future internet

MDPI

Editorial

Data Science and Knowledge Discovery

Filipe Portela [1,2]

[1] Algoritmi Research Centre, University of Minho, 4800-058 Guimarães, Portugal; cfp@dsi.uminho.pt
[2] IOTECH—Innovation on Technology, 4785-588 Trofa, Portugal

Abstract: Nowadays, Data Science (DS) is gaining a relevant impact on the community. The most recent developments in Computer Science, such as advances in Machine and Deep Learning, Big Data, Knowledge Discovery, and Data Analytics, have triggered the development of several innovative solutions (e.g., approaches, methods, models, or paradigms). It is a trending topic with many application possibilities and motivates the researcher to conduct experiments in these most diverse areas. This issue created an opportunity to expose some of the most relevant achievements in the Knowledge Discovery and Data Science field and contribute to such subjects as Health, Smart Homes, Social Humanities, Government, among others. The relevance of this field can be easily observed by its current achieved numbers: thirteen research articles, one technical note, and forty-six authors from fifteen nationalities.

1. Introduction

The importance and impact of Data Science (DS) in the decision process are significantly increasing. DS is an interdisciplinary field that combines various areas, including Computer Science, Machine Learning, Math and Statistics, domain/business knowledge, software development, and traditional research. DS applies scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data as a research topic.

Knowledge Discovery (KD) is the basis of Data Science and consists of creating knowledge from structured and unstructured sources (e.g., text, data, and images). The output needs to be in a readable and interpretable format. It must represent knowledge in a manner that facilitates inferencing. This new trend is being explored in several areas, such as education, health, accounting, energy, and public administration. In this context, this Special Issue arises as an excellent opportunity to provide scientific knowledge and disseminate the findings and achievements through several communities.

This Special Issue discusses this trending topic and presents innovative solutions to show the importance of Data Science and Knowledge Discovery to researchers, managers, industry, society, and other communities. Through invited and open call submissions, a total of fourteen excellent articles have been accepted, following a rigorous review process that required a minimum of three reviews and at least one revision round for each paper.

2. Contributions

The first paper, written by Theodora A. Maniou and Andreas Veglis [1] and entitled Employing a Chatbot for News Dissemination during Crisis: Design, Implementation, and Evaluation, presents some benefits of using chatbots by journalists and media professionals. It shows the advantages of implementing chatbots in news platforms during a crisis when the audience's need for timely and accurate information rapidly increases. This study was evaluated using two metrics: the technical effort of creating a functional and robust news chatbot and, the second, users' perception regarding the appropriation of this news chatbot. The participants involved in the case study agreed that the COVINFO Reporter's accessibility was very good, and they experienced no problems navigating the chatbot.

check for updates

Citation: Portela, F. Data Science and Knowledge Discovery. *Future Internet* **2021**, *13*, 178. https://doi.org/10.3390/fi13070178

Received: 27 May 2021
Accepted: 5 July 2021
Published: 7 July 2021

The second paper, entitled Visualization, Interaction and Analysis of Heterogeneous Textbook Resources, and written by Christian Scheel, Francesca Fallucchi, and Ernesto William De Luca [2], proposes a Component Metadata Infrastructure (CMDI)-based approach. The authors used this approach for data rescue and reuse, where data is retroactively joined into one repository, minimizing future research projects' implementation efforts. While the data is precious, it cannot be used by any service—except the prepared tool. With this approach, the authors want to increase data understanding, sustainability and reusability, and reduce data silos.

Antonio Maria Rinaldi, Cristiano Russo, and Cristian Tommasino wrote the third paper of this issue: Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features [3]. In this work, the authors analyzed studies in the semantic research field based on ontologies. They considered that, although modern search engines provide visual queries, it is not easy to find systems that allow searching from a particular domain of interest and that perform such searches by combining text and visual questions. Then, the authors proposed a novel approach for semantic image retrieval that included a possible combination with multimedia document analysis. This paper presents the method developed and several results to show its performance compared with the state of the art.

Alan Ponce and Raul Alberto Ponce Rodriguez wrote the fourth paper, an Analysis of the Supply of Open Government Data [4]. This work presents an analysis based on the index of the release of open government data, published in 2016 by the Open Knowledge Foundation, which shows a significant variability in the country's supply of open data. The authors used several linear regression models to explain the cross-country differences. This work provides evidence that the country's civil liberties, government transparency, quality of democracy, efficiency of government intervention, and economies of scale in the provision of public goods, as well as the size of the economy, are the most statistically important reasons for differences in the supply of open government data.

The following paper, the fifth one, is entitled Geospatial Assessment of the Territorial Road Network by Fractal Method [5] and was written by Mikolaj Karpinski, Svitlana Kuznichenko, Nadiia Kazakova, Oleksii Fraze-Frazenko, and Daniel Jancarczyk. This paper proposes an approach to the geospatial assessment of a territorial road network based on the fractal theory. The method allows calculation of the fractal dimension based on a combination of box-counting and GIS analysis. The authors created a geoprocessing script tool for the GIS software system ESRI ArcGIS 10.7 using the spatial pattern of the transport network of the Ukraine territory and other countries of the world. The study results help to better understand the different aspects of the development of transport networks, their changes over time, and the impact on the socioeconomic indicators of urban development.

The sixth paper was written by José Paulo Lousado and Sandra Antunes and is entitled S. Monitoring and Support for Elderly People Using LoRa Communication Technologies: IoT Concepts and Applications [6]. This paper is the second article motivated by the SARS-CoV-2 virus (COVID-19) and aims to show an implementation of low-cost technologies, which make it possible to answer a fundamental question: how can near real-time monitoring and follow-up of the elderly and their health conditions, as well as their homes, especially for those living in isolated and remote areas, be provided within their care and protect them from risky events? The proposed system uses low-cost devices for communication and data processing, supported by Long-Range (LoRa) technology, and incorporates various sensors, both personal and in residence. It allows family members, neighbors, and authorized entities (including security forces) to have access to the health condition of system users and the habitability of their homes, as well as their urgent needs. This article shows that it is possible to implement sensor networks to monitor the elderly using the LoRa gateway and other low-cost infrastructures.

The seventh publication, entitled About Rule-Based Systems: Single Database Queries for Decision Making, is a technical note written by Piotr Artiemjew, Lada Rudikova, and Oleg Myslivets [7]. It explores the implementation of artificial intelligence systems for

manipulating data and the surrounding world in a more complex way. In this work, the authors addressed the possibility of placing the rule-based learned model of decision support in a SQL database environment. They propose a universal solution for any IF-THEN rule induction algorithm to place the previously trained model in the database and apply it by employing single queries.

Sook-Ling Chua, Lee Kien Foo, and Hans W. Guesgen wrote the eighth paper—Predicting Activities of Daily Living with Spatio-Temporal Information [8]. This paper is framed in smart homes and shows the importance of having spatial and temporal information for reasoning. The authors created a method for predicting user activities given the spatial and temporal information and explained how it could be represented for activity recognition. The method was evaluated using three publicly available smart-home datasets and achieved an average accuracy of more than 81%.

The following article, the ninth one, addresses the Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce and was written by Laith T. Khrais [9]. This article explores the use of AI in e-commerce, where ethical soundness is a contentious issue, especially regarding the concept of explainability. The study adopted the use of word cloud analysis, voyance analysis, and concordance analysis to gain a detailed understanding of the idea of explainability as has been utilized by researchers in the context of AI. Motivated by a corpus analysis, the authors formulated the Explainable Artificial Intelligence (XAI) model that provides insights into the decision points, variables, and data used to produce recommendations. This study also suggests that the Machine Learning models should be improved by making them interpretable and comprehensible, and allowing them to deploy explainable XAI systems.

Jing Wang, Zhong Cheng Wu, Fang Li, and Jun Zhang present the tenth article entitled Data Augmentation approach to Distracted Driving Detection [10]. This work addresses the behavior problem associated with distracted driving that leads to vehicle crashes. To address this problem, the authors proposed a Data Augmentation method based on the driving operation area using the convolutional neural network classification model. The classification's result achieved a 96.97% accuracy using the distracted driving dataset. This method is helpful to detect drivers in actual application scenarios and identify dangerous driving behaviors. It helps to give an early warning of unsafe driving behaviors and avoid accidents.

The eleventh paper, entitled Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks was written by Aleksandr Romanov, Anna Kurtukova, Alexander Shelupanov, Anastasia Fedotova, and Valery Goncharov [11]. The authors explored the advantages and disadvantages of various approaches that can determine the author of a natural language text. Some of the examples found were used to identify authors of suicide notes, conduct forensic exams, and detect plagiarism. This article describes the process of identifying the author of Russian-language texts using support vector machine (SVM) and deep neural network architectures, as well as convolutional neural networks (CNN) with attention networks and transformers. The results show that all the considered algorithms are suitable for solving the authorship identification problem, but SVM offers the best accuracy. The average accuracy of SVM reaches 96%.

Nuno Marques da Costa, Nelson Mileu, and André Alves present article number twelve, entitled Dashboard COMPRIME_COMPRI_MOv: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal [12]. This article, the third in this Special Issue about the pandemic, shows a set of dashboards to disseminate information and multi-scale knowledge of COVID-19. As a result, the authors developed a system for monitoring the evolution of the pandemic. The constructed platform dynamically and interactively brings together a diverse set of variables and indicators that reflects the evolutionary behavior of the pandemic from a multi-scale perspective in Portugal. The authors mention that this approach proves to be crucial to guarantee everyone's access to information while simultaneously emerging as an epidemiological surveillance tool.

This tool can assist public authorities in terms of ensuring competent decision-making by defining control policies and fighting the spread of new coronavirus strains.

The article Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities, the thirteenth in this Special Issue, was written by Albert Weichselbraun, Philipp Kuntschik, Vincenzo Francolino, Mirco Saner, Urs Dahinden, and Vinzenz Wyss [13], and addressed the application of data-driven research methods to disciplines such as the Social Sciences and Humanities. The authors presented a case study that demonstrates the potential of the proposed method in the domain of Communication Science by creating approaches that aid domain experts in locating, tracking, analyzing, and finally, better understanding the dynamics of media criticism. The paper shows that data-driven research approaches require a tighter integration with the methodological framework of the target discipline to provide a significant impact on the target discipline.

The last article is the fourth study about covid19. Ana Teresa Ferreira, Carlos Fernandes, José Vieira, and Filipe Portela present the study Pervasive Intelligent Models to Predict the Outcome of COVID-19 Patients [14]. The authors induced intelligent models capable of predicting and supporting clinical decisions to predict if the patient will die or recover from COVID-19. The best scenario is composed of all comorbidities, symptoms, and ages. The best model achieved a sensitivity of 95.20%, accuracy of 90.67%, and specificity of 86.08%. The models were deployed as a service and are part of a clinical decision support system named ioCOVID19, which is available for authorized users anywhere and anytime.

Funding: This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Maniou, T.; Veglis, A. Employing a Chatbot for News Dissemination during Crisis: Design, Implementation and Evaluation. *Future Internet* **2020**, *12*, 109. [CrossRef]
2. Scheel, C.; Fallucchi, F.; De Luca, E. Visualization, Interaction and Analysis of Heterogeneous Textbook Resources. *Future Internet* **2020**, *12*, 176. [CrossRef]
3. Rinaldi, A.; Russo, C.; Tommasino, C. A Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features. *Future Internet* **2020**, *12*, 183. [CrossRef]
4. Ponce, A.; Ponce Rodriguez, R. An Analysis of the Supply of Open Government Data. *Future Internet* **2020**, *12*, 186. [CrossRef]
5. Karpinski, M.; Kuznichenko, S.; Kazakova, N.; Fraze-Frazenko, O.; Jancarczyk, D. Geospatial Assessment of the Territorial Road Network by Fractal Method. *Future Internet* **2020**, *12*, 201. [CrossRef]
6. Lousado, J.; Antunes, S. Monitoring and Support for Elderly People Using LoRa Communication Technologies: IoT Concepts and Applications. *Future Internet* **2020**, *12*, 206. [CrossRef]
7. Artiemjew, P.; Rudikova, L.; Myslivets, O. About Rule-Based Systems: Single Database Queries for Decision Making. *Future Internet* **2020**, *12*, 212. [CrossRef]
8. Chua, S.; Foo, L.; Guesgen, H. Predicting Activities of Daily Living with Spatio-Temporal Information. *Future Internet* **2020**, *12*, 214. [CrossRef]
9. Khrais, L. Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce. *Future Internet* **2020**, *12*, 226. [CrossRef]
10. Wang, J.; Wu, Z.; Li, F.; Zhang, J. A Data Augmentation Approach to Distracted Driving Detection. *Future Internet* **2021**, *13*, 1. [CrossRef]
11. Romanov, A.; Kurtukova, A.; Shelupanov, A.; Fedotova, A.; Goncharov, V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. *Future Internet* **2021**, *13*, 3. [CrossRef]
12. Marques da Costa, N.; Mileu, N.; Alves, A. Dashboard COMPRIME_COMPRI_MOv: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal. *Future Internet* **2021**, *13*, 45. [CrossRef]
13. Weichselbraun, A.; Kuntschik, P.; Francolino, V.; Saner, M.; Dahinden, U.; Wyss, V. Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities. *Future Internet* **2021**, *13*, 59. [CrossRef]
14. Ferreira, A.; Fernandes, C.; Vieira, J.; Portela, F. Pervasive Intelligent Models to Predict the Outcome of COVID-19 Patients. *Future Internet* **2021**, *13*, 102. [CrossRef]