applied sciences

MDPI

Review

Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review

Ganesh Kumar [1,*], Shuib Basri [1], Abdullahi Abubakar Imam [1,2], Sunder Ali Khowaja [3], Luiz Fernando Capretz [4] and Abdullateef Oluwagbemiga Balogun [1,5]

[1] Computer and Information Science Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar 32610, Perak, Malaysia; shuib_basri@utp.edu.my (S.B.); abdullahi_g03618@utp.edu.my (A.A.I.); abdullateef_16005851@utp.edu.my (A.O.B.)
[2] Department of Computer Science, Ahmadu Bello University, Zaria 1044, Nigeria
[3] Department of Telecommunication Engineering, Faculty of Engineering and Technology, University of Sindh, Jamshoro 76090, Pakistan; sandar.ali@usindh.edu.pk
[4] Department of Electrical and Computer Engineering, Western University, London, ON N6A 5B9, Canada; lcapretz@uwo.ca
[5] Department of Computer Science, University of Ilorin, Ilorin 1515, Nigeria
* Correspondence: ganesh_17005106@utp.edu.my

Abstract: As data size increases drastically, its variety also increases. Investigating such heterogeneous data is one of the most challenging tasks in information management and data analytics. The heterogeneity and decentralization of data sources affect data visualization and prediction, thereby influencing analytical results accordingly. Data harmonization (DH) corresponds to a field that unifies the representation of such a disparate nature of data. Over the years, multiple solutions have been developed to minimize the heterogeneity aspects and disparity in formats of big-data types. In this study, a systematic review of the literature was conducted to assess the state-of-the-art DH techniques. This study aimed to understand the issues faced due to heterogeneity, the need for DH and the techniques that deal with substantial heterogeneous textual datasets. The process produced 1355 articles, but among them, only 70 articles were found to be relevant through inclusion and exclusion criteria methods. The result shows that the heterogeneity of structured, semi-structured, and unstructured (SSU) data can be managed by using DH and its core techniques, such as text preprocessing, Natural Language Preprocessing (NLP), machine learning (ML), and deep learning (DL). These techniques are applied to many real-world applications centered on the information-retrieval domain. Several assessment criteria were implemented to measure the efficiency of these techniques, such as precision, recall, F-1, accuracy, and time. A detailed explanation of each research question, common techniques, and performance measures is also discussed. Lastly, we present readers with a detailed discussion of the existing work, contributions, and managerial and academic implications, along with the conclusion, limitations, and future research directions.

Keywords: data harmonization; heterogeneous data; text preprocessing

Citation: Kumar, G.; Basri, S.; Imam, A.A.; Khowaja, S.A.; Capretz L.F.; Balogun, A.O. Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review. *Appl. Sci.* **2021**, *11*, 8275. https://doi.org/10.3390/app11178275

Academic Editor: Manuel Armada

Received: 17 August 2021
Accepted: 30 August 2021
Published: 6 September 2021

1. Introduction

Big Data play a vital role in the assessment of massive data produced every second by real-world applications, using tools and algorithms [1]. Some of real-life application domains of Big Data are healthcare, telecommunication, financial firms, retail, law enforcement, marketing, new product development, banking, energy and utilities, insurance, education, agriculture, and urban planning, as discussed in Reference [2]. Nowadays, data are being produced in various formats, ranging from structured and semi-structured to unstructured (SSU) generated from heterogeneous resources [3,4]. The disparate nature of data cannot be processed with simple tools and techniques [2,5], and this creates a challenge for decision-makers to make decisions based on the scattered data. Emerging technologies, such as the Internet of Things (IoT), Industry 4.0 (I4.0), and extended reality (XR), produce distinct kinds of information via heterogeneous sources and real-world applications that create heterogeneity issues [6], in IoT integration, security, analytics challenges, and computational time [7–9]. Among them, data harmonization (DH), which describes the uniform representation of heterogeneous data, was proposed in References [10,11].

IoT is a system that deals with interrelated computing objects, such as unique tags, RFID, or machine interactions, and that can transfer data without human and machine involvement [12]. As technology evolves, the IoT has further grown into the Industrial IoT (IIoT), which deals with heterogeneous data produced by real-world applications, industrial products, and devices, such as privacy authentication logs of IIoT devices [13], business architecture devices data [14], and heterogeneous IIoT devices data [6]. In addition, I4.0 deals with IoT-based automation, technologies, and decision-making that help decision-makers to make decisions based on the disparate nature of data produced [15]. Applications of I4.0 in higher education, predictive maintenance [16,17], food logistics [18], knowledge management [19], business [20], and supply chain [21]. The main problem faced by these applications is related to managing the heterogeneous data produced in bulk by employing I4.0 and IIoT. The data produced by industries include digital data for manufacturing purposes, unstructured data for predictive maintenance, customer data for food logistics, customer reviews for knowledge management, business data for the supply chain, and manufacturing data for the supply chain. XR deals with the real and virtual environment with the help of a machine and human interaction [22]. XR is improving heterogeneous manufacturing data in the digital world. The tools must be advanced so that user acceptance and better usability of products are achieved [23]. AI can be effectively used to address the disparate nature of manufacturing data to deliver the best appearance to the XR industry [24].

To resolve the problems mentioned earlier, the disparity of data needs to be reviewed in detail, so that data harmonization models, tools, techniques, algorithms, and their performance can be evaluated for extensive heterogeneous textual information. Although related work was carried out in multimodalities for text, image, audio, and video [25–27], there were no such studies highlighting the work associated with textual data, data harmonization core techniques, and performance measurement. Multiple studies have been conducted which deal with applications such as sentiment analysis, text similarity, word embedding, and emotion recognition in conjunction with the help of classification and clustering techniques. Therefore, solving real-world application problems, such as those of a medical and healthcare nature, needs data to be harmonized and uniformly presented, so that decisions can be carried out efficiently. Based on the needs and contributions of emerging technologies and real-world application domains, we aimed to conduct a systematic review of the literature that could demonstrate the heterogeneity issues faced by real-world applications, data harmonization as a solution architecture for the disparate nature of data, techniques that can deal with large textual heterogeneous datasets, and performance assessment of models.

In this SLR, we have proposed to solve the problems mentioned above by doing a systematic review of the literature on textual data domains that deal with heterogeneity. The disciplines added were data integration, curation, and harmonization and were performed and contributed by the research community in their novel ideas. In addition, we selected articles that deal with the core textual data techniques and performance measurement techniques. The state-of-the-art SLR contains heterogeneity issues, textual data harmonization, data-processing techniques, and models' performance measurement methods. The research questions were drawn to emphasize the domains that focus on the heterogeneity issue, how the data harmonization approach will help, the core techniques that can deal with sizeable textual data, and which algorithms are suitable based on efficiency. The objectives of the proposed SLR are to understand the issues of heterogeneity faced by industries; to mold the heterogeneity issue by replacing DH; and to keep an acceptable level of ML, DL, and NLP core techniques and performance measurement techniques for concerning sizeable textual data. This study will contribute towards Big Data variety and the data analytics research community that helps in representation, visualization, and prediction of the heterogeneous information produced in the disparate form.

This paper is organized as follows: In Section 2, an illustration of research methodology is presented that comprises three steps: planning, conducting, and reporting of the review. The results of selected articles, research questions, standard techniques, data formats, and performance techniques are discussed in Section 3. In Section 4, a discussion of existing solutions is presented, along with their contributions, managerial and academic implications, and the conclusion by including limitations, future research directions, and the scientific contribution of this review.

2. Research Methodology

In this Systematic Literature Review (SLR), the guidelines were followed from References [28,29]. The research process is divided into three phases. In the first planning phase, the stages of defining research questions, developing, and validating review protocols are covered. In the second phase, identification and selection of relevant studies, data extraction, and the information synthesis process are covered; and in the third phase, writing and validating the review are reported. Figure 1 illustrates the flow of three phases.

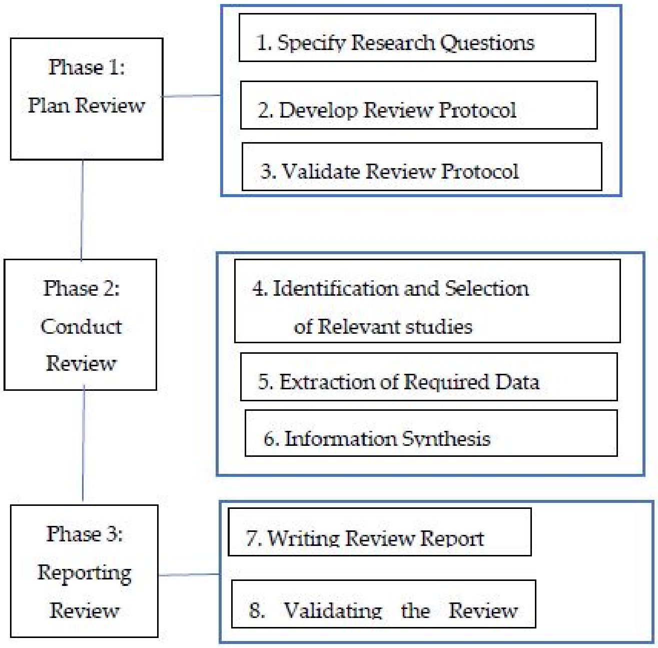

Figure 1. SLR process.

2.1. Plan Review

In this first phase of research methodology, the significant research questions and development of review protocols are specified with the proper searching strategy.

2.1.1. Research Questions

In this SLR, the following research questions are set, and possibly all questions are later answered with proper solutions.

1. RQ #1: Which of the domains are mainly focused on researching heterogeneity?

The motivation behind this research question is to find the significance of heterogeneity or heterogeneous data produced by industries, websites, and hospitals from models, frameworks, or applications. Heterogeneous data comprise structured, semi-structured, and unstructured, which is very tough to manage, and data are important for organizations, industries, and firms.

2. RQ #2: How does data harmonization resolve the issues of heterogeneity?

This research question is linked to heterogeneity because we need to check the solution to it. The motivation for this research question is to check the models, frameworks, and applications that manage the disparate nature of data with the latest tools and techniques. Moreover, a different strategy was adopted to gather relevant data by using terms such as "data integration", "data mapping", and "data fusion".

3. RQ #3: Which techniques are being used to solve the harmonization issue for large textual datasets?

The purpose of this research question is to identify the textual data and the techniques used for solving the issues of storing, managing, and uniform representation. Textual data can be used for semantic, syntactic, and schematic representation. Furthermore, it is used for predictive analysis and information retrieval. Thus, in-depth techniques for text data retrieval, data formats, and performance measures are later reviewed.

4. RQ #4: Which deep learning algorithms are well-suited with respect to efficiency for large sequential datasets

In this research question, the main target is to identify the algorithms' performance for sequential text data processing. The textual data are completely based on heterogeneous data, data harmonization, and techniques.

2.1.2. Review Protocols

The development and validation of the review protocol highlight the searching of related articles with the appropriate keywords and the literature sources.

2.1.3. Searching Keywords

To guarantee that the review closely covers data harmonization and relevant techniques for heterogeneous data, we tried to limit our search to the most relevant search term. Thus, we started with the keywords, and then we went through the following steps:

- Extracting the major distinct terms from our research questions;
- Using different spellings of the terms;
- Updating our search terms with keywords from relevant papers.

We used the main alternatives and added "OR operator" and "AND operator" to get the maximum amount of directly relevant works in the literature, as shown in Table 1.

Table 1. Inclusion and exclusion criteria description.

ID	Keywords
1	"Data Harmonization" AND ("Heterogeneous Data" OR "Heterogeneity") AND ("Textual Data" OR "Text Data") AND ("text Preprocessing "OR "Preprocessing")
2	"Data Harmonization" AND ("Heterogeneous Data" OR "Heterogeneity") AND ("Textual Data" OR "Text Data") AND ("Text Preprocessing "OR "Preprocessing") AND (Techniques OR Algorithm)
3	"Data Integration" AND ("Heterogeneous Data" OR "Heterogeneity") AND ("Textual Data" OR "Text Data") AND ("text Preprocessing "OR "Preprocessing")
4	"Data Integration" AND ("Heterogeneous Data" OR "Heterogeneity") AND ("Textual Data" OR "Text Data") AND ("text Preprocessing "OR "Preprocessing") AND (Techniques OR Algorithm)
5	"Data Fusion" AND ("Heterogeneous Data" OR "Heterogeneity") AND ("Textual Data" OR "Text Data") AND ("text Preprocessing "OR "Preprocessing")

6	“Data Fusion” AND (“Heterogeneous Data” OR “Heterogeneity”) AND (“Textual Data” OR “Text Data”) AND (“text Preprocessing “OR “Preprocessing”) AND (Techniques OR Algorithm)
7	(“Data Harmonization” OR “Data Integration” OR “Data Fusion”) AND (“Heterogeneous Data” OR “Heterogeneity”) AND (“Textual Data” OR “Text Data”) AND (“text Preprocessing “OR “Preprocessing”)
8	(“Data Harmonization” OR “Data Integration” OR “Data Fusion”) AND (“Heterogeneous Data” OR “Heterogeneity”) AND (“Textual Data” OR “Text Data”) AND (“text Preprocessing “OR “Preprocessing”) AND (Techniques OR Algorithm)
9	(“Data Harmonization” OR “Data Integration” OR “Data Fusion”) AND (“Heterogeneous Data” OR “Heterogeneity”) AND (“Textual Data” OR “Text Data”) AND (Techniques OR Algorithm)
10	(“Data Harmonization” OR “Data Integration” OR “Data Fusion”) AND (“Heterogeneous Data” OR “Heterogeneity”) AND (“Textual Data” OR “Text Data”)

2.1.4. Literature Resources

- Primary review studies: Web of Science, Scopus, ACM Digital Library, Springer, Science Direct, and IEEE Explorer databases were chosen for selection of relevant articles. These databases have maximum coverage of quality articles in our domain, such as ISI and Scopus indexed articles. The search term was constructed by using the advanced search features provided by each of these databases. Our search included the period from 2015 to 2020.

2.2. Conduct Review

In this phase, we conducted the review according to the research questions, keywords, and protocols. This phase mostly emphasizes the inclusion and exclusion of articles, according to Table 2a, 2b.

Table 2. (**a**) Inclusion criteria description and (**b**) exclusion criteria description.

Inclusion Criteria
The research was relevant to heterogeneous data sources.
The research was directly related to the data.
The research was related to text preprocessing and NLP applications.
The research used performance measurement techniques.
The research was conducted using ML, DL and NLP techniques related to textual data.
For duplicate publications of the same study, the newest and most complete one was selected. This is recorded for only one study whose related work appeared two times.
Exclusion Criteria
Studies that were irrelevant to data harmonization and domain were skipped. They showed up in our search due to the misuse of the term “harmonization” to describe traditional chemical and music work. Table A1 shows selected studies.

2.2.1. Study Selection

The whole process of study selection is illustrated in Figure 2. A total of 1355 articles appeared in the online search. By applying filtration with title, keyword, inclusion, and exclusion criteria, a total of 155 papers were short-listed. Inclusion and exclusion criteria are defined in Table 2a, 2b. Among them, 33 articles were repeated in other databases, and 22 articles were from different domains, such as chemistry, music, and other languages. At the end, 30 articles are removed from the list after going through full reading.

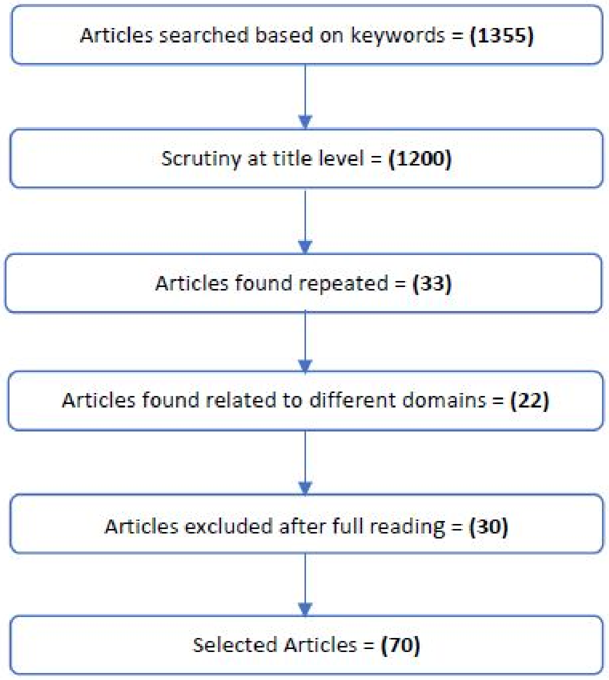
Articles searched based on keywords = (1355)
Scrutiny at title level = (1200)
Articles found repeated = (33)
Articles found related to different domains = (22)
Articles excluded after full reading = (30)
Selected Articles = (70)

Figure 2. Process of identifying relevant studies.

Table 2a, 2b describes the selection criteria of the relevant articles according to keywords. The duplicate articles and articles that do not cover all research questions are excluded.

Table 3 illustrates the quality checklist questions for evaluation of studies. The questions are mainly designed for selection of the studies that are more relevant, detailed, and that cover all research questions.

Table 3. Quality checklist.

No.	Questions
1	Did the studies focus on the heterogeneous nature of data?
2	Was the study explaining the harmonization of large textual data?
3	Was there any model proposed for textual data harmonization?
4	Is study focusing on the core techniques of ML, DL and NLP for large textual data?
5	Is the study discussing model performance using core techniques?

2.2.2. Data Extraction

In order to obtain the data which are needed to address our research questions and contributions, we used the data-extraction methods highlighted in Table 4.

Table 4. Data extraction.

Study
Study Research Problem Contributions
RQ1: Heterogeneous Data
RQ2: Data Harmonization
RQ3: Industrial Textual Data and Techniques
RQ4: Sequential Data Techniques

2.2.3. Information Synthesis

At this stage, the extracted data were aggregated to answer the research questions. For our research questions, we used the narrative synthesis method. Accordingly, we used tables and charts to present our results.

2.3. Report Review

Data extracted from the primary studies was used to answer our four research questions. The guidelines of References [29,30] were closely followed in the reporting of results.

3. Results

The summary of selected studies in the detailed arrangement is presented in Table A1 (Appendix A). A total of 70 studies were included in this review. Of those, 14 studies highlighted the RQ1, 25 studies covered data harmonization, 23 studies focused on the techniques, and 8 studies showed the performance measure of sequential text, as shown in Table 5.

Table 5. RQ studies.

RQ	Studies
Heterogeneity	14
Data Harmonization	25
Industrial Textual Data	23
Sequential Data Performance	08

Figure 3 illustrates the number of studies per year. There were a smaller number of studies in 2015, and then the related research studies grew in 2016/2017, such as 15 and 18, respectively. In 2018 and 2019, the number of studies found was 16 and 15, and in 2020, four studies found.

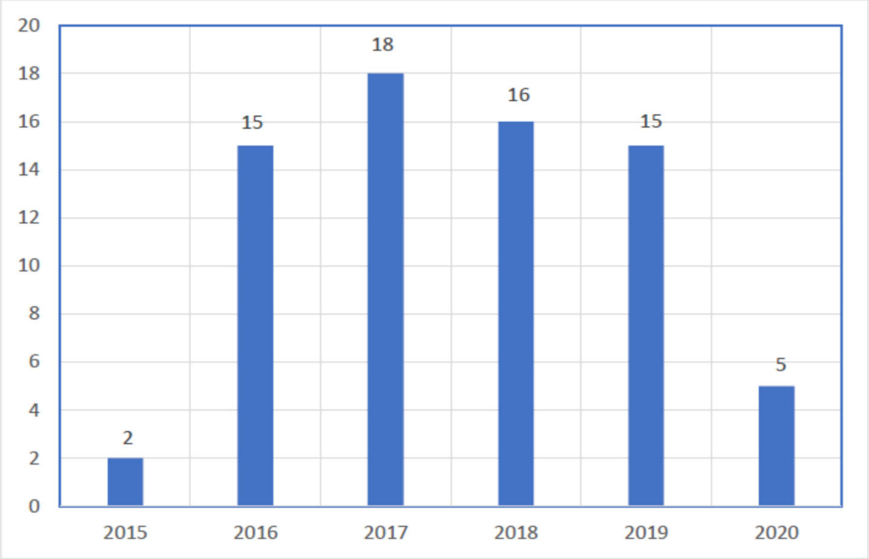

Figure 3. Studies selected per year.

Figure 4 illustrates the worldwide map of countries who contributed in research related to the research questions mentioned above. The research discussion on each research question is described in Table A1 (Appendix A).

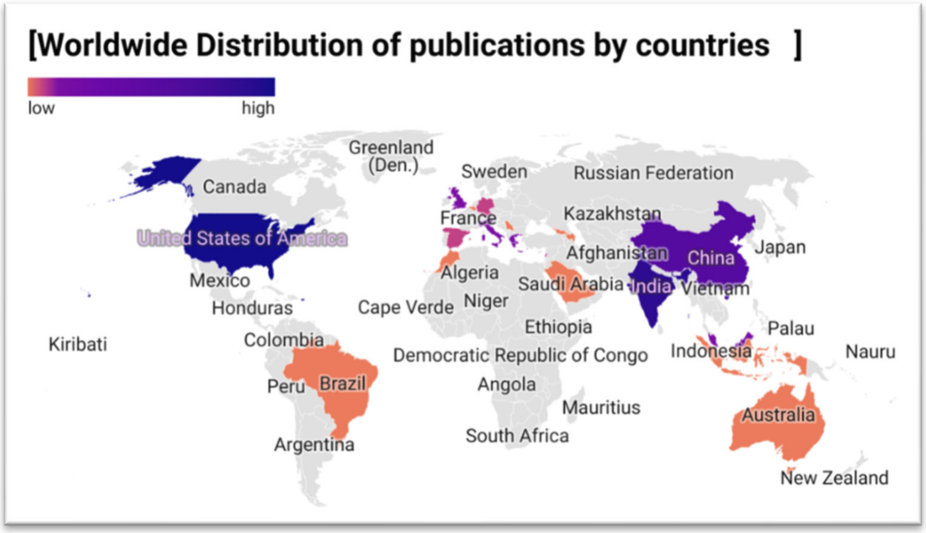

Figure 4. Chlorophet map showing the worldwide distribution of publications by countries.

3.1. *Which of the Domains Are Mainly Focused on Researching Heterogeneity?*

In this review, 14 studies discussed heterogeneity or heterogeneous datasets. Here we present a list of contributions for heterogeneous data.

Initially, the heterogeneity issues were presented by the researchers Silverio, Cavallo, De Rosa, and Galasso [31], in which the centralization of cardiovascular disease (CVD) is needed because no such system has been developed for all CVDs worldwide that will help patients and staff with responsible treatment. A framework was developed of all collected CVDs and compared based on performance ratio. In addition, a survey was conducted by a research team of Verma, Agrawal, Patel, and Patel [32] on the challenges faced by heterogeneous data in structured, semi-structured, and unstructured (SSU) formats. Various types of data are covered in SSU data, such as text, audio, images, video, and social media. In this, it is stated that Big Data Analytics (BDA) play an important role in Industrial Revolution 4.0, particularly for Big Data Analysts for making decisions, analysis, visualization, and prediction of challenges. The studies show the following challenges faced by industries: as predictive analysis, social media analytics, content-based analytics, text analytics, audio analytics, and video analytics. Moreover, Ali, Neagu, and Trundle [33] have highlighted in their work that heterogeneity has become an issue for large textual datasets used for data mining. For this, different machine learning (ML) techniques, such as KNN, SVM, NB, and ANN, are used for the classification of data, as well as for similarity, semantic-based information retrieval uses different algorithms. The result shows the accuracy of similar pairs to be better than existing classification techniques.

Likewise, the researchers Sivarajah, Kamal, Irani, and Weerakkody [34] emphasized that Big Data are growing rapidly in every type of their characteristics. Because of this issue, sometimes it is difficult to decide from large volumetric data. Multiple forms of interview forms, case studies, and experiment data files were used to check the assessment score and experts suggested that for timely decisions, data harmonization is needed, and it will remove heterogeneity from large data files. Likewise, in Bangalore city, a model of city bus heterogeneous data was proposed by Jaybal, Ramanathan, and Rajagopalan [35] to predict the bus timings, previous stops, and next stops on the map. The data formats were GPS, Web Data, CSV, and text formats. On the contrary, data heterogeneity of different Electronic Health Records (EHR), using deep learning (DL) techniques, was discussed by authors Shickel, Tighe, Bihorac, and Rashidi [36] to improve digitization of EHR. Chinese hospital data are used to create deep EHR projects with the help of DL algorithms. Moreover, in Big Data (BD), massive data are produced in unknown and untuneful patterns, thus creating a data heterogeneity problem, as highlighted by Gheisari, Wang, and Bhuiyan [37]. In this study, different DL, ML, and BD techniques were highlighted to remove heterogeneity and make data useful.

Furthermore, data mining of heterogeneous data plays an important role in getting proper knowledge and information from huge datasets, as presented by Kalra and Lal [38] in their work. By using ML classification and clustering techniques, it is possible to fetch data. Moreover, the authors Kolhatkar, Patil, Kolhatkar, and Paranjape [39] expressed their views on online education systems, such as MOOC and Moodle. There are issues related to structured and unstructured data formats. It is difficult to manage and store heterogeneous data. To remove the heterogeneity issue, a conceptual model was proposed using software applications and RDBMS. In addition, the prominent authors Sindhu and Hegde [5] contributed with their work by proposing that the handling of large and complex data creates the problem of heterogeneity and is solved by using different BD techniques and text-mining algorithms. The conversion of unstructured into structured format performance is calculated in time. Moreover, real-time healthcare data are generated from sensors and gadgets in unstructured formats, which is not possible in structured formats with simple tools. Ismail, Shehab, and El-Henawy [40] stated in their work that healthcare data need to be monitored, visualized, and predicted time by time for these BD techniques, such as Hadoop and MapReduce, used for developing the EHR. Moreover, for the development of heterogeneity, Zhang et al. [41] highlighted the use of large-scale urban multisource heterogeneous data integrated by using tensor decomposition to solve the issues of urban town and to make it a smart town.

Additionally, Elsharkawy, Ahmed, and Salem [42] believe that the heterogeneity and complexity are two main issues highlighted for the solution and integration of clinical data. The data are generated in different formats, such as RDF, XML, and images. The issues were solved by using ontology and semantic techniques. Recently, the authors Arora and Goyal [3] emphasized that the various frameworks of heterogeneity and heterogeneous datasets can be used for solving the issues of heterogeneity, and it was shown that heterogeneity is always due to the unusual format of data and lack of integration resources, expertise, and techniques. Table 6 shows the advantages and disadvantages of studies selected for Research Question 1.

Table 6. RQ1 domains' advantages and disadvantages.

Study Reference	Domain	Advantages	Disadvantages
[31]	Healthcare	Millions of E HR combined	Limited data access to selected smallest patient populations
[32]	General Purpose	diverse, massive, and complex data	Requires new norm of integration
[33]	Text similarity	Helps in classifying similar objects. Increase performance of ML techniques	It is considered inconsistent and ad-hoc.
[34]	General Purpose	The natural property of BD. Combine and manage	variety of inconsistent data create problems
[35]	Information Retrieval	Utilize data from a variety of sources. Single unified data source	Loss of information due to semantic, syntactic, and schematic difference
[36]	Healthcare	Electronic health record became unique	Difficult to manage different form
[37]	General Purpose	DL helps in solving the integration of heterogeneous data	Conflicting information shared by resources
[38]	Information Retrieval	Efficient data extraction Effectiveness of data	Meaningful data extraction from the huge database
[39]	Education	Helps in distribution across clusters	Need More memory for processing

[5]	Healthcare	Need to manage data efficiently and collaborative way	Need to be distributed and parallel computing systems and database
[40]	Healthcare	Multiple different sources data with a unified view	Data processing Speed and quality of data analytics
[41]	Infrastructure	Extremely large multisource infrastructure containing vehicles, residents, and smart card	Some models are only indirectly relevant to a particular phenomenon of interest
[42]	Healthcare	Semantic-based integration and semantic-based medical retrieval	Clinical records without linguistic standard
[3]	General Purpose	Heterogeneous data can be solved by using RDBMS,concept lattice, and MapReduce	Performance of all heterogeneous data is not calculated

3.2. How Does the Data Harmonization Resolve the Issues of Heterogeneity?

In this section, 25 studies were selected which discuss data harmonization, data integration, and data fusion. The details of each study are discussed below.

Initially, heterogeneous oil and gas data are unorganized, which is difficult to manage. For that data harmonization was proposed by Danyaro and Liew [43], using semantic web and BD tools. Where performance of the precision, recall, and F-score found better than existing techniques. In addition, agriculture data are stored in clusters, and it is difficult to handle heterogeneous data. Therefore, a uniform format was reported by (Sambrekar, Rajpurohit, and Joshi [44], using Couchbase and NoSQL, and it was found that the time duration for fetching records is fast. Apart from this, different frameworks have been developed by different organizations to make decisions, but no framework has been proposed for value creation. In this study, Saggi and Jain [45] created a framework for value creation from SSU data also in-depth issues of heterogeneity, harmonization, and BD techniques were highlighted. It shows the importance of data integration for industrial data, decisions, product reviews, and visualization of future strategies. Artificial intelligence, ML, and cloud computing will be helpful for BD Analysts. Moreover, Li, Chai, and Chen [46] summarized that the heterogeneous data in industry are produced easily but are difficult to store, manage, and audit. In this study, the issue of heterogeneity of large firms was solved using a NoSQL-based data integration model. Furthermore, health data are very important for patient treatment, monitoring, and satisfaction. Health data are generated by all institutes by using open-source web data, but no such online module has been proposed for integration of all web-based centralized. In their study, Hong, Wang, et al. [47] revealed a Web-based FHIR visualization tool, using a standard structured format API. Again, Lopes, Bastião, and Oliveira [48] described that the file sharing between users was difficult for heterogeneous data. Therefore, a real-time integration and interoperability model was developed by using PostgreSQL to facilitate different users.

In addition, Yuan, Holtz, Smith, and Luo [49] mentioned that the child-patient disorder/condition data were complex and unmanageable due to manual work and human involvement. To overcome this issue, different preprocessing, NLP, and ML tools are used to create patient data in digital form and without any biases. The performance of the autism spectrum is calculated using precision and recall. Furthermore, Daniel [50] also emphasized on the issues and challenges faced by educational institutes and researchers are highlighted, such as data integration and sharing between campuses and branches. Besides this, text-free or unstructured data in healthcare data create issues for managing and storing. Therefore, data fusion was suggested by the Kraus et al. [10] to manage the heterogeneous data. Moreover, in an online learning system, data need to be integrated and efficient for smart educational systems. Data processing and storage of audio, video, images, and text formats was developed by Dahdouh, Dakkak, Oughdir, and Messaoudi [51] with the help of Hadoop, MapReduce, and Spark. As a result, it helps in taking a smart decision within seconds. Additionally, Patel and Sharma [52] explained the various issues of data harmonization in this survey. Before that, data warehousing and OLAP were used, which do not support huge datasets of open source and unstructured formats. In the end, different BD and ML techniques are suggested for dealing with huge data. Consequently, in the oil and gas industry, data are generated in operational formats from different clusters at a time, which needs data integration to collect data in a centralized place for making timely decisions identified by Alguliyev, Aliguliyev, and Hajirahimova [53].

Wang [54] mentioned that the disparate data are generated in unstructured formats, such as sensors and text, which describe heterogeneous behavior. For this reason, a data integration model was developed to solve the technical and quality problems of BDA. The model was developed using ML and DL techniques so that BD analysts could visualize, analyze, and make decisions from disparate data. Additionally, Chondrogiannis et al. generated a tool for clinical data in a heterogeneous form and for integration of data, an ontology-based tool suggested to arrange data in a structured format. Moreover, patient cohort and biomedical data play an important role for previous health treatment and analysis, and data provided by patients in a heterogeneous structure need to be harmonized, as argued by Kourou et al. [11], so that, in an online tool, all patient data are available to medical staff during analysis. In this survey, different cohort harmonization techniques were highlighted, which will help in healthcare applications, such as ML, DL, and Ontology techniques. In addition, in an urban town, so many issues related to basic needs was mentioned by Souza et al. [55]; the objective of that study was to make the urban town into a smart urban town. Data are generated by different departments in JSON, string, and maps. To make smart decisions, all data must be integrated.

Furthermore, the patient stays in hospital data with different codes were not publicly available to make an health records into an EHR reported by authors (Scheurwegs, Luyckx, Luyten, Daelemans, and Van den Bulcke [56]. By using Naïve base and Random Forest on the UZA dataset, the patient classification was performed. Similarly, the researchers Jayaratne et al. [57], in their study, stated that the web-portal-based patient data produced by many healthcare hospitals in different formats were difficult to decide due to decentralization. To solve this issue, an automated and centralized web portal was developed which helps with online decisions. In contrast, the research team of Hong, Wen, Stone, et al. [58] analyzed that the patients with obesity and comorbidities were monitored after discharge from hospitals. The objective of this study is to develop a patient-centric system for FHIR using NLP toolkits and ML algorithms from the Mayo Clinic, MIMIC III, and i2b2 datasets. The overall performance of this system is measured in precision, recall, and F-Score. In addition, the same authors, Hong et al. [59] proposed a model for the quality and performance-based data integration for information extraction, using NLP, ML, and Bag of Words (BoW). Moreover, Hong et al. [60] used a Mayo Clinic dataset with the help of NLP toolkits for making a digital FHIR system. In contrast, Chen, Zhong, Yuan, and Hu [61] conducted a review and suggested a unified model for SSU data, using MapReduce. Besides that, XML-based OGOLOD datasets were accessed by using ontology tools for a semantic oriented data harmonization model that was presented by Carmen Legaz-García, Miñarro-Giménez, Menárguez-Tortosa, and Fernández-Breis [62].

In Saudi Arabia, patient health data generated in public and private hospitals are not shared and integrated with the health information system due to a lack of heterogeneity. Therefore, the Banu, Kuppuswamy, and Sasikala [63] team proposed a NLP and BDA-based systems. Lastly, online FHIR-based web portals were developed by using NLP techniques and open-source tools on the Mayo Clinic dataset to centralize the data generated in a heterogeneous format that was revealed by the researchers Hong, Wen, Shen, et al. [64]. The contributions of all studies in all domains are discussed in Table 7.

Table 7. RQ2 domain and contributions.

Study Reference	Domain	Contributions
[43]	Oil and Gas	High performance measure
[44]	Agriculture	High performance, high availability, and high scalability, using the latest techniques
[45]	General-Purpose	Data generation, storing, fetching, analysis, visualization, and decision-making
[46]	Banking	Helps in auditing the multisource data
[47]	Healthcare	Facilitate for navigation of HL7 FHIR core resources
[48]	General-Purpose	Delivering automatic services to interoperable system
[49]	Healthcare	Helps in developing an automatic system for disordered patient
[50]	Education	To motivate researchers and academicians about the latest techniques
[10]	Healthcare	Useful for decisions of scientific, clinical, and administrative work
[51]	Education	Facilitate in online learning, storage, processing, and academic activities
[52]	General-Purpose	Recommendation system, opinion mining, and parallelism can be targeted
[53]	Oil and Gas	Helpful for decision-makers during exploration, drilling, and production
[54]	General-Purpose	It will facilitate for fetching data and performance measure
[65]	Healthcare	Helpful for disease prevention, tracking, and policy making
[11]	Healthcare	Helps in boosting statistical power of sustainable and robust data
[55]	Infrastructure	Geographic based smart city for aggregation, visualization, and analysis
[56]	Healthcare	Helps in predicting the clinical codes of patient stays
[57]	Healthcare	Helps in patient-centered care decision-making among stakeholders
[58]	Healthcare	Helps in finding the patient having obesity and comorbidities
[59]	Healthcare	Helps in developing patient diagnostic criteria and representation
[61]	General-Purpose	Support in integration, storage, computation, and visualization
[62]	Healthcare	Open biomedical repositories can be developed in semantic web formats
[60]	Healthcare	Normalizing and integration of structured and unstructured EHR data
[63]	Healthcare	Helps health information system to keep a record of patients' data
[64]	Healthcare	Helps in standardizing the clinical data normalization

3.3. Which Techniques Are Being Used for Solving the Harmonization Issue for Large Textual Datasets?

In previous studies, SSU heterogeneous data were used in the form of text, images, audio, video, and social media formats. The BD and BDA literature reviews proposed so many models and frameworks for data harmonization or integration. Among them, textual data play an important role in semantic, syntactic, and schematic data from large datasets. In different industries, different approaches are used by BD analysts to meet the demands of users and owners.

In this section, 16 studies were selected that highlight the core techniques and their contributions in terms of performance, time, and accuracy in data harmonization, data integration, and data fusion. The details of each study are discussed below.

At first, Tekli [66] found that, in the entertainment industry, the feedback given by the audience in form of large sentences and getting semantic meaning from XML documents is very challenging. Additionally, Sanyal, Bhadra, and Das [67] pointed out that, by using business intelligence tool sentence-similarity retrieved, the technique proposed for the IT Ecosystem has been adopted by business firms. Apart from that, in the health sector, data are also important for harmonization, as noted by Adduru et al. [68]. They also discussed how the dataset contains many clinical codes and how it is difficult to get information and text classification to solve the issue. NLP techniques, such as N-Gram, Jaccard Similarity, Word2Vec, and different DL approaches, are used to create a paraphrasing dataset from clinical data. Similarly, the research team of Mujtaba et al. [69] revealed, in a clinical-text-classification review that the approaches for textual data play an important role, especially in supervised ML techniques. Likewise, a medical prescription is a document of proof about a patient's health history recorded during the diagnosis, but sometimes it is difficult to understand the semantics of prescribed medicines was presented by Yanshan Wang et al. [70]. In this study, the Mayo Clinic dataset was utilized with the help of NLP techniques to find the semantic and similarity scores of medical texts. On the contrary, a study was proposed by Chen, Hao, Hwang, Wang, and Wang [71] that states that the healthcare communities manage healthcare data on web-based portals but are not available to all medical practitioners. For the prediction of chronic diseases, ML classification algorithms, such as CNN, NB, KNN, and DT, are used for analysis. Besides that, the authors Pathak and Lal [72] focused on open-source files-based heterogeneous datasets developed by using Modified IDF cosine similarity for information retrieval. A very detailed and descriptive survey was carried out by the authors Torfi, Shirvani, Keneshloo, Tavvaf, and Fox [73]. In this survey, different datasets of open-source NLP tasks, using different DL methods on BERT models, were discussed to summarize text and word embedding. In addition, Wu, Zhao, and Li [74] proposed that phrases of NLP models be vectorized by using the phrase2Vec model to overcome the issues of BoW and preprocessing. In the same way, the authors Moscatelli et al. [75] stated that patient data are very critical and sharing them is possible with high-security algorithms. By using NoSQL, MongoDB, and NLP techniques on XLS, CSV, and TXT files, data acquisition and simulation are possible. Similarly, Chen, Du, Kim, Wilbur, and Lu [76] also emphasized that, with the use of advanced technology, the health sector can be upgraded. Furthermore, health records can be in the digital form of clinical data and support multiple formats, but it is not easy to fetch similar data for digital records without the latest techniques in text mining. DL-based entities fetched from STS datasets combine rich features. Despite this, Mahlawi and Sasi [77] found that, from the large number of Enron email datasets, data are extracted by using NLP and sentiment analysis to make them available in a structured format. Furthermore, the authors Eke, Norman, Shuib, and Nweke [78] noted that the other parts of NLP are also important. In that, lexical analysis and ML-based emotional behavior detected from the text messages were used to check the level of criticism or hurt level from the Sarcasm dataset. Moreover, biomedical text mining was performed by using text preprocessing, clustering, classification, and information-extraction techniques mentioned by Allahyari et al. [79]. This led authors García, Ramírez-Gallego, Luengo, Benítez, and Herrera [80] to focus on Indian regional multilingual data processed with the help of natural language-processing techniques. Finally, Harish and Rangan [81] suggested that text be processed through ML and DL algorithms for semantics. BD processing for huge data is performed by using BD tools and libraries.

The contributions, techniques, and domains of all studies are discussed in Table 8.

Table 8. Domain, techniques, and contributions.

Study Reference	Domain	Techniques	Contributions
[66]	Entertainment	Sentence Similarity	Using computational identification of the meaning of data in context from XML large datasets
[67]	Business	Business Intelligence	To extract values from the organized and refined data also helps stakeholders to make decisions
[68]	General	Sentence Similarity	Helps in training deep learning models for clinic paraphrase generation and simplification
[69]	General	Feature Extraction	To calculate the performance of expert-driven and fully automated features on free-text clinical reports
[70]	Healthcare	Sentence Similarity	Support reduction in cognitive burden and improvement in the clinical decision-making process
[71]	Healthcare	Word Embedding	Analysis of illness will help early disease detection, patient care, and community services

[72]	General	Information Retrieval	Document retrieval from large datasets of disparate formats
[73]	General	Deep Learning	Automatic semantic analysis and data drove strategies for Computer Vision, ASR, and NLP
[74]	General	Phrase Embedding	Helps in the integrity of semantic units and vectorization of similar words
[75]	Healthcare	Preprocessing	To precise historical analysis of clinical activities of patient
[76]	Healthcare	Sentence Similarity	To get clinic semantic textual similarities such as lexical patterns, word semantic, and named entities
[77]	Email	Preprocessing	Unstructured email extracted in a structured format
[78]	Social media	NLP	To find the sentiment from the message concerning content and level of hurting or criticizing
[79]	General, Health	Knowledge discovery Database	All tasks and techniques related to textual data, such as IR, NLP, IE, text summarization, unsupervised and supervised ML, opinion mining, and biomedical text mining, are discussed
[80]	General	NLP, BD tools	Data with long instances of text processed using tools and libraries
[81]	Multilingual	NLP, ML	Multilingual text preprocessing issues by using text mining and processing discussed

3.4. How NN Algorithms Are Well-Suited with Respect to Efficiency for Large Sequential Datasets

In this section, 8 studies were selected which highlight the performance of Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) used for sequential data. The details of each study are discussed below.

At first, the researchers Yin et al. [82] and Ouyang et al. [83], in both surveys, discussed the use of NLP and DL techniques for fake-news detection and sequential data. By using techniques, it is found that the accuracy of model is up to 93%. Moreover, a comparison of CNN and RNN reveals that RNN is better than CNN. The techniques can be used for sentimental, relational, textual entitlement, answer selection, QA path query, and POS tagging was pointed by Lopez and Kalita [84]. Additionally, the authors Chai and Li [85] selected the studies that work for the Chinese community. In that, Chinese language based Clinical NER's performance was increased by using NLP techniques with DL. Similarly, the other techniques such as RNN with DL always shows better results which was presented by authors Oshikawa, Qian, and Wang [86]. In addition, with the help of NLP in the different domains, the sequential data performance is optimum also highlighted by Young, Hazarika, Poria, and Cambria [87,88]. lastly, a survey was conducted by the authors Jing and Xu Jing and Xu [89,90] which depicts the performance of RNN with the addition of NLP is at it shows the performance at its peak.

The contributions, techniques, and domains of all studies are discussed in Table 9.

Table 9. Model Performance Techniques.

Study Reference	**Domain**	**Techniques**	**Contributions**
[82]	General	CNN, RNN for NLP	RNN perform better
[83]	Healthcare	RNN, N-Gram	RNN performance better by using N-gram
[84]	General	Compared with the existing Algorithm of CNN	RNN outperformed

[85]	General	Used in many NLP and audio-video functionality	Better for sequential text
[86]	Fake News	RNN for larger data sets of fake news	93% accuracy
[87]	General	CNN, RNN	RNN is better as per recent studies
[88]	Cancer, healthcare	DL classifier is better than conventional classifier	Model accuracy is better by using RNN
[89]	General	FFNNLM, RNNLM	RNN Language model is best
[90]	Medical, General	CNN, DBN, RNN	RNN is better in terms of NLP

3.5. Common Techniques

In this paper, different heterogeneous data and data harmonization approaches are discussed. The core techniques used in studies are presented in Table 10. The NLP technique is used all together in some studies, and also in some studies as a separate technique. For larger datasets, different techniques are highlighted that were used to help different domains. The major techniques include N-Gram, Bag of words, Bag of Phrases, TFIDF, cosine similarity, Jaccard similarity, Jaro Winkler, word2Vec, Phrase2Vec, Doc2Vec. It also helps with Name Entity Recognition (NER), text summarization, predictive analysis, word embedding, and semantic-based feature extraction from large heterogeneous datasets.

Along with NLP techniques, different machine learning algorithms for classification and clustering are highlighted in Table 10. Moreover, for training and testing of sequential data, deep learning algorithms are used for better performance and efficiency. Recurrent Neural networks performed better than CNN with the help of NLP and BD techniques. Other than that, BD, database, and web-based techniques are also discussed in different studies. For making structured formats, NoSQL, RDBMS, SQL, PostgreSQL, and ETL were used, and from the studies, it was found that the performance of structured data using these techniques is better. For web-based data, XML and ontology tools are used to fetch the data and place them in a structured format. Table 10 contains the core techniques of ML, DL, and NLP. Core techniques are also added that are based on studies relevant to research questions.

Table 10. Common techniques.

Category	Techniques	Studies' References
Storage Technology	NoSQL	[3,44,46,53,67]
	HDFS	[4,5,40,51,53,67,75,80,91]
	PostgreSQL	[48]
	DWH	[53]
	OLAP	[53]
	ETL	[91]
	SQL	[91]
	RDBMS	[39]
Web-Based Processing Technology	Semantic web	[43,62,66]
	Ontology	[11,42,62,65]
	Web	[57,64]
	Open-source	[47]
Platform Technology	Couch base	[44]
	MapReduce	[4,51,53,61,67,80]
	Spark	[51,67,80]
Processing Technology	ML	[37,38,54,58,59,69,78,79,81]
	DL	[36,37,54,68,79,81]
	Tensor	[41,61]

	KNN	[33,71]
	SVM	[33,49]
	NB	[33,56,71]
	RF	[56]
	DT	[71]
	K-Means	[74]
	ANN	[33]
	CNN	[71,92]
	RNN	[92]
Information Processing Technology	NLP	[5,49,58,60,63,70,74–77]
	N-Gram	[49,68]
	TFIDF	[49,72]
	Jaccard Similarity	[68]
	Word2Vec	[68]
	BoW	[59]
	Cosine Similarity	[72]
	Text Summarization	[73]
	Word embedding	[73]
	JaroWinkler	[93]
	Soft TFIDF	[93]
	Doc2Vec	[49]
	Text preprocessing	[79–81]

3.6. Data Formats

In Table 11, data formats or sources of data are presented. Large heterogeneous datasets and representations are found in different formats. Based on these formats, structured, semi-structured, and unstructured data are categorized. The structured data formats that are used in the studies are SQL, XLS, and String. On the other hand, semi-structured data formats that are used are CSV, JSON, XML, URI, RDF, GIS, and GPS. Unstructured data formats are categorized as sensor data, video, audio, images, text, industrial data, and social media files.

Table 11. Data formats.

Techniques	**Studies' References**
URI	[43]
RDF	[43,47]
Sensor	[40,44,54]
TXT	[5,35,44,75,79,93]
CSV	[35,44,48,75]
XML	[44,47,62,66]
SSU	[3,31,33,34,36,37,41,45,46,54,58,60,61,63,68–71,73,74,76,77]
GIS	[46]
JSON	[47,48,55]
SQL	[48]
OCR	[49]
INDUSTRIAL	[2,53,67]
VIDEO	[51,92]
AUDIO	[51,92]
IMAGE	[51,92]
WEB	[11,35,37,39,52,56,59,64,72,91]
TEXTUAL	[52]

XLS	[5,65,75]
GPS	[35]
STRING	[55]

3.7. Performance Measure Techniques

The performance measure is important for all types of models, tools, techniques, and algorithms used in industrial projects and data-fetching mechanisms. Decision-making, visualization, and prediction are the key roles played by BD analysts and their role in making the system efficient and effective. In this paper, different types of performance are measured in different studies, which are shown in Table 12. Among all, precision, recall, and F-score are dominant in all types of BD-, ML-, DL-, and NLP-based models and frameworks. Accuracy is used by many researchers, where business intelligence tools are used only by business firms which was highlighted by [31].

Ratio comparison and similarity score are calculated in studies where there is a comparison of pairs or similar sentences. Correlation of terms used for linking different health records of similar codes can be found in Reference [76]. Time is used for conversion or fetching records from larger datasets [5,51]. Based on the performance measurement techniques, it is found that it will help BD analysts to use a data harmonization model, relevant tools, effective techniques, and efficient algorithms that are used for disparate nature of data domains and real-time applications.

Table 12. Performance measure techniques.

Study Reference	Precision	Recall	F-1	Accuracy	BI	Score	Correlation	Time	Ratio
[43]	✓	✓	✓	×	×	×	×	×	×
[66]	✓	✓	✓	×	×	×	×	×	×
[49]	✓	✓	×	×	×	×	×	×	×
[51]	×	×	×	×	×	×	×	×	✓
[67]	×	×	×	×	✓	×	×	×	×
[31]	×	×	×	×	×	✓	×	×	×
[33]	×	×	×	✓	×	×	×	×	×
[68]	×	×	×	✓	×	×	×	×	×
[70]	×	×	×	×	×	×	✓	×	×
[34]	×	×	×	×	×	×	✓	×	×
[35]	×	×	×	×	×	×	✓	×	×
[55]	×	×	×	×	×	×	✓	×	×
[56]	×	×	✓	×	×	×	×	×	×
[71]	×	✓	✓	✓	×	×	×	×	×
[58]	✓	✓	✓	×	×	×	×	×	×
[59]	×	✓	✓	✓	×	×	×	×	×
[5]	×	×	×	×	×	×	×	×	✓
[41]	×	×	×	✓	×	×	×	×	×
[60]	×	×	✓	×	×	×	×	×	×
[42]	×	✓	✓	✓	×	×	×	×	×
[72]	×	×	×	×	×	✓	×	×	×
[73]	×	×	✓	✓	×	×	×	×	×
[74]	×	×	✓	×	×	×	×	×	×
[75]	✓	✓	✓	✓	×	×	×	×	×
[76]	×	×	×	×	×	×	×	✓	×
[64]	×	×	✓	×	×	×	×	×	×
[78]	✓	✓	✓	✓	×	×	×	×	×
[93]	×	×	×	×	×	×	✓	×	×

✓ Technique mentioned in the study. × Technique is not mentioned in the study.

4. Discussion

The studies selected in this SLR are more from the medical and healthcare domains, discussing the heterogeneity issue and unstructured data. The main reason behind the heterogeneity issue is the heterogeneous data produced by healthcare devices, gadgets, and diagnostics reports. Due to the disparate nature of data, the need for integration or harmonization arises with the uniform representation of scattered data. In addition, textual data related to core techniques and their performance measurement were essential to be discussed in detail. The insight of this study mainly focuses on the type of SSU data; the industry focusing on the heterogeneity issue; the data harmonization or integration approach used by domains; and information-related core techniques, algorithms, tools, and models, and their performance assessment. Moreover, other aspects related to data harmonization and textual information retrieval were discussed to help the research community with up-to-date research related to textual data, such as conversion, integration, curation, and mapping of data, as well as text similarity, word embedding, sentence similarity, phrase embedding, and preprocessing techniques. All research questions, such as how harmonization solves the issues of heterogeneity, techniques, and performance, are discussed in Section B. Moreover, the recently implemented standard methods by researchers and data analysts for different purposes are shown in Table 10. The disparate nature of data produced in multiple formats, shown in Table 11, shows how commonly SSU textual data files are processed for desired output. The performance measurement technique discussed in Table 12 shows how techniques such as precision, recall, F-1, accuracy, and time will help researchers and managers to evaluate the performance of the core techniques, tools, algorithms, and models.

As we figured out by searching using different search databases, no such data harmonization SLR is proposed and presented that mainly focuses on textual data (SSU). Related work with data harmonization was performed by the research team at Stanford University in 2011. They proposed a multimodal deep learning technique for multimodalities to enhance the performance of audio/video networks in Reference [94]. With the advent of technology and tools, multimodalities were used by various researchers for big multimodal data by using text, audio, visual, and physiological signals [27]. The replacement of the data harmonization keyword was proposed, such as data integration and fusion [95]. A survey paper has been published that deals with affective computing [26] and related areas, such as opinion mining, emotion, and sentiment analysis. In that unorganized and unstructured data generated from consumer feedback is used to check the modalities' feasibility, physiological data helps them with anti-spoofing. Textual data are widely used for information retrieval, sentiment analysis, and text similarity, and a review was conducted on multimodal fusion for affective computing that mainly focuses on sentiment analysis in 2017. Text, audio, and video were data inputs, along with feature and decision format fusion techniques. Sentiment analysis was performed based on supervised and unsupervised approaches, and NLP, ML, and DL techniques were used. Meanwhile, textual modality, heterogeneity, and performance measurement of methods were unnoticed.

Later, different studies related to the proposed SLR were discussed. They are described below. In Reference [25], a multimodal machine learning survey characterizes the fusion modalities for multimodalities and issues such as representation, translation, alignment, fusion, and co-learning. Still, no parameter related to heterogeneity was discussed. In Kourou et al. (2018), different harmonization methods for healthcare case studies mainly concentrate on the percentage of data harmonized. Moreover, in Reference [96], sentiment was predicted with the help of the multimodal approach. Punctuation predicted from conversational speech, using semi-supervised multimodal fusion techniques, is presented in Reference [97]. The hierarchical fusion technique was used for sentiment analysis using TAF data and social images [98–100]. Moreover, a fake-news-detection framework was proposed in References [100,101] for Twitter and image multimodalities. In Reference [102], movie content similarity was detected by using text, auditory, and visual information. As in Reference [103], emotions with the recommendation of traveling place framework proposed using DL techniques on text, pictures, and video data. The existing related work mainly focuses on analyzing single modalities other than textual data. Limited to core techniques, neither integration nor performance measurement is contrary to the proposed SLR.

Based on the literature review and results discussed above, many studies are presented, which focus on research related to research questions. The insight from these results will help in futuristic applications and domains such as zero-shot network, transfer learning, IIoT, Industry 4.0, and extended reality applications. These technologies produce data in bulk that are disparate, so the data need to be processed before storage and presentation. The techniques which fall under the umbrella of machine learning, deep learning, and NLP will be beneficial for the performance and extraction of information retrieval. Performance evaluation techniques such as precision, recall, F-1, accuracy, and time will help data analysts and managers in the selection of relevant tools and techniques.

4.1. Contributions

To the best of our knowledge, this is the state-of-the-art SLR that discusses heterogeneous textual datasets. The main objective of this study was to perform an in-depth review of the heterogeneity issue; the data harmonization or integration approach for massive and scattered information; computational techniques that can process and manage textual information in an efficient manner; and performance measurement of tools, techniques, and models. Previously, various related studies, such as multimodalities [25–27], discuss the disparate nature of data, but no study explains the issues stated in the research questions.

Based on the results and discussion, the knowledge delivered from this SLR will be helpful for researchers and data analysts who are dealing with the enormous types of industrial text files and datasets. As mentioned, the most recent machine learning, deep learning, and NLP techniques would be helpful in fetching, representation, and visualization of data that are in the form of multiple formats. In addition, performance measurement techniques will also assist in selecting optimal techniques, tools, models, and frameworks. Besides this, data harmonization will help with futuristic applications, such as IIoT, Industry Revolution 4.0, extended reality, and zero-shot network domain data.

4.2. Implications for Practice

Massive data produced by domains such as healthcare, banking, insurance, law, infrastructure, education, oil and gas, telecommunication, and entertainment were managed by managers and data analysts with existing tools and techniques. Textual datasets contain raw data and information; retrieving helpful information from applications such as user feedback, semantic similarity, textual similarity, word embedding, and emotion recognition is a challenging task for decision-makers. With the advent of the latest tools and technology, it is the responsibility of managers and data analysts to process, store, fetch, and efficiently represent data so that the decision-maker can perceive all the data and make decisions appropriately. To solve heterogeneity, heterogeneous data must be harmonized and presented so that effective decisions can be made. For uniform representations, managers and data analysts need to select the appropriate SSU data formats, such as Microsoft Excel Spreadsheet file (XLS), text file (TXT), JavaScript Object Notation (JSON), etc. Moreover, efficient, and effective machine learning, deep learning, and NLP techniques for textual data will help with faster training and testing approaches. As a result, the performance measure can be obtained through precision, recall, F-1, accuracy, and time.

Academically, data harmonization or data integration is a hot topic under the umbrella of Big Data variety, because it deals with the disparate nature of data and produces information in bulk. The information retrieval from SSU data formats needs time, and for fetching the records, it needs to be uniformly represented. For this purpose, data harmonization can be an excellent approach to characterize, predict, and visualize. As for the consequences of academic performance, data harmonization is a domain where research can move forward.

5. Conclusions

Big Data describe an occurrence in the complex and dynamic growth of data, and it is challenging to manage the variety of data with simple tools and techniques. The main objective of this study was to review the heterogeneity issues, data harmonization approach, core techniques for textual data processing, and their performance measurement. Core techniques, such as NLP, ML, and DL, were applied to real-world applications and reviewed in detail, so that they can cover heterogeneity, harmonization, and sequential large textual datasets. It is assumed that the heterogeneity issue is solved by using data harmonization or data integration approaches for real-world applications. In addition, there is a requirement to adopt high-performance techniques, optimized algorithms, and efficient measurement techniques.

The main contribution of this study relates to Big Data variety and data analytics, in that it solves the most crucial issue of data heterogeneity: managers and data analysts. Before presentation or visualization, data must be uniformly managed and stored with the DH approach and the latest analytical techniques. Compared with existing related work, the multimodalities [25–27], specifically, focus on multiple data formats, such as text, image, audio, video, and visual representations. Along with this, information retrieval and classification techniques were used, but no such point was discussed as proposed in this study's research questions.

Limitations and Future Research Directions

The most important limitation of this SLR is data availability and context of the domain, as harmonization keyword uses in many applications. Another limitation is the biases in the selection of articles, SLRs, and surveys. Some relevant articles were not available, as the paper acceptance and publication phase was between 2015 and 2020. A minor limitation is the selection of studies in the English language.

Big Data and data analytics play an essential role in futuristic technologies, such as IIoT, extended reality, Industrial Revolution 4.0, transfer learning, and blockchain. Data harmonization can uniformly represent massive data produced by domains and application platforms for solving heterogeneity. Based on the detailed discussion and in-depth review of all the articles, techniques, tools, and models, the following are the essential suggestions for moving forward. In medical health, update the FHIR based EHR model by using NLP and DL techniques on SSU data. Information retrieval should be from domains such as emotion, affective computing, sentiment analysis, and content similarity from SSU datasets. Measurement of the effects of emotion recognition from textual data shows that real-time fusion methods can be used to fuse information extracted from raw data. It will also help in getting optimal solutions from clustering techniques and core NLP techniques. It will play an important role in the automation of the education system based on learners' choices and semantic meaning. Moreover, with Ontology and XML, DH can be used for web-based applications, models, and frameworks in healthcare, business, finance, education, oil and gas, and other industries, which will help them to retrieve information from users' semantic meaning and predictive analysis. Moreover, automatic text summarization, phrase similarity using bag of phrases, vocabulary generation, and the semantic meaning of long sentences using the DH approach will be performed for heterogeneous textual datasets. In addition, it will help users from any context, domain, and application.

Author Contributions: Conceptualization, G.K., S.B., A.A.I., and S.A.K.; methodology, G.K., S.B., A.A.I., L.F.C. and. S.A.K.; validation, G.K., S.B., A.A.I., and S.A.K.; formal analysis, G.K., S.B., and

A.O.B.; investigation; resources, G.K., S.B., A.A.I., A.O.B., and S.A.K.; data curation, G.K.; writing—original draft preparation, G.K. and S.B.; writing—review and editing G.K., S.B., A.A.I., A.O.B., L.F.C.; and S.A.K.; visualization G.K., S.B., A.A.I., A.O.B., and S.A.K.; supervision, S.B. and S.A.K.; project administration, S.B.; funding acquisition, S.B.,L.F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was partially supported with the funding of Universiti Teknologi PETRONAS, under the Yayasan Universiti Teknologi PETRONAS (YUTP) fundamental Research Grant Scheme (YUTP-FRG/015LC0297).

Institutional Review Board Statement: Not Applicable.

Institutional Review Board Statement: Not Applicable.

Data Availability Statement: Not Applicable.

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

Abbreviations

Abbreviation	Full Name
RFID	Radio Frequency Identification
AI	Artificial Intelligence
SLR	Systematic Literature Review
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
RQ	Research Question
ACM	Association of Computing Machinery
IEEE	Institute of Electrical, Electronics Engineering
ISI	Institute for Scientific Information
KNN	K- Nearest Neighbors
SVM	Support Vector Machine
NB	Naïve Base
ANN	Artificial Neural Network
GPS	Global Positioning System
CSV	Comma Separated Value
MOOC	Massive Open Online Course
RDBMS	Realtime Database Management System
RDF	Resource Description Framework
XML	Extensible Markup language
EHR	Electronics Health Record
BD	Big Data
NoSQL	Not only SQL
FHIR	Fast Healthcare Interoperability Resource
OLAP	Online Analytical Processing
JSON	JavaScript Object Notation
XLS	Microsoft Excel Spreadsheet
TXT	Text
IR	Information Retrieval

IE	Information Entity
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
FFNNLM	Feed Forward Neural Network Language Model
RNNLM	Recurrent Neural Network Language Model
DBN	Deep Belief Network
TFIDF	Term Frequency Inverse Document Frequency
ETL	Extract Transform Load
URI	Uniform Resource Identifier
GIS	Geographic Information System
BoW	Bag of Words

Appendix A

This section comprises Table A1, which illustrates the selection of studies based on inclusion and exclusion process.

Table A1. Selected studies.

Study	Research Problem	Outcome
(Y. Wang et al., 2020) [52]	The semantic textual similarity of clinical text	Clinical similar text selected based on semantic behavior by using STS and NLP
(Torfi, Shirvani, Keneshloo, Tavvaf, and Fox, 2020) [75]	DL methods used for tasks and models	Different NLP basic tasks, their application, and using DL domains highlighted to enhance the system with high performance
(Wu, Zhao, and Li, 2020) [76]	Advancement of BoW and NLP	Performance of NLP and BoW enhanced by BoP and P2V to solve the problem of phrase embedding
(Eke, Norman, Shuib, and Nweke, 2020) [82]	Classification of sarcasm using NLP	Level of hurt or criticism in message text classification and NLP techniques discussed
(Harish and Rangan, 2020) [93]	Processing of low resource languages using language processing task	Raw text processed using text preprocessing, ML, and NN
(Hong, Wang, et al., 2019) [35]	FHIR visualization design, development, and evaluation	Visualization tool developed for HL7 FHIR users to add profiling
(Daniel, 2019) [39]	Data integration and sharing educational data	Use of Big Data in education especially ontological, technical, ethics, privacy, and lack of expertise discussed
(Dhayne, Haque, Kilany, and Taher, 2019) [44]	Healthcare data integration techniques and techniques	Issues with healthcare data integration, heterogeneity, and heterogeneous data are discussed
(Silverio, Cavallo, De Rosa, and Galasso, 2019) [45]	Heterogeneity of high volume, wide variety, and speed	Public health improvement, drug surveillance, integrating pharmacy, data integrity, data security, and legal issues were discussed for CVD
(Chondrogiannis, Andronikou, Karanastasis, and Varvarigou, 2019) [49]	Clinical data uniformity	Ontology-based data harmonization of clinical data terms expressed in a common frame

(Mujtaba et al., 2019) [51]	Key aspects of clinical text	Supervised ML-based clinical text classification, feature extraction, representation selection technique discussed with future suggestions
(Jayaratne et al., 2019) [57]	Data Integration of patient to take a decision	Sports injury data integration platform developed for patient-centered healthcare and clinical decision support
(Hong, Wen, Stone, et al., 2019) [64]	Phenotyping framework for FHIR	ML algorithms and discharge summaries of patient conversion developed using NLP on the i2b2 dataset
(Ismail, Shehab, and El-Henawy, 2019) [69]	Solution for healthcare heterogeneous data	Healthcare system proposed for collecting and managing data from different healthcare devices to store and manage using BD tools and techniques
(Maheshwari, 2019) [4]	Issues associated with a variety of SSU data	Large heterogeneous data storing, processing, and transferring discussed
(Arora and Goyal, 2019) [3]	The various framework developed for SSU data	Heterogeneity and dimensionality issues discussed for data visualization
(Hong, Wen, Shen, et al., 2019) [80]	To develop pipeline using HL7 fast health	FHIR based Clinical unstructured and structured data integration model developed for EHR data normalization using NLP toolkits
(Chai and Li, 2019) [87]	Information processing from video recognition	Advanced applications used for information process discussed using RNN and CNN
(Guan et al., 2019) [90]	Cancer patient progress from free text	Information extracted and classified using NLP, DL, and ML techniques from cancer patient data
(Jing and Xu, 2019) [91]	Word sequence calculation from large data	Using NLP and ML the sequence of words and vocab identified
(Sambrekar, Rajpurohit, and Joshi, 2018) [32]	Conversion of unstructured data is not properly performed	SSU data conversion in addition it will provide better efficiency, scalability, and performance
(Saggi and Jain, 2018) [33]	Value creation from the different perspective of BDA	An integrated framework for Big Data and value creation from data
(Kraus et al., 2018) [10]	Data harmonization which is not available in healthcare	Data integration framework proposed to cover data harmonization, semantic enrichment, and data analysis process for medical data
(Dahdouh, Dakkak, Oughdir, and Messaoudi, 2018) [40]	An integrated online learning system	The architecture of BD for online learning system proposed with storage, processing, benefits for professionals, students, and teachers discussed in detail.
(Ali, Neagu, and Trundle, 2018) [48]	Classification of heterogeneous data	By applying classifier and algorithm on pairwise similarity to enhance quality and performance
(Adduru et al., 2018) [50]	Clinical text simplification using deep learning	DL based clinical text simplification and paraphrasing dataset developed
(Kourou et al., 2018) [11]	Data harmonization of biomedical data and cohort	Data harmonization and integration cohorts discussed with open challenges of biomedical data

(Jaybal, Ramanathan, and Rajagopalan, 2018) [54]	Semantic, syntactic, and schematic view of data	Bus fleet operations analysis, diagnosis, and improvement of schedules discussed, and operation cost reduction proposed
(Hong et al., 2018) [71]	FHIR based digital data system using NLP	Unstructured and structured healthcare data integrated by using NLP tool to form a mapping of similar codes of medication data
(L. Zhang, Xie, Xidao, and Zhang, 2018) [74]	Multisource fusion using DL techniques	Multisource heterogeneous data-based data fusion model proposed to solve the issue of heterogeneity
(Moscatelli et al., 2018) [77]	Patient data sharing is critical	Clinical data of the patient and their precise historical analysis framework developed by using ML and BD tools
(Q. Chen, Du, Kim, Wilbur, and Lu, 2018) [79]	To get similarity core between clinical notes	DL models discussed for clinical semantic textual data similarity.
(Prasetya, Wibawa, and Hirashima, 2018) [83]	Measurement of text similarity algorithm	Lexical and semantic similarity performance measured between pairs
(Oshikawa, Qian, and Wang, 2018) [88]	Problems with fake news generation	Performance of fake news datasets, the technique of NLP for identification of fake news discussed
(Young, Hazarika, Poria, and Cambria, 2018) [89]	DL models and methods for NLP	RNN role in NLP applications such as Information retrieval, summarization, and their performance highlighted
(S. Patel and Patel, 2018) [4]	Usage of ML algorithm and performance	Heterogeneous data types highlighted where RNN and CNN are used for information retrieval
(Danyaro and Liew, 2017) [31]	A large amount of data are unorganized	Semantic-based integration model for O&G
(J. A. Patel and Sharma) [41]	Data harmonization of various heterogeneous data	Data quality, scalability, heterogeneity, and efficiency highlighted for disparate nature of data generated in form of Big Data
(L. Wang, 2017) [45]	The technical and quality problem of BDA	Technical challenges of data value, data mining, ML, and DL methods discussed for disparate data
(Sivarajah, Kamal, Irani, and Weerakkody, 2017) [53]	BD challenges for technology	BD characteristics issues such as heterogeneity, process challenges, and different textual analytics techniques discussed
(Souza et al., 2017) [55]	Urban planning issues of smart city	Heterogeneous-data-type-based integration of data from multiple departments integrated to make smart city
(Shickel, Tighe, Bihorac, and Rashidi, 2017) [58]	Data heterogeneity of EHR using DL	DL Techniques discussed for EHR data for larger healthcare datasets
(Gheisari, Wang, and Bhuiyan, 2017) [59]	BD and DL challenges for research trends	Data analytics, semantic indexing, preprocessing, and data governance research problems highlighted for BD
(M. Chen, Hao, Hwang, Wang, and Wang, 2017) [60]	Effective prediction of chronic disease using ML algorithms	Incomplete medical data in Chinese chronic diseases detected using CNN from structured and unstructured data
(Klašnja-Milićević, Ivanović, and Budimac, 2017) [62]	Perspective trends for education learning using BD	BD tools based online educational framework proposed for research and professionals

(Kolhatkar, Patil, Kolhatkar, and Paranjape, 2017) [63]	How to store and manage Unstructured data	Student activities, sentiment analysis, and predictive analysis suggested for educational heterogeneous data scope
(Sindhu and Hegde, 2017) [5]	Handling heterogeneity among large data	Conversion of unstructured data into structured data using text mining and HDFS for clinical text
(Pathak and Lal, 2017) [73]	Information retrieval from heterogeneous data files	Vector space model developed for information retrieval from large documents using MIDF cosine similarity
(Banu, Kuppuswamy, and Sasikala, 2017) [78]	HIS for Saudi hospital data management	Data integration model proposed for hospitals and HIS for Saudi Arabia using NLP and BD
(Mahlawi and Sasi, 2017) [81]	Structured data extraction from email	Unstructured email conversion into a structured format for knowledge extraction using NLP and text mining approach.
(Yin, Kann, Yu, and Schütze, 2017) [84]	Performance of CNN and RNN against NLP	For sequential and text matching RNN perform better than CNN
(Ouyang, Li, Jin, Li, and Zhang, 2017) [85]	To find limitations and type of medical entities for CNER	Performance analysis of rich context information with the help of medical vocab and POS using DL and NLP
(Lopez and Kalita, 2017) [86]	Enhanced CNN performance	Latest trends, techniques, and application of NLP and DL highlighted with performance measure and type of datasets
(Allahyari et al., 2017) [92]	Useful information extraction from the large volume of data	Text mining approaches, text preprocessing, clustering, classification, information extraction techniques discussed in detail with justification
(Tekli, 2016) [37]	XML based semi-structured semantic analysis	Textual data presented with a focus on semi-structured XML and ongoing challenges for XML disambiguation, semantic meaning, and combination
(Yuan, Holtz, Smith, and Luo, 2016) [38]	To make a digital system by converting unstructured and semi-structured data useful	By using NLP and ML information extracted from medical forms of ASD patients and result evaluated by experts with 91% recall
(Bhadani and Jothimani, 2016) [2]	Advancement in Big Data and web.2.0	The latest tools, sources of Big Data, techniques, software applications, and technical limitations were discussed in detail
(Sanyal, Bhadra, and Das, 2016) [42]	Data processing using BD technologies	A conceptual model proposed for value creation and decision-making from huge Big Data
(Alguliyev, Aliguliyev, and Hajirahimova, 2016) [43]	Integration of transactional data and business analytics	Industrial data integration model proposed for large transactional data and visualization
(Verma, Agrawal, Patel, and Patel, 2016) [47]	SSU related issues	Predictive, social, text, audio, video analytics related issues were suggested for different industries
(Scheurwegs, Luyckx, Luyten, Daelemans, and Van den Bulcke, 2016) [56]	Structured and unstructured patient data are not accessible	The patient stays clinical code in structured and TXT format integrated to predict the type of isolation

(Kalra and Lal, 2016) [61]	Research challenges of heterogeneous data	Data-mining techniques discussed for SSU data produced in different formats to overcome the issue of data heterogeneity
(Hong et al., 2016) [65]	To convert textual data into the structured format	With the help of NLP and ML techniques, QDM developed for clinical diagnostics report will help in standard criteria development
(Z. Chen, Zhong, Yuan, and Hu, 2016) [66]	To develop a universal model to represent visual analysis	IBD model proposed which facilitates representation, management, and visualization.
(del Carmen Legaz-García, Miñarro-Giménez, Menárguez-Tortosa, and Fernández-Breis, 2016) [67]	By using semantic web-based data integration	Semantic web-based online and open dataset for biomedical data created using the integration of XML data
(Anagnostopoulos, Zeadally, and Exposito, 2016) [68]	Classification of 4v's of BD	BD integration framework discussed to highlight the challenges, tools, and techniques which will help stakeholders
(D. Zhang et al., 2016) [70]	Heterogeneity of urban cyber-physical system	Disparate nature of data in China integrated using the cyber-physical system to facilitate the urban system with 29% better performance
(Elsharkawy, Ahmed, and Salem, 2016) [72]	Semantic-based integration and information retrieval	Semantic-based health data integration and to enhance the performance of precision medicine
(García, Ramírez-Gallego, Luengo, Benítez, and Herrera, 2016) [94]	Processing of huge data analysis	Preprocessing techniques and libraries highlighted for huge data
(Li, Chai, and Chen, 2015) [34]	Heterogeneous data sources integration	A synchronized business audit data integration model developed using middleware technology
(Lopes, Bastião, and Oliveira, 2015) [36]	Automate real-time data integration	A platform provided for data and service for original data sources

References

1. Avci, C.; Tekinerdogan, B.; Athanasiadis, I.N. Software architectures for big data: A systematic literature review. *Big Data Anal.* **2020**, *5*, 1–53, doi:10.1186/s41044-020-00045-1.
2. Bhadani, A.K.; Jothimani, D. Big data: Challenges, opportunities, and realities. In *Effective Big Data Management and Opportunities for Implementation*; IGI Global: Hershey, PA, USA, 2016; pp. 1–24.
3. Arora, Y.; Goyal, D. Review of data analysis framework for variety of big data. In *Emerging Trends in Expert Applications and Security*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 55–62.
4. Maheshwari, H.; Verma, L.; Chandra, U. Overview of Big Data And Its Issues. *IJRECE*, **2019**, *7*, 256.
5. Sindhu, C.; Hegde, N.P. Handling Complex Heterogeneous Healthcare Big Data. *Int. J. Comput. Intell. Res.* **2017**, 13, 1201–1227.
6. Younan, M.; Houssein, E.H.; Elhoseny, M.; Ali, A.A. Challenges and recommended technologies for the industrial internet of things: A comprehensive review. *Measurement* **2020**, *151*, 107198, doi:10.1016/j.measurement.2019.107198.
7. Wang, Y.; Jan, M.N.; Chu, S.; Zhu, Y. Use of Big Data Tools and Industrial Internet of Things: An Overview. *Sci. Program.* **2020**, *2020*, 1–10, doi:10.1155/2020/8810634.
8. Jaidka, H.; Sharma, N.; Singh, R. Evolution of iot to iiot: Applications & challenges. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC). Available at SSRN 3603739, 18 May 2020.
9. Ralph, B.; Stockinger, M. Digitalization and digital transformation in metal forming: Key technologies, challenges and current developments of industry 4.0 applications. In Proceedings of the XXXIX. Colloquium on Metal Forming, 21–25 March 2020, Leoben, Austria
10. Kraus, J.M.; Lausser, L.; Kuhn, P.; Jobst, F.; Bock, M.; Halanke, C.; Hummel, M.; Heuschmann, P.; Kestler, H.A. Big data and precision medicine: Challenges and strategies with healthcare data. *Int. J. Data Sci. Anal.* **2018**, *6*, 241–249, doi:10.1007/s41060-018-0095-0.

11. Kourou, K.D.; Pezoulas, V.C.; Georga, E.I.; Exarchos, T.P.; Tsanakas, P.; Tsiknakis, M.; Varvarigou, T.; De Vita, S.; Tzioufas, A.; Fotiadis, D.I.I. Cohort Harmonization and Integrative Analysis from a Biomedical Engineering Perspective. *IEEE Rev. Biomed. Eng.* **2018**, *12*, 303–318, doi:10.1109/rbme.2018.2855055.
12. Stoyanova, M.; Nikoloudakis, Y.; Panagiotakis, S.; Pallis, E.; Markakis, E.K. A Survey on the Internet of Things (IoT) Forensics: Challenges, Approaches, and Open Issues. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1191–1221, doi:10.1109/comst.2019.2962586.
13. Xiong, H.; Wu, Y.; Jin, C.; Kumari, S. Efficient and Privacy-Preserving Authentication Protocol for Heterogeneous Systems in IIoT. *IEEE Internet Things J.* **2020**, *7*, 11713–11724, doi:10.1109/jiot.2020.2999510.
14. Sahu, A.K.; Sahu, A.K.; Sahu, N.K. A Review on the Research Growth of Industry 4.0: IIoT Business Architectures Benchmarking. *Int. J. Bus. Anal. IJBAN* **2020**, *7*, 77–97.
15. Khan, M.; Wu, X.; Xu, X.; Dou, W. Big data challenges and opportunities in the hype of Industry 4.0. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
16. James, Y.; Szymanezyk, O. The Challenges of Integrating Industry 4.0 in Cyber Security—A Perspective. *Int. J. Inf. Educ. Technol.* **2021**, *11*, 242–247, doi:10.18178/ijiet.2021.11.5.1518.
17. Sajid, S.; Haleem, A.; Bahl, S.; Javaid, M.; Goyal, T.; Mittal, M. Data science applications for predictive maintenance and materials science in context to Industry 4.0. *Mater. Today Proc.* **2021**, *45*, 4898–4905.
18. Jagtap, S.; Bader, F.; Garcia-Garcia, G.; Trollman, H.; Fadiji, T.; Salonitis, K. Food Logistics 4.0: Opportunities and Challenges. *Logistics* **2020**, *5*, 2, doi:10.3390/logistics5010002.
19. Sedkaoui, S.; Khelfaoui, M. Industry 4.0 and knowledge management practices. Volto Já–Senior Exchange Program: From Idea To Implementation, *ICOMTT*, **2020**, *978-972-95259-6-4*, p. 47.
20. De Vass, T.; Shee, H.; Miah, S. IoT in Supply Chain Management: Opportunities and Challenges for Businesses in Early Industry 4.0 Context. *Oper. Supply Chain Manag. Int. J.* **2021**, *14*, 148–161, doi:10.31387/oscm0450293.
21. Shao, X.-F.; Liu, W.; Li, Y.; Chaudhry, H.R.; Yue, X.-G. Multistage implementation framework for smart supply chain management under industry 4.0. *Technol. Forecast. Soc. Chang.* **2021**, *162*, 120354, doi:10.1016/j.techfore.2020.120354.
22. Andrade, T.; Bastos, D. Extended reality in iot scenarios: Concepts, applications and future trends. In Proceedings of the 2019 5th Experiment International Conference (Exp. at'19), Funchal, Portugal, 12–14 June 2019; IEEE: Piscataway, NJ, USA; pp. 107–112.
23. Chuah, S.H.-W. Why and who will adopt extended reality technology? Literature review, synthesis, and future research agenda. Literature Review, Synthesis, and Future Research Agenda (13 December 2018), 2018. Available online at SSRN: https://ssrn.com/abstract=3300469 or http://dx.doi.org/10.2139/ssrn.3300469 (access on 28 August 2021)
24. Gong, L.; Fast-Berglund, A.; Johansson, B. A Framework for Extended Reality System Development in Manufacturing. *IEEE Access* **2021**, *9*, 24796–24813, doi:10.1109/access.2021.3056752.
25. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443, doi:10.1109/tpami.2018.2798607.
26. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125, doi:10.1016/j.inffus.2017.02.003.
27. Shoumy, N.J.; Ang, L.-M.; Seng, K.P.; Rahaman, D.; Zia, T. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *J. Netw. Comput. Appl.* **2020**, *149*, 102447, doi:10.1016/j.jnca.2019.102447.
28. Keele, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Technical Report, Ver. 2.3 EBSE Technical Report; EBSE: Goyang-si, Korea, 2007.
29. Wang, D.; Miwa, T.; Morikawa, T. Big Trajectory Data Mining: A Survey of Methods, Applications, and Services. *Sensors* **2020**, *20*, 4571, doi:10.3390/s20164571.
30. Kitchenham, B.; Pfleeger, S.; Pickard, L.; Jones, P.; Hoaglin, D.; El-Emam, K.; Rosenberg, J. Preliminary Guidelines for Empirical Research in Software Engineering. *IEEE Trans. Softw. Eng.* **2002**, *28*, 721–734.
31. Silverio, A.; Cavallo, P.; De Rosa, R.; Galasso, G. Big Health Data and Cardiovascular Diseases: A Challenge for Research, an Opportunity for Clinical Care. *Front. Med.* **2019**, *6*, 36, doi:10.3389/fmed.2019.00036.
32. Verma, J.P.; Agrawal, S.; Patel, B.; Patel, A. Big data analytics: Challenges and applications for text, audio, video, and social media data. *Int. J. Soft Comput. Artif. Intell. Appl. IJSCAI* **2016**, *5*, 41–51.
33. Ali, N.; Neagu, D.; Trundle, P. Classification of Heterogeneous Data Based on Data Type Impact on Similarity. In Proceedings of the UK Workshop on Computational Intelligence, Nottingham, UK, 5–7 September 2018; Springer: Cham, Switzerland, 2018; pp. 252–263.
34. Sivarajah, U.; Kamal, M.M.; Irani, Z.; Weerakkody, V. Critical analysis of Big Data challenges and analytical methods. *J. Bus. Res.* **2017**, *70*, 263–286, doi:10.1016/j.jbusres.2016.08.001.
35. Jaybal, Y.; Ramanathan, C.; Rajagopalan, S. Hdsanalytics: A data analytics framework for heterogeneous data sources. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, Goa, India, 11–13 January 2018; pp. 11–19.
36. Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 1589–1604, doi:10.1109/jbhi.2017.2767063.

37. Gheisari, M.; Wang, G.; Alam Bhuiyan, Z. A Survey on Deep Learning in Big Data. In Proceedings of the 22017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou, China, 21–24 July 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 2, pp. 173–180.
38. Kalra, M.; Lal, N. Data mining of heterogeneous data with research challenges. In Proceedings of the 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, India, 18–19 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–6.
39. Kolhatkar, S.; Patil, M.; Kolhatkar, S.; Paranjape, M. Emergence of Unstructured Data and Scope of Big Data in Indian Education. *Emergence* 2017, *8*.
40. Ismail, A.; Shehab, A.; El-Henawy, I.M. Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations. In *Security in Smart Cities: Models, Applications, and Challenges*; Springer: Cham, Switzerland, 2019; pp. 27–45.
41. Zhang, D.; Zhao, J.; Zhang, F.; He, T.; Lee, H.; Son, S.H. Heterogeneous Model Integration for Multi-Source Urban Infrastructure Data. *ACM Trans. Cyber-Phys. Syst.* **2016**, *1*, 1–26, doi:10.1145/2967503.
42. Elsharkawy, B.; Ahmed, H.; Salem, R. Semantic-based Approach for Solving the Heterogeneity of Clinical Data. *IJCI. Int. J. Comput. Inf.* **2016**, *5*, 35–45, doi:10.21608/ijci.2016.33955.
43. Danyaro, K.U.; Liew, M.S. A Proposed Methodology for Integrating Oil and Gas Data Using Semantic Big Data Technology. In International Conference of Reliable Information and Communication Technology; Springer: Cham, Switzerland, 2017; pp. 30–38.
44. Sambrekar, K.; Rajpurohit, V.S.; Joshi, J. A Proposed Technique for Conversion of Unstructured Agro-Data to Semi-Structured or Structured Data. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–5.
45. Saggi, M.K.; Jain, S. A survey towards an integration of big data analytics to big insights for value-creation. *Inf. Process. Manag.* **2018**, *54*, 758–790, doi:10.1016/j.ipm.2018.01.010.
46. Li, C.; Chai, W.; Chen, L. An Integration Model of Multi-Source Heterogeneous Audit Data. In Proceedings of the 2015 International Conference on Electronic Science and Automation Control, Zhengzhou, China, 15–16 August 2015; Atlantis Press: Amsterdam, The Netherlands, 2015.
47. Hong, N.; Wang, K.; Wu, S.; Shen, F.; Yao, L.; Jiang, G. An Interactive Visualization Tool for HL7 FHIR Specification Browsing and Profiling. *J. Healthc. Informa. Res.* **2019**, *3*, 329–344, doi:10.1007/s41666-018-0043-8.
48. Lopes, P.; Bastiao, L.; Oliveira, J.L. i2x: An Automated Real-Time Integration and Interoperability Platform (Short Paper). In Proceedings of the 2015 IEEE 8th International Conference on Service-Oriented Computing and Applications (SOCA), Rome, Italy, 19–21 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 26–30.
49. Yuan, J.; Holtz, C.; Smith, T.H.; Luo, J. Autism spectrum disorder detection from semi-structured and unstructured medical data. *EURASIP J. Bioinform. Syst. Biol.* **2016**, *2017*, 3, doi:10.1186/s13637-017-0057-1.
50. Daniel, B.K. Big Data and data science: A critical review of issues for educational research. *Br. J. Educ. Technol.* **2019**, *50*, 101–113, doi:10.1111/bjet.12595.
51. Dahdouh, K.; Dakkak, A.; Oughdir, L.; Messaoudi, F. Big data for online learning systems. *Educ. Inf. Technol.* **2018**, *23*, 2783–2800, doi:10.1007/s10639-018-9741-3.
52. Patel, J.A.; Sharma, P. Big Data Harmonization–Challenges and Applications. *Int. J. Recent Innov. Trends Comput. Commun.* **2017**, *5*, 206–208.
53. Alguliyev, R.M.; Aliguliyev, R.M.; Hajirahimova, M. Big data integration architectural concepts for oil and gas industry. In Proceedings of the 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 12–14 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–5.
54. Wang, L. Big Data Analytics for Disparate Data. *Am. J. Intell. Syst.* **2017**, 7, 39–46.
55. Souza, A.; Pereira, J.; Oliveira, J.; Trindade, C.; Cavalcante, E.; Cacho, N.; Batista, T.; Lopes, F. A data integration approach for smart cities: The case of natal. In Proceedings of the 2017 International Smart Cities Conference (ISC2), Wuxi, China, 14–17 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
56. Scheurwegs, E.; Luyckx, K.; Luyten, L.; Daelemans, W.; Van den Bulcke, T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *J. Am. Med. Inform. Assoc.* **2016**, *23*, e11–e19, doi:10.1093/jamia/ocv115.
57. Jayaratne, M.; Nallaperuma, D.; De Silva, D.; Alahakoon, D.; Devitt, B.; Webster, K.E.; Chilamkurti, N. A data integration platform for patient-centered e-healthcare and clinical decision support. *Futur. Gener. Comput. Syst.* **2019**, *92*, 996–1008, doi:10.1016/j.future.2018.07.061.
58. Hong, N.; Wen, A.; Stone, D.J.; Tsuji, S.; Kingsbury, P.R.; Rasmussen, L.V.; Pacheco, J.A.; Adekkanattu, P.; Wang, F.; Luo, Y.; et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J. Biomed. Inform.* **2019**, *99*, 103310, doi:10.1016/j.jbi.2019.103310.
59. Hong, N.; Li, D.; Yu, Y.; Xiu, Q.; Liu, H.; Jiang, G. A computational framework for converting textual clinical diagnostic criteria into the quality data model. *J. Biomed. Inform.* **2016**, *63*, 11–21, doi:10.1016/j.jbi.2016.07.016.
60. Hong, N.; Wen, A.; Shen, F.; Sohn, S.; Liu, S.; Liu, H.; Jiang, G. Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data. *AMIA Summits Transl. Sci. Proc.* **2018**, *2018*, 74.
61. Chen, Z.; Zhong, F.; Yuan, X.; Hu, Y. Framework of integrated big data: A review. In Proceedings of the 2016 IEEE International Conference on Big Data Analysis (ICBDA), Hangzhou, China, 12–14 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–5.

62. Legaz-García, M.D.C.; Giménez, J.A.M.; Menárguez-Tortosa, M.; Fernández-Breis, J.T. Generation of open biomedical datasets through ontology-driven transformation and integration processes. *J. Biomed. Semant.* **2016**, *7*, 1–17, doi:10.1186/s13326-016-0075-z.
63. Rasitha, G.; Kuppuswamy, P.; Sasikala, N. Implementation of Big Data in Health Information Systems: Sample Approaches in Saudi Hospital. *Int. J. Comput. Appl.* **2017**, *160*, 1–4, doi:10.5120/ijca2017912917.
64. Hong, N.; Wen, A.; Shen, F.; Sohn, S.; Wang, C.; Liu, H.; Jiang, G. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. *JAMIA Open* **2019**, *2*, 570–579, doi:10.1093/jamiaopen/ooz056.
65. Chondrogiannis, E.; Andronikou, V.; Karanastasis, E.; Varvarigou, T. A Novel Approach for Clinical Data Harmonization. In Proceedings of the 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan, 27 February–2 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
66. Tekli, J. An Overview on XML Semantic Disambiguation from Unstructured Text to Semi-Structured Data: Background, Applications, and Ongoing Challenges. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1383–1407, doi:10.1109/tkde.2016.2525768.
67. Sanyal, M.K.; Bhadra, S.K.; Das, S. A Conceptual Framework for Big Data Implementation to Handle Large Volume of Complex Data. In *Information Systems Design and Intelligent Applications*; Springer: New Delhi, India, 2016; pp. 455–465.
68. Adduru, V.; Hasan, S.A.; Liu, J.; Ling, Y.; Datla, V.V.; Qadir, A.; Farri, O. Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In *KHD@ IJCAI*; **2018**.
69. Mujtaba, G.; Shuib, N.L.M.; Idris, N.; Hoo, W.L.; Raj, R.G.; Khowaja, K.; Shaikh, K.; Nweke, H.F. Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst. Appl.* **2019**, *116*, 494–520, doi:10.1016/j.eswa.2018.09.034.
70. Wang, Y.; Afzal, N.; Fu, S.; Wang, L.; Shen, F.; Rastegar-Mojarad, M.; Liu, H. MedSTS: A resource for clinical semantic textual similarity. *Lang. Resour. Eval.* **2020**, *54*, 57–72, doi:10.1007/s10579-018-9431-1.
71. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease Prediction by Machine Learning Over Big Data from Healthcare Communities. *IEEE Access* **2017**, *5*, 8869–8879, doi:10.1109/access.2017.2694446.
72. Pathak, B.; Lal, N. Information retrieval from heterogeneous data sets using moderated IDF-cosine similarity in vector space model. In Proceedings of the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 1–2 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3793–3799.
73. Torfi, A.; Shirvani, R.A.; Keneshloo, Y.; Tavvaf, N.; Fox, E.A. Natural Language Processing Advancements by Deep Learning: A Survey. *arXiv* **2020**, arXiv:2003.01200.
74. Wu, Y.; Zhao, S.; Li, W. Phrase2Vec: Phrase embedding based on parsing. *Inf. Sci.* **2020**, *517*, 100–127, doi:10.1016/j.ins.2019.12.031.
75. Moscatelli, M.; Manconi, A.; Pessina, M.; Fellegara, G.; Rampoldi, S.; Milanesi, L.; Casasco, A.; Gnocchi, M. An infrastructure for precision medicine through analysis of big data. *BMC Bioinform.* **2018**, *19*, 351, doi:10.1186/s12859-018-2300-5.
76. Chen, Q.; Du, J.; Kim, S.; Wilbur, W.J.; Lu, Z. Combining rich features and deep learning for finding similar sentences in electronic medical records. *In Proceedings of the BioCreative/OHNLP Challenge*, **2018**; pp. 5–8.
77. Mahlawi, A.Q.; Sasi, S. Structured data extraction from emails. In Proceedings of the 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), Thiruvananthapuram, India, 20–22 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 323–328.
78. Eke, C.I.; Norman, A.A.; Shuib, L.; Nweke, H.F. Sarcasm identification in textual data: Systematic review, research challenges and open directions. *Artif. Intell. Rev.* **2020**, *53*, 4215–4258, doi:10.1007/s10462-019-09791-8.
79. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv* **2017**, arXiv:1707.02919.
80. García, S.; Ramírez-Gallego, S.; Luengo, J.; Benítez, J.M.; Herrera, F. Big data preprocessing: Methods and prospects. *Big Data Anal.* **2016**, *1*, 9, doi:10.1186/s41044-016-0014-0.
81. Harish, B.S.; Rangan, R.K. A comprehensive survey on Indian regional language processing. *SN Appl. Sci.* **2020**, *2*, 1–16, doi:10.1007/s42452-020-2983-x.
82. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative study of cnn and rnn for natural language processing. *arXiv* **2017**, arXiv:1702.01923.
83. Ouyang, E.; Li, Y.; Jin, L.; Li, Z.; Zhang, X. Exploring n-gram character presentation in bidirectional RNN-CRF for chinese clinical named entity recognition. In *CEUR Workshop Proceedings*; **2017**; *Volume 1976*, pp. 37–42.
84. Lopez, M.M.; Kalita, J. Deep Learning applied to NLP. *arXiv* **2017**, arXiv:1703.03091.
85. Chai, J.; Li, A. Deep Learning in Natural Language Processing: A State-of-the-Art Survey. In Proceedings of the 2019 International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, 7–10 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
86. Oshikawa, R.; Qian, J.; Wang, W.Y. A survey on natural language processing for fake news detection. *arXiv* **2018**, arXiv:1811.00770.
87. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75, doi:10.1109/mci.2018.2840738.
88. Guan, M.; Cho, S.; Petro, R.; Zhang, W.; Pasche, B.; Topaloglu, U. Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes. *JAMIA Open* **2019**, *2*, 139–149, doi:10.1093/jamiaopen/ooy061.

89. Jing, K.; Xu, J. A survey on neural network language models. *arXiv* **2019**, arXiv:1906.03591.
90. Patel, S.; Patel, A. Deep Leaning Architectures and its Applications: A Survey. *Int. J. Comput. Sci. Eng.* **2018**, *6*, 1177–1183, 2018.
91. Klasnja-Milicevic, A.; Ivanović, M.; Budimac, Z. Data science in education: Big data and learning analytics. *Comput. Appl. Eng. Educ.* **2017**, *25*, 1066–1078, doi:10.1002/cae.21844.
92. Zhang, L.; Xie, Y.; Xidao, L.; Zhang, X. Multi-source heterogeneous data fusion. In Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 26–28 May 2018; *IEEE: Piscataway*, NJ, USA, **2018**; pp. 47–51.
93. Prasetya, D.; Wibawa, A.P.; Hirashima, T. The performance of text similarity algorithms. *Int. J. Adv. Intell. Inform.* **2018**, *4*, 63–69, doi:10.26555/ijain.v4i1.152.
94. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the International Conference on Machine Learning (ICML), Bellevue, WA, USA, 28 June–2 July 2011.
95. Wang, X.; Branciamore, S.; Gogoshin, G.; Ding, S.; Rodin, A.S. New Analysis Framework Incorporating Mixed Mutual Information and Scalable Bayesian Networks for Multimodal High Dimensional Genomic and Epigenomic Cancer Data. *Front. Genet.* **2020**, *11*, 648, doi:10.3389/fgene.2020.00648.
96. Shirzad, A.; Zare, H.; Teimouri, M. Deep Learning approach for text, image, and GIF multimodal sentiment analysis. In Proceedings of the 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 29–30 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 419–424.
97. Sunkara, M.; Ronanki, S.; Bekal, D.; Bodapati, S.; Kirchhoff, K. Multimodal Semi-Supervised Learning Framework for Punctuation Prediction in Conversational Speech. *arXiv* **2020**, arXiv:2008.00702.
98. Xu, J.; Huang, F.; Zhang, X.; Wang, S.; Li, C.; Li, Z.; He, Y. Sentiment analysis of social images via hierarchical deep fusion of content and links. *Appl. Soft Comput.* **2019**, *80*, 387–399, doi:10.1016/j.asoc.2019.04.010.
99. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133, doi:10.1016/j.knosys.2018.07.041.
100. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S. SpotFake: A Multi-modal Framework for Fake News Detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 39–47.
101. Kumar, A.; Garg, G. Sentiment analysis of multimodal twitter data. *Multimed. Tools Appl.* **2019**, *78*, 24103–24119, doi:10.1007/s11042-019-7390-1.
102. Bougiatiotis, K.; Giannakopoulos, T. Enhanced movie content similarity based on textual, auditory and visual information. *Expert Syst. Appl.* **2018**, *96*, 86–102, doi:10.1016/j.eswa.2017.11.050.
103. Nie, W.; Ding, H.; Song, D.; Long, X. Location emotion recognition for travel recommendation based on social network. *Signal Image Video Process.* **2019**, *13*, 1259–1266, doi:10.1007/s11760-019-01457-w.