visual Interfaces—AVI '12, Capri Island, Naples, Italy, 21–25 May 2012; Tortora, G., Levialdi, S., Tucci, M., Eds.; ACM: New York, NY, USA, 2012; p. 74, ISBN 9781450312875.
111. Cho, I.; Dou, W.; Wang, D.X.; Sauda, E.; Ribarsky, W. VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 210–219. [CrossRef]
112. Heimerl, F.; Han, Q.; Koch, S.; Ertl, T. CiteRivers: Visual Analytics of Citation Patterns. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 190–199. [CrossRef] [PubMed]
113. Gad, S.; Javed, W.; Ghani, S.; Elmqvist, N.; Ewing, T.; Hampton, K.N.; Ramakrishnan, N. ThemeDelta: Dynamic Segmentations over Temporal Topic Models. *IEEE Trans. Vis. Comput. Graph.* **2015**, *21*, 672–685. [CrossRef] [PubMed]

114. Sun, G.; Wu, Y.; Liu, S.; Peng, T.-Q.; Zhu, J.J.H.; Liang, R. EvoRiver: Visual Analysis of Topic Coopetition on Social Media. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1753–1762. [CrossRef] [PubMed]
115. Dou, W.; Wang, X.; Skau, D.; Ribarsky, W.; Zhou, M.X. LeadLine: Interactive visual analysis of text data through event identification and exploration. In Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), Seattle, WA, USA, 14–19 October 2012; pp. 93–102, ISBN 978-1-4673-4753-2.
116. Fulda, J.; Brehmel, M.; Munzner, T. TimeLineCurator: Interactive Authoring of Visual Timelines from Unstructured Text. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 300–309. [CrossRef]
117. Janicke, S.; Focht, J.; Scheuermann, G. Interactive Visual Profiling of Musicians. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 200–209. [CrossRef]
118. El-Assady, M.; Gold, V.; Acevedo, C.; Collins, C.; Keim, D. ConToVi: Multi-Party Conversation Exploration using Topic-Space Views. *Comput. Graph. Forum* **2016**, *35*, 431–440. [CrossRef]
119. Hoque, E.; Carenini, G. ConVis: A Visual Text Analytic System for Exploring Blog Conversations. *Comput. Graph. Forum* **2014**, *33*, 221–230. [CrossRef]
120. Oesterling, P.; Scheuermann, G.; Teresniak, S.; Heyer, G.; Koch, S.; Ertl, T.; Weber, G.H. Two-stage framework for a topology-based projection and visualization of classified document collections. In Proceedings of the 2010 IEEE Symposium on Visual Analytics Science and Technology (VAST), Salt Lake City, UT, USA, 25–26 October 2010; pp. 91–98, ISBN 978-1-4244-9488-0.
121. Thom, D.; Kruger, R.; Ertl, T. Can Twitter Save Lives? A Broad-Scale Study on Visual Social Media Analytics for Public Safety. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 1816–1829. [CrossRef]
122. Mashima, D.; Kobourov, S.G.; Hu, Y. Visualizing Dynamic Data with Maps. *IEEE Trans. Vis. Comput. Graph.* **2012**, *18*, 1424–1437. [CrossRef] [PubMed]
123. Thom, D.; Bosch, H.; Koch, S.; Worner, M.; Ertl, T. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In Proceedings of the 2012 IEEE Pacific Visualization Symposium (PacificVis), Songdo, Korea, 28 February–2 March 2012; pp. 41–48, ISBN 978-1-4673-0866-3.
124. Siddiqui, A.T. Data Visualization: A Study of Tools and Challenges. *Asian J. Technol. Manag. Res.* **2021**, *11*, 18–23.