Data Harmonization

Subjects: Computer Science, Information Systems
Contributor: Ganesh Kumar

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.

Keywords: data harmonization ; heterogeneous data ; text preprocessing

1. Introduction

Big Data plays a vital role in the assessment of massive data produced every second by real-world applications, using tools and algorithms [1]. Some of the 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][8][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][26][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.

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

In this part, 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 [28], using semantic web and BD tools. Where the performance of the precision, recall, and F-score were 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 [29], 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 [30] 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 [31] summarized that the heterogeneous data in the 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. [32] revealed a Web-based FHIR visualization tool, using a standard structured format API. Again, Lopes, Bastião, and Oliveira [33] 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 [34] 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 [35] also emphasized 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 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 [36] with the help of Hadoop, MapReduce, and Spark. As a result, it helps in taking a smart decision within seconds. Additionally, Patel and Sharma [37] 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 [38].

Wang [39] 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 were mentioned by Souza et al. [40]; 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 a health record into an EHR reported by authors (Scheurwegs, Luyckx, Luyten, Daelemans, and Van den Bulcke [41]. By using Naïve base and Random Forest on the UZA dataset, the patient classification was performed. Similarly, the researchers Jayaratne et al. [42], 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. [43] 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. [44] 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. [45] used a Mayo Clinic dataset with the help of NLP toolkits for making a digital FHIR system. In contrast, Chen, Zhong, Yuan, and Hu [46] 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 [47].

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 [48] team proposed 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 researcher Hong, Wen, Shen, et al. [49]. The contributions of all studies in all domains are discussed in **Table 7**.

Table 1. RQ2 domain and contributions.

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

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 have been 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 [51] 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 [52] pointed out that, by using the 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. [53]. 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. [54] 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. [55]. 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 [56] 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 [57] 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 [58]. 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 [59] 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. [60] 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 [61] 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, Malawi and Sasi [62] found that, from a 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 [63] 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. [64]. This led authors García, Ramírez-Gallego, Luengo, Benítez, and Herrera [65] to focus on Indian regional multilingual data processed with the help of natural-language-processing techniques. Finally, Harish and Rangan [66] 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

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

In this section, 8 studies have been selected that 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. [67] and Ouyang et al. [68], 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 the model is up to 93%. Moreover, a comparison of CNN and RNN reveals that RNN is better than CNN. The techniques that can be used for sentimental, relational, textual entitlement, answer selection, QA path query, and POS tagging were pointed by Lopez and

Kalita [69]. Additionally, the authors Chai and Li [70] 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 show better results which were presented by authors Oshikawa, Qian, and Wang [71]. 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 [72][73]. lastly, a survey was conducted by the authors Jing and Xu Jing and Xu [74] [75] 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 2**.

Table 2. Model Performance Techniques.

Study Reference	Domain	Techniques	Contributions
[67]	General	CNN, RNN for NLP	RNN perform better
[68]	Healthcare	RNN, N-Gram	RNN performance better by using N-gram
[69]	General	Compared with the existing Algorithm of CNN	RNN outperformed
[70]	General	Used in many NLP and audio-video functionality	Better for sequential text
[71]	Fake News	RNN for larger data sets of fake news	93% accuracy
[72]	General	CNN, RNN	RNN is better as per recent studies
[73]	Cancer, healthcare	DL classifier is better than conventional classifier	Model accuracy is better by using RNN
[74]	General	FFNNLM, RNNLM	RNN Language model is best
[75]	Medical, General	CNN, DBN, RNN	RNN is better in terms of NLP

References

1. Avci, C.; Tekinerdogan, B.; Athanasiadis, I.N. Software architectures for big data: A systematic literature review. Big Dat a Anal. 2020, 5, 1–53.
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 Application s 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, 1 201–1227.
6. Younan, M.; Houssein, E.H.; Elhoseny, M.; Ali, A.A. Challenges and recommended technologies for the industrial intern et of things: A comprehensive review. Measurement 2020, 151, 107198.
7. Wang, Y.; Jan, M.N.; Chu, S.; Zhu, Y. Use of Big Data Tools and Industrial Internet of Things: An Overview. Sci. Progra m. 2020, 2020, 1–10.
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 online: https://ssrn.com/abstract=3603739 (accessed on 18 May 2020).
9. Ralph, B.; Stockinger, M. Digitalization and digital transformation in metal forming: Key technologies, challenges and cu rrent developments of industry 4.0 applications. In Proceedings of the XXXIX, Colloquium on Metal Forming, Leoben, A ustria, 21–25 March 2020.
10. Kraus, J.M.; Lausser, L.; Kuhn, P.; Jobst, F.; Bock, M.; Halanke, C.; Hummel, M.; Heuschmann, P.; Kestler, H.A. Big dat a and precision medicine: Challenges and strategies with healthcare data. Int. J. Data Sci. Anal. 2018, 6, 241–249.

11. Kourou, K.D.; Pezoulas, V.C.; Georga, E.I.; Exarchos, T.P.; Tsanakas, P.; Tsiknakis, M.; Varvarigou, T.; De Vita, S.; Tzio ufas, A.; Fotiadis, D.I.I. Cohort Harmonization and Integrative Analysis from a Biomedical Engineering Perspective. IEE E Rev. Biomed. Eng. 2018, 12, 303–318.
12. Stoyanova, M.; Nikoloudakis, Y.; Panagiotakis, S.; Pallis, E.; Markakis, E.K. A Survey on the Internet of Things (IoT) For ensics: Challenges, Approaches, and Open Issues. IEEE Commun. Surv. Tutor. 2020, 22, 1191–1221.
13. Xiong, H.; Wu, Y.; Jin, C.; Kumari, S. Efficient and Privacy-Preserving Authentication Protocol for Heterogeneous Syste ms in IIoT. IEEE Internet Things J. 2020, 7, 11713–11724.
14. Sahu, A.K.; Sahu, A.K.; Sahu, N.K. A Review on the Research Growth of Industry 4.0: IIoT Business Architectures Benc hmarking. 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 t he 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. Edu c. Technol. 2021, 11, 242–247.
17. Sajid, S.; Haleem, A.; Bahl, S.; Javaid, M.; Goyal, T.; Mittal, M. Data science applications for predictive maintenance an d 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 Chal lenges. Logistics 2020, 5, 2.
19. Sedkaoui, S.; Khelfaoui, M. Industry 4.0 and knowledge management practices. Volto Já–Senior Exchange Program: F rom Idea To Implementation. In Proceedings of the International Conference on Management Technology and Tourism, ICOMTT, Santarém, Portugal, 6–7 February 2020; p. 47.
20. De Vass, T.; Shee, H.; Miah, S. IoT in Supply Chain Management: Opportunities and Challenges for Businesses in Earl y Industry 4.0 Context. Oper. Supply Chain Manag. Int. J. 2021, 14, 148–161.
21. Shao, X.-F.; Liu, W.; Li, Y.; Chaudhry, H.R.; Yue, X.-G. Multistage implementation framework for smart supply chain ma nagement under industry 4.0. Technol. Forecast. Soc. Chang. 2021, 162, 120354.
22. Andrade, T.; Bastos, D. Extended reality in iot scenarios: Concepts, applications and future trends. In Proceedings of th e 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. 2018. Available onlin e: https://ssrn.com/abstract=3300469 or http://dx.doi.org/10.2139/ssrn.3300469 (accessed on 28 August 2021).
24. Gong, L.; Fast-Berglund, A.; Johansson, B. A Framework for Extended Reality System Development in Manufacturing. I EEE Access 2021, 9, 24796–24813.
25. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans. Pattern An al. Mach. Intell. 2018, 41, 423–443.
26. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fu sion. Inf. Fusion 2017, 37, 98–125.
27. Shoumy, N.J.; Ang, L.-M.; Seng, K.P.; Rahaman, D.; Zia, T. Multimodal big data affective analytics: A comprehensive su rvey using text, audio, visual and physiological signals. J. Netw. Comput. Appl. 2020, 149, 102447.
28. Danyaro, K.U.; Liew, M.S. A Proposed Methodology for Integrating Oil and Gas Data Using Semantic Big Data Technol ogy. In International Conference of Reliable Information and Communication Technology; Springer: Cham, Switzerland, 2017; pp. 30–38.
29. Sambrekar, K.; Rajpurohit, V.S.; Joshi, J. A Proposed Technique for Conversion of Unstructured Agro-Data to Semi-Stru ctured 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.
30. Saggi, M.K.; Jain, S. A survey towards an integration of big data analytics to big insights for value-creation. Inf. Proces s. Manag. 2018, 54, 758–790.
31. Li, C.; Chai, W.; Chen, L. An Integration Model of Multi-Source Heterogeneous Audit Data. In Proceedings of the 2015 I nternational Conference on Electronic Science and Automation Control, Zhengzhou, China, 15–16 August 2015; Atlanti s Press: Amsterdam, The Netherlands, 2015.
32. Hong, N.; Wang, K.; Wu, S.; Shen, F.; Yao, L.; Jiang, G. An Interactive Visualization Tool for HL7 FHIR Specification Br owsing and Profiling. J. Healthc. Informa. Res. 2019, 3, 329–344.

33. 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.
34. 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.
35. Daniel, B.K. Big Data and data science: A critical review of issues for educational research. Br. J. Educ. Technol. 2019, 50, 101–113.
36. Dahdouh, K.; Dakkak, A.; Oughdir, L.; Messaoudi, F. Big data for online learning systems. Educ. Inf. Technol. 2018, 23, 2783–2800.
37. Patel, J.A.; Sharma, P. Big Data Harmonization–Challenges and Applications. Int. J. Recent Innov. Trends Comput. Commun. 2017, 5, 206–208.
38. 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.
39. Wang, L. Big Data Analytics for Disparate Data. Am. J. Intell. Syst. 2017, 7, 39–46.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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, 32.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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 IJCAI. 2018. Available online: https://www.semanticscholar.org/paper/Towards-Dataset-Creation-And-Establishing-Baselines-Adduru-Hasan/469b714845898aa23c53485ca91bd41aecbe8de3 (accessed on 28 August 2021).

54. Mujtaba, G.; Shuib, N.L.M.; Idris, N.; Hoo, W.L.; Raj, R.G.; Khowaja, K.; Shaikh, K.; Nweke, H.F. Clinical text classificati on research trends: Systematic literature review and open issues. Expert Syst. Appl. 2019, 116, 494–520.
55. 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.
56. Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease Prediction by Machine Learning Over Big Data from Healthc are Communities. IEEE Access 2017, 5, 8869–8879.
57. Pathak, B.; Lal, N. Information retrieval from heterogeneous data sets using moderated IDF-cosine similarity in vector s pace 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.
58. Torfi, A.; Shirvani, R.A.; Keneshloo, Y.; Tavvaf, N.; Fox, E.A. Natural Language Processing Advancements by Deep Lea rning: A Survey. arXiv 2020, arXiv:2003.01200.
59. Wu, Y.; Zhao, S.; Li, W. Phrase2Vec: Phrase embedding based on parsing. Inf. Sci. 2020, 517, 100–127.
60. Moscatelli, M.; Manconi, A.; Pessina, M.; Fellegara, G.; Rampoldi, S.; Milanesi, L.; Casasco, A.; Gnocchi, M. An infrastr ucture for precision medicine through analysis of big data. BMC Bioinform. 2018, 19, 351.
61. 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. Available online: http s://www.researchgate.net/publication/327402060_Combining_rich_features_and_deep_learning_for_finding_similar_se ntences_in_electronic_medical_records (accessed on 28 August 2021).
62. 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: P iscataway, NJ, USA, 2017; pp. 323–328.
63. Eke, C.I.; Norman, A.A.; Shuib, L.; Nweke, H.F. Sarcasm identification in textual data: Systematic review, research chall enges and open directions. Artif. Intell. Rev. 2020, 53, 4215–4258.
64. 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.
65. García, S.; Ramírez-Gallego, S.; Luengo, J.; Benítez, J.M.; Herrera, F. Big data preprocessing: Methods and prospects. Big Data Anal. 2016, 1, 9.
66. Harish, B.S.; Rangan, R.K. A comprehensive survey on Indian regional language processing. SN Appl. Sci. 2020, 2, 12 04.
67. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative study of cnn and rnn for natural language processing. arXiv 2017, a rXiv:1702.01923.
68. Ouyang, E.; Li, Y.; Jin, L.; Li, Z.; Zhang, X. Exploring n-gram character presentation in bidirectional RNN-CRF for chine se clinical named entity recognition. CEUR Workshop Proc. 2017, 1976, 37–42.
69. Lopez, M.M.; Kalita, J. Deep Learning applied to NLP. arXiv 2017, arXiv:1703.03091.
70. Chai, J.; Li, A. Deep Learning in Natural Language Processing: A State-of-the-Art Survey. In Proceedings of the 2019 In ternational Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, 7–10 July 2019; IEEE: Piscataw ay, NJ, USA, 2019; pp. 1–6.
71. Oshikawa, R.; Qian, J.; Wang, W.Y. A survey on natural language processing for fake news detection. arXiv 2018, arXi v:1811.00770.
72. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. I EEE Comput. Intell. Mag. 2018, 13, 55–75.
73. 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.
74. Jing, K.; Xu, J. A survey on neural network language models. arXiv 2019, arXiv:1906.03591.
75. Patel, S.; Patel, A. Deep Leaning Architectures and its Applications: A Survey. Int. J. Comput. Sci. Eng. 2018, 6, 1177–1 183.