Application of Latent Dirichlet Allocation Topic Model in Identifying 4IR Research Trends

Muthoni Masinde(✉)

Central University of Technology, Bloemfontein, Free State, South Africa
muthonimasinde@gmail.com

Abstract. The dynamic nature of the technologies associated with the fourth Industrial Revolution (4IR) presents complex scenarios for researchers, practitioners and policymakers alike. To this end, reaching decisions such as what technology to invest/train in could be made easier through a 4IR technology trend predictive tool. In this paper, we apply Latent Dirichlet Allocation (LDA) topic model to identify and predict trends in 4IR technologies. The LDA models were developed and trained using text composed of abstracts, titles and keywords retrieved from 11,7314-IR related to the 2012 to 2022 publications in the Web of Science database. The effectiveness of the resulting tool was then evaluated using text from email message distributed to subscribers of the IEEE's Tccc-announce mailing list. From the results, our model correctly identifies trends in the following 4IR technologies and applications domains: Internet of Things, Artificial Intelligence/Machine Learning, Big Data/Data Analytics, Augmented Reality, Smart Manufacturing, Supply Chains, Sustainability and Circular Economy. By plotting and visualizing these trends over time (2019 to 2022), the validation text confirms our tool's ability to identify the trajectory developments as identified by other similar tools such as Bibliometric Analysis.

Keywords: Latent Dirichlet Allocation (LDA) · Fourth Industrial Revolution (4IR) · Technology Trends · Bibliometric Analysis and Topic Models

1 Introduction

1.1 An Overview of the Fourth Industrial Revolution

Compared to previous industrial revolutions, the fourth industrial revolution (4IR) has received tremendous attention in its short period of existence [1–3]. The main differentiator of 4IR from the other three industrial revolutions, is the adoption of cyber-physical-systems (CPS) [4]. This is because CPS supports seamless integration of physical and computational worlds, which in turn enables the implementation of features such as adaptability, safety, security and scalability. While the research community has produced large number of publications that documents developments in systems, business models and methodologies, 4IR has received equal attention within the business community seeking to better the world of business through smarter, efficient, adaptive, secure

M. Masinde and A. Bagula (Eds.): AFRICATEK 2022, LNICST 503, pp. 74–94, 2023.
https://doi.org/10.1007/978-3-031-35883-8_6

products/services and environments [3]. Prior to 2020, 4IR was mostly associated with digital integration and intelligent engineering while its top components were CPS, additive manufacturing, virtual and augmented reality, cloud computing, big data analytics and data science.

The core enabling technologies of 4IR include: (1) internet of things (IoT) and related technologies such as Radio-frequency identification (RFID), sensors, actuators, mobile devices (and associated communication technologies such as Wi-Fi, Near-Field Communication (NFC)); (2) cloud computing; (3) CPS; and (4) industrial integration, enterprise architecture and enterprise application integration that every organization requires in order to transit to 4IR. In the latter, new business models are required in order to support the inevitable integration of 4IR. According to Xu et al. [4], these integrations will trigger changes in enterprise architecture, ICT integration and processes.

In terms of 4IR's application domains, smart manufacturing sector has dominated - with applications such as digital twin shop-floor [5], intelligent manufacturing [6, 7] and CPS in manufacturing [8]. There are also many applications in the logistics sector in form of smart supply chain, with examples such as those reported in [9–11]. In the last three years (2020 onwards), the focus of 4IR research includes innovative and business models [12–14], blockchain technologies [15, 16], application of augmented reality [17, 18] in different domains, human factors [17, 19–21] and sustainability [20, 21]. In relation to business models and sustainability, digital platforms [12, 22, 23] and circular economy [24, 25] are two strongly emerging areas.

The use of bibliometric analysis as way to identify trends from literature is widely documented [26, 27], even in the 4IR sub-field [2, 24, 28]. For instance, in Muhuri et al. [2], the authors carried out a bibliometric analysis of publications from the Web of Science (WoS) and Scopus databases using the search phrase "Industry 4.0" based on publications dating until 2017. However, research that uses topic models to assess trends in 4IR are limited [29, 30]. This is the gap this paper aims to fill.

1.2 Topic Modeling

In general, topic modeling aims to demonstrate inter-links in discrete data by discovering structural relations and meaning in voluminous information and data [31, 32]. Application domains for topic modeling span all spheres of life, for example, in political sciences and medical sciences. Others are in source code analysis, opinion and aspect mining [32]. Topic modeling finds its empirical grounding is in computer science's subfields of text mining and natural language processing. Topic modelling tools support statistical analysis of collection of documents [31]. The basic assumption is that a topic is a list of words where the latter refers to unstructured text from sources such as email, tweets and books. Topic models assume any part of the text is combined by selecting words from probable baskets of words – each basket in this case corresponds to a topic.

1.3 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is one of the most popular technique for topic modelling [32]. Furthermore, LDA is a probabilistic model of corpus in which each document is represented as a probabilistic distribution over Latent topics. In other words, each topic

is a distribution over words and a document is a mixture of topics. For a given application, the topic distribution in all documents share a common Dirichlet prior. Mathematical representation of the main terms (Corpus, Document, Topic and Number of Words) can be found in publications such as these ones [30–37]. One of the strengths of LDA is its ability to work with very large corpus and to identify sub-topics for technology area composed of many sub-topics. LDA achieves this by first generating terms in a set of documents then going further to generate a vocabulary to discover hidden topics. Given the nested nature of the 4IR topic, LDA was found to be the most appropriate. For parameter estimation inference, LDA applies either expectation propagation, variational method or Gibbs Sampling. Gibbs Sampling is the most commonly used; it employs the Monte Carlo Markov-Chain algorithm. Besides, LDA is supervised learning algorithm [32, 36].

LDA has been widely researched and has found intense application in social media analytics [32]. Different extensions of LDA have also emerged over the years, each with enhanced features. Some of these features include ability to capture correlation among topics, to classify documents, analyse documents in different languages and to analyse the temporal evaluation of topics in very large collection of documents. Some examples of these extensions are: (1) Dynamic Topic Model (DTM), which can vision the topic trend and (2) Labelled LDA (LLDA), which is supervised algorithm. Others are (3) Maximum Entropy Discrimination LDA (Med LDA) [34, 35] which applies hierarchical Bayesian Model concept and (4) Relational Topic Model (TRM) which focuses on networks of text data. Besides, there is a wide range of tools and software for implementing LDA modeling such as those listed in [32]. For this paper, the MATLAB implementation of LDA was used.

1.4 Dynamic Topic Models

Over and above the function of LDA, dynamic topic models capture how the meaningful patterns of words change over time [31]. In the implementation of Dynamic LDA (D-LDA), the use of probabilistic time series allow the topics to vary smoothly over time. The weakness of D-LDA has been reported as its inability to capture rare words as well as long tail of language data [31]. This problem has however been solved through Embedded Topic Model (ETM) where continuous representation of words is made use of [31, 38]. In [31], Dynamic ETM (D-ETM) [31] is introduced to address the problem that ETM cannot analyse a corpus whose topics shift over time. D-ETM works by building on word embedding topic models and dynamic topic models. Given the dynamic nature of 4IR technologies, a variation of D-ETM is considered the best option.

2 Data and Methods

2.1 Data Sources

The following two sources of data were used to extract the text for the LDA topic modeling.

The TCCC-ANNOUNCE Archives. One of the functions of the IEEE's Technical Committee on Computer Communications (TCCC) is to provide the members with a forum for technical discussions and interactions (https://tccc.committees.comsoc.org/) [39]. To this end, the Committee runs mailing lists (https://tccc.committees.comsoc.org/mailing-list/) such as the tccc-announce and tccc-discuss. The tccc-announce is used for announcements related to on-topic call for papers (CFPs) and faculty or research job openings. The key requirement for these announcements is to have a primary focus on networking and communication. Consequently, most of the topics covered by these CFPs cover the 4IR technologies. As of August 2022, the mailing list had 4,599 subscribers. Besides, the author of this paper has been receiving these announcements since 2011. The text data (typically containing emails of announcements) used in this paper was retrieved from tccc-announce archive as follows:

- File 1 – a combination of 5 archive files consisting the List's home at Columbia University from 2001 to 2013
- File 2 – an export of selected (2019 to August 2022) of tccc-announce emails received by the author.

Web of Science Core Collection. In order to identify the kind of research being carried out under the general topic of Fourth Industrial Revolution (4IR), publications were extracted from the Web of Science (WoS) core collection (https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/) using the search phrase: *ALL = (("4th industrial revolution" or "4IR" OR "Industry 4.0") and ("technologies" or "technology"))*. After applying filters such publication year (2012 to 2022) and language (English), a total of 11,731 publications were extracted. From these, text strings used for text processing in MATLAB were created by concatenating the following headings: ArticleTitle, SourceTitle, BookSeriesTitle, BookSeriesSubtitle, ConferenceTitle, AuthorKeyword, KeywordsPlus, Abstract and MeetingAbstract. These entries were stored in a comma separated value (CSV) file. A second file was exported in the Research Information System (RIS) format and later used for Bibliometric Analysis presented in Sect. 2.3

2.2 Text Pre-processing

The text data files were pre-processed using existing functions (e.g. *preprocessText(inputText)*) in MATLAB. Through this, common pre-processing functions were performed, including: converting data to lowercase, tokenization, erasing of punctuation marks and Lemmatization. The flow of these steps is presented in Fig. 1 below.

As shown in Fig. 1, the text files (e.g. "Tccc2010–2013.txt") were converted to strings of text, which were then passed over to the text pre-processing functions. The 'remove stop words' removes a list of stop words (such as "and", "of", and "the") from the input string. Given that the tccc-announce text files contains email announcements, the frequency of occurrence for common written-speech words (such as "international", "conference", "discussing", "correct", "grammar" "spelling", "indicate" and "contribution") was very high. Further, since such words would affect the results of analysing the text using topic models, these words were manually identified and removed using the function '*removeWords'* as shown in step 4 in the Fig. 1 below.

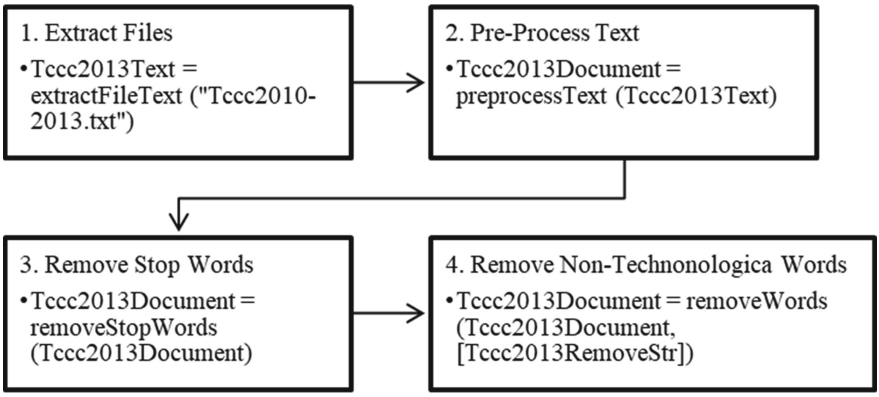


Fig. 1. Text pre-processing steps followed

2.3 Bibliometric Analysis

Bibliometric analysis has been widely used as a quantitative method for studying different aspects of research publications [26, 27]. One of these aspects is the identification of the main research areas within a scientific field. On the other hand, one of the most commonly used bibliometric analysis software is VOSviewer that is capable of creating co-occurrence or co-authorship maps from network text files such Research Information System (RIS) [40, 41]. For this paper, the bibliometric analysis was performed using RIS file containing 11,577 publications from the WoS. The maps generated by the VOS Viewer Software were used to determine the main technologies that were featured under various themes of 4IR. To minimize the number of items included in the map, the minimum number of occurrence for a keyword was set at 10; this resulted in in 961 items clustered in 11 clusters shown in Fig. 2 below. From this, the following themes are identifiable:

2.3.1 System Intelligence Design Aspects of 4IR Technologies

With 225 items, this constitutes the largest cluster and is represented by colour red (in Fig. 2). The identified technologies are algorithms related to Artificial Intelligence, Machine Learning, Deep Learning and Artificial Neural Networks. Further, the applications of these algorithms in domains such as smart manufacturing, digital twin, additive manufacturing, 3-D printing, is depicted. Besides, the application of the algorithms in the following dimensions is illustrated: classification, optimization, simulation, sustainable manufacturing, data analysis, data mining, and predictive maintenance.

2.3.2 Adoption and Innovation in Emerging Technologies

This is the second largest cluster with 210 items and depicted in colour green in the co-occurrence map. The dominant technologies include augmented reality, virtual reality, autonomous vehicles and collaborative robotics. The focus is mostly on application domains such as manufacturing, food industry, health and education. On the other hand, the dimensions highlighted are automation, safety, flexibility, productivity and digitalization.

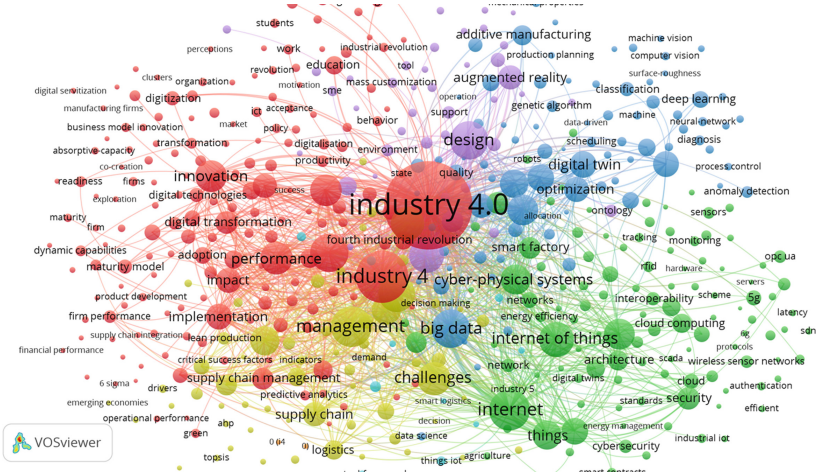


Fig. 2. Bibliometric analysis map for the 2012 to 2022 WoS publications focusing on 4IR

2.3.3 Developments Within IoT

In third position is the blue cluster that has 208 items that relate to the evolution of the core technology – Internet of Things (IoT). The highlighted baseline technologies are cloud computing, wireless sensor networks, wireless communication and radio frequency identification (RFID). Alongside these supporting technologies are the overarching issues of security, privacy and interoperability. The associated emerging technology included are Industrial IoT (IIoT), Blockchain, 5G, edge computing and fog computing. Emerging applications areas identified are smart city, smart grid and factory automation.

2.3.4 Enhanced and Futuristic Business Management

Fourth in popularity is the cluster represented in colour red, which has 159 items. The main management themes covered are smart supply chain management, lean production, digital transformation (including ditigilization), big data analytics, business models (with a focus on small and medium enterprises (SMEs)), e-commerce, performance management, integration, agility and servitization.

2.3.5 Technologies for Supporting Sustainability Agenda

At number 5, and represented by colour purple, are 90 items related to sustainability frameworks and models. The items depict 4IR as a current and future catalyst for implementing and managing sustainability. Some of the aspects include sustainable development, sustainable supply chain, sustainable environment, cleaner production, renewable energy, resilience, efficiency, risk management, reverse logistics, and emission reduction. As can be seen in Fig. 3 below, the theme of circular economy is strongly connected in the map.

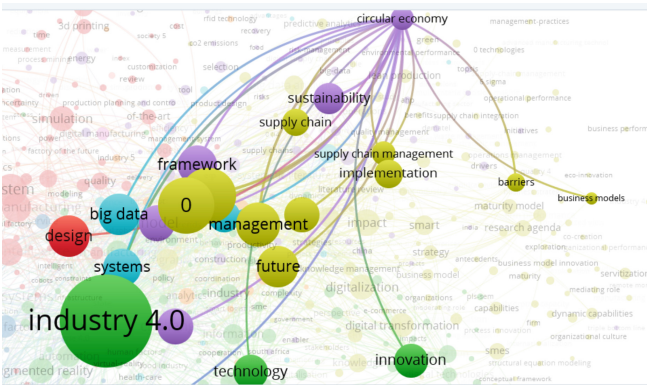


Fig. 3. Bibliometric analysis map for the 2012 to 2022 WoS publications focusing on circular economy

2.3.6 Big Data and Related Systems

The next significant cluster (light blue) with only 39 items represens Big Data, and related technologies and systems such as Data Science, Intelligence and Context and Systems. Due to many articles applying these technologies for studying the Covid-19 pandemic, it is not surprising that 'Cod-19' emerged as term in this cluster. Due to its common application in construction projects related to big data, the Building Information Model (BIM) also featured strongly under this cluster.

2.3.7 Other Minor Technologies

The much smaller Cluster 7 (Orange) had 26 items such as Blockchain Technology, Smart Logistics, Analytics, Evolution, and Opportunities. Clusters 8, 9, 10 and 11 had only one item each, which are chain, digital manufacturing and product life cycle respectively. These five clusters were not included in the subsequent text processing steps. Based on the remaining six clusters discussed above, the identified 4IR technologies and application areas were identified for further trend topic model analysis as discussed in the subsequent sections of this paper.

2.4 Analysis of the WoS Text Data Using Topic Models

Latent Dirichlet Allocation (LDA) [32] model was applied to the pre-processed text as depicted in Fig. 4 below and explained in the following paragraphs.

Step 1: Extraction CSV files from WoS: based on the Publication Year column, the WoS publications were divided into 5 files containing 2211, 2802, 2364, 2669 and 1531 publications for years 2012 to 2018, 2019, 2020, 2021 and 2022 respectively. The percentage distribution of the number publications for each of the five files is shown in Fig. 5 below. This split was meant to create some form of time-series on which to compare the developments in each of the 4IR technologies identified in Sect. 2.3 above.

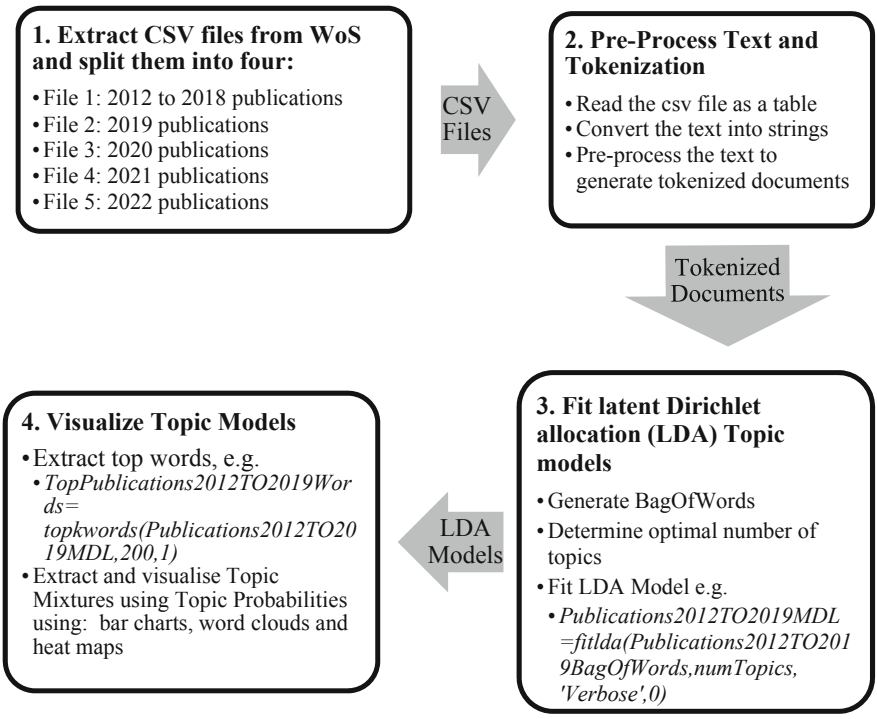


Fig. 4. LDA topic modelling steps followed

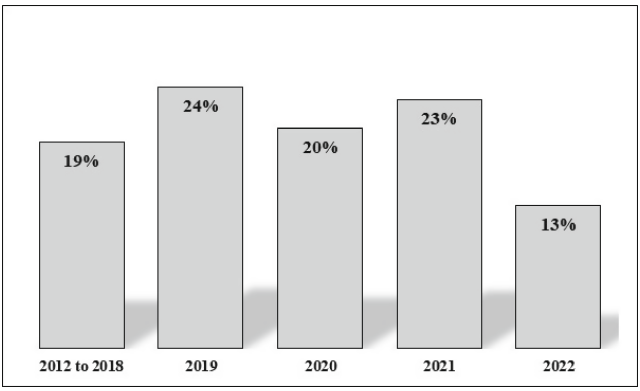


Fig. 5. Distribution of WoS publications by year of publication

Step 2: Text Pre-Processing and Tokenization: Pre-defined MATLAB functions for text pre-processing (*preprocessTex()* and *removeStopWords()*) were applied to each of the 5 files.

Step 3: Latent Dirichlet allocation (LDA) Topic models: the resulting tokenized documents (one for each publication in each of the 5 files) were then used to generate bag of individual words (*bagOfWords()*) and two-words phrases (*bagOfNgrams()*). For optimal performance, the number of topics were identified through *goodness-of-fit* of LDA models with varying number of topics. This was achieved by calculating the perplexity of a held-out set of documents for each of the four files (excluding the 2012 to 2018 one). The lowest perplexity value was selected [37] as this indicated how well the models described the set of documents in each file.

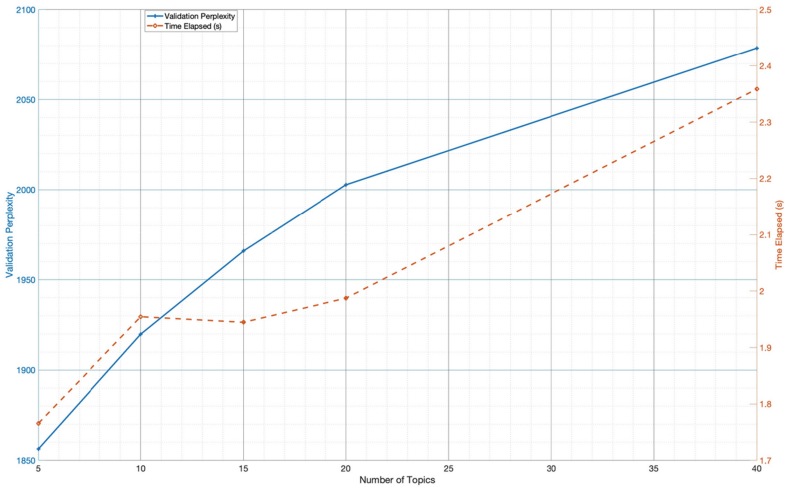


Fig. 6. Goodness-of-fit of LDA topic models for 2020 WoS publications

From Fig. 6 above for example, 10 was the optimal number of topics for the 2020 publications file. Similar number was obtained for the 2019, 2021 and 2022 files. For the 2012 to 2018 file however, different values (ranging from 15 to 25) were obtained using different LDA solvers ("cgs" "avb" "cvb0" and "savb"). The diversity of topics as well as lack of concentration of the 4IR could explain this variance. In line with literature reviewed, which revealed that 2019 was the main year when 4IR publications surged, this file (for 2012 to 2018) was omitted from the subsequent analysis.

Step 4: Visualize Topic Model: based on their probabilities, different topic mixtures were visualized using bar charts, heat maps and word clouds. Further details of this are presented under the results section of this paper.

2.5 Analyze Tccc-Announce Text Using Longer Phrases

To reflect the phrases that represent both the 4IR technologies (e.g. "Internet of Things" and "Machine Learning"), the MATLAB tool for analyzing text data using Multiword Phrases was applied to the tccc-announce files. In this case, two-words phrase was selected using the function *bagOfNgrams()*. An example of the results from this mapping are shown in Fig. 7 below.

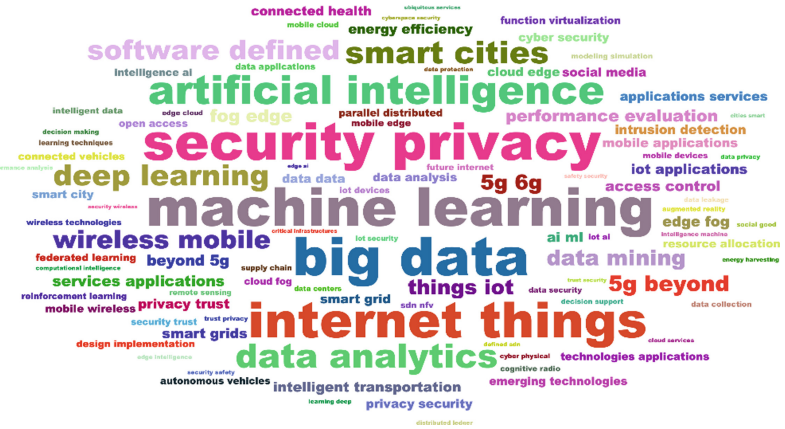


Fig. 7. Word cloud for tccc-announce emails for 2021

From Fig. 7 above, the same trends identified from the WoS dataset as subjected to the bibliometric analysis, is also identified from the tccc-announce dataset. This further affirmed the choice of the trends used in the text processing. Similar matching (with the bibliometric analysis) trends were identified for text containing tccc-announce emails sent during the years 2019, 2020 and 2022.

3 Results

3.1 LDA Topic Models for WoS

To enable contextualization of the words included in the LDA topic models for the 4 files (containing WoS publications) considered, 5 and 10 top key words for the 2-word-phrase and single-word LDA topic models respectively, were extracted using the predefined MATLAB function *topkwords()*. The results are discussed below.

3.1.1 LDA Topic Models Mixtures

In order to compare keywords/phrases from the LDA topic models mixtures with those from the bibliometric analysis, the topic probabilities were plotted using both the two-word and one-word phrases bag of words. Sample results of this are shown in Figs. 8 and 9 below. As shown in Fig. 8, two-word topics mixtures with highest probabilities are 1, 5, 2 and 6 respectively while topics 9, 3, 10 and 8 have the lowest values. On the other hand, topics 10, 6 and 8 (see Fig. 9) have the highest probabilities in the one-word-phrase topics while all the other topics have much lower values of probabilities. An inspection of the top words in these topics, as shown in Tables 1 and 2 below, confirms (see Topics 1 and 5 in Table 1 and Topics 8 and 10 in Table 2) that they are the same top words (in each cluster) as identified through the bibliometric analysis as presented in Sect. 2.3 of this paper. It is also worth noting that some of the words included under Topic 6 in Table 2 are not necessarily 4IR-related but are common words used in publications.

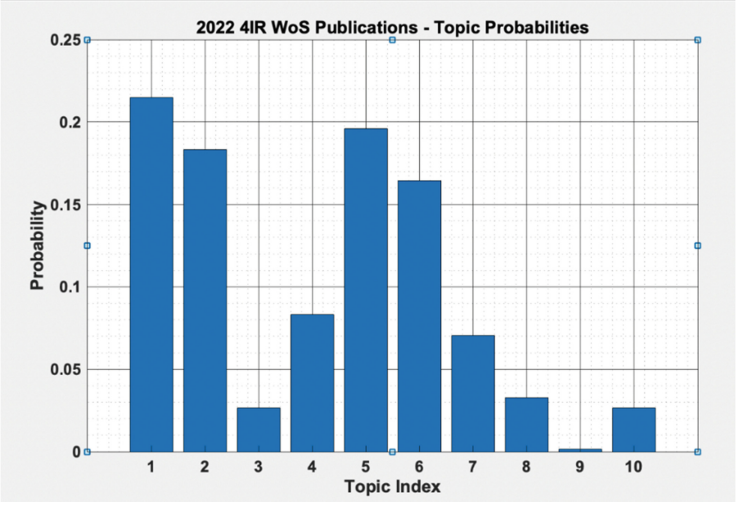


Fig. 8. LDA topic models for two-words mixtures for 2022 WoS publications

Table 1. Top words of the LDA topic models for double words for 2022 WoS publications

1	machine learning, industry 40, computer science, industry 50, sustainable development
2	computer science, smart manufacturing, science engineering, industry 40, c 2022
3	augmented reality, deep learning, quality management, chemistry engineering, operations management
4	internet things, artificial intelligence, big data, industrial internet, blockchain technology
5	industry 40, digital twin, digital twins, automation control, control systems
6	industry 40, supply chain, literature review, industry 4, 4 0
7	industry 40, 40 technologies, circular economy, 4 0, business economics
8	industry 40, smart factory, business model, waste management, international conference
9	digital transformation, industrial revolution, 3d printing, methodology approach, fourth industrial
10	industry 40, artificial intelligence, internet things, materials science, fourth industrial

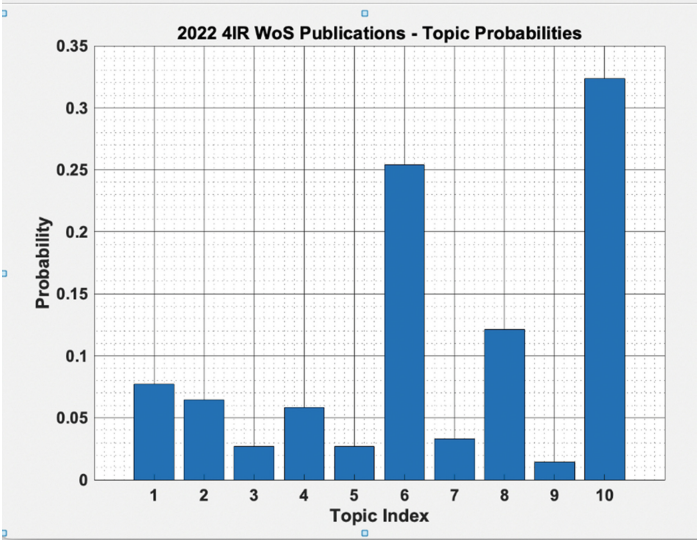


Fig. 9. LDA topic models for single words mixtures for 2022 WoS publications

3.1.2 Correlations for Topic Words Probabilities

Topic words' probabilities were created using the MATLAB function *corrcoef()* and the results presented using heat maps similar to the one shown in Fig. 10 below. From such maps, it is possible to visualize the correlations between the various subthemes/technologies under 4IR. For instance, there is a high (value of 0.2724) correlation between issues of sustainability and production in manufacturing sector. This indicates high focus on research into ways of making manufacturing more sustainable.

Table 2. Top words of the LDA topic models for single words for 2022 WoS publications

1	research, supply, management, chain, business, industry, study, technologies, literature, 40
2	industry, 40, industrial, technologies, development, revolution, smart, information, intelligence, science
3	industry, 40, digital, manufacturing, technologies, study, transformation, management, companies, implementation
4	systems, system, process, manufacturing, production, data, quality, model, control, proposed
5	environment, technology, design, development, science, application, smart, efficiency, present, product
6	new, time, based, different, results, approach, paper, problem, case, scheduling
7	science, materials, maintenance, reality, computer, c, virtual, human, high, due
8	internet, data, iot, things, network, applications, industrial, blockchain, networks, security
9	food, performance, results, covid19, technology, model, health, waste, factors, analysis
10	learning, education, energy, construction, engineering, 3d, skills, training, additive, printing

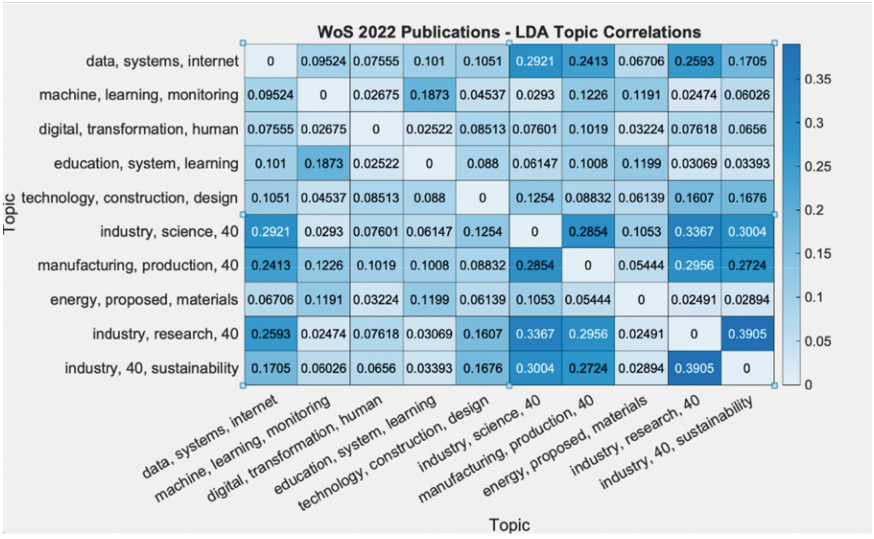


Fig. 10. LDA topic models for single words correlation for 2022 WoS publications

3.2 Fourth Industrial Revolution (4IR) Trends

Based on the top words identified through the LDA Top Models presented in sections above, the following trends were identified and matched with those identified through the bibliometric analysis.

3.2.1 System Intelligence Design Aspects of 4IR Technologies

As reflected by the words and phrases shown in Table 3 below, system intelligence design aspects of 4IR have evolved over the last 4 years and their applications in domain areas such as 3D-printing, energy, education, smart cities and construction has been amplified. Aspects of training/skilling and digital transformation have also come into limelight.

Table 3. Emerging system intelligence design aspects of 4IR technologies

Year	Two-word-phrase bag of words	Single-word-phrase bag of words
2019	machine learning, cloud computing, emerging technologies, decision making, 3d printing	data, machine, method, accuracy, detection, based, learning
2020	digital twin, internet things, business model, smart cities	system, control, sensors, applications, twin, 3d, assembly, robot, technology
	machine learning, artificial intelligence, neural network, cloud computing	learning, engineering, education, skills, technology, reality
2021	data, machine, learning, method, process, model, system, proposed, time, algorithm	
2022	digital twin, automation control, control systems	learning, education, energy, construction, engineering, 3d, skills, training, additive, printing
	digital transformation, 3d printing,	

3.2.2 Developments Within IoT

Over the last 4 years, research into developments within IoT, such as cloud computing and cyber-physical systems (CPS) have stabilized while more focus is now on trends such as industrial internet of things (IIoT) and blockchain technologies. As can be seen from the list shown in Table 4 below, issues of security and privacy still remain a concern and are still heavily researched on.

Table 4. Emerging developments within Internet of Things

Year	Two-word-phrase bag of words	Single-word-phrase bag of words
2019	big data, supply chain, cyberphysical systems	data, internet, industrial, smart, systems,
2020	cyberphysical systems, artificial intelligence, industrial internet	data, internet, things, network, industrial, communication, security, cloud
2021	artificial intelligence, machine learning	network, industrial, security, applications, computing, control
2022	artificial intelligence, big data, industrial internet, blockchain technology	data, network, applications, industrial, blockchain, networks, security
	artificial intelligence, materials science	

3.2.3 Adoption and Innovation in Emerging Technologies

As detailed in Table 5 blow, the adoption and innovations under virtual reality and augmented reality have persisted. Applications domains for 4IR innovations have been extended from smart manufacturing to include education, health (in particular, a surge in research around the Covid-19 pandemic) agriculture and the general food industry.

Table 5. Adoption and innovation in emerging technologies

Year	Two-word-phrase bag of words	Single-word-phrase bag of words
2019	engineering industry, virtual reality, digital economy, conference proceedings	design, manufacturing, process, system, production, systems, control, virtual, product, industry
	augmented reality, cyber physical, physical system, data analysis, sensor networks	
2020	computer science, science engineering, augmented reality, engineering education	
2021	education, learning, work, health, intelligence, new, covid19, human	
	smart manufacturing, augmented reality, manufacturing industry, digital transformation	environmental, science, technology, study, food, analysis, agriculture
2022	augmented reality, deep learning, quality management, chemistry engineering, operations management	science, materials, maintenance, reality, virtual, human
		food, performance, results, covid19, technology, model, health, waste, factors, analysis

3.2.4 Enhanced and Futuristic Business Management

In this sub-sector, 4IR technologies have been largely used in managing supply chains. In the last 2 to 3 years however, there is a huge focus on redefinition of business models both for pre-adoption and post-adoption of 4IR. Newest in this area are topics under circular economy and expansion of the domain to include waste management and construction industry. Some of these words and phrases are listed in Table 6 below.

Table 6. Adoption and enhanced and futuristic business management

Year	Two-word-phrase bag of words	Single-word-phrase bag of words
2019	supply, chain, technology, environmental, study, energy, performance, logistics, impact, risk	
2020	supply chain, digital transformation, internet things	management, research, manufacturing, business, supply, chain
2021	supply chain, big data, supply chains, chain management	supply, chain, technologies, management, data, technology
	deep learning, business economics, business model, construction industry	
2022	supply chain, smart factory, business model, waste management	supply, management, chain, business, industry, technologies
	smart factory, business model, waste management	circular economy, business, economics

3.2.5 Big Data and Related Systems

Similar to the bibliometric analysis, only a few words and phrases were identified through the LDA Topic Models. The list in Table 7 below show that developments in big data and data analytics are in tandem with developments in other sub-fields of 4IR. For instance, its application in industrial IoT (IIoT) and in blockchain technologies.

Table 7. Big data and related systems

Year	Two-word-phrase bag of words	Single-word-phrase bag of words
2019	model, computer, data,	
2020	big data, business economics, data analytics	data, model, machine, method, maintenance, approach, based, system
	iot, data, internet, things, network, industrial, communication, security, cloud	
2021	science, computer, systems, data, emerging, model	
2022	systems, system, process, manufacturing, production, data, quality, model, control,	
	internet, data, iot, things, network, applications, industrial, blockchain, networks, security	

3.3 Predict Top LDA Topics of Tccc Emails Documents

From the results presented in the section above, it can be confirmed that there is a strong correlation between the results of both the bibliometric analysis and LDA Topic modeling of the WoS publications on one hand and the content from the email communication within the Tccc mailing list on the other hand. Subsequently,

this section describes how the LDA Topic models of WoS were used to predict the focus of publications that may result from the tccc-announce communications. This was achieved using the *predict()* function of MATLAB, e.g. *topicIdx = predict(Publications2022MDL,Tccc2021Documents)* which predicts the Topic Model from the 2022 WoS publications that matches the tccc email content contained in the documents named "*Tccc2021Documents*".

It is a well-known trend that call for papers (conference, journal articles and workshops) published in year^{n}, will results in publications in year $^{n+1}$. For this reason, emails sent in 2021 were used to predict LDA Topics using the LDA Models generated using 2022 WoS publications. For instance, the text of the email shown in Fig. 12 below was extracted and pre-processed following the steps shown in Fig. 1. The *predict()* function was then applied on the processed documents; the results are show in Fig. 11 below. The documents were matched with LDA Topic 8 of the 2022 publications. For comparison and verification purposes, the *predict()* function was also performed using the 2019, 2020 and 2021 WoS publications. A visual inspection of the content of the word cloud shown in Fig. 11 confirms that the email text conforms more to the 2022WoS publications.

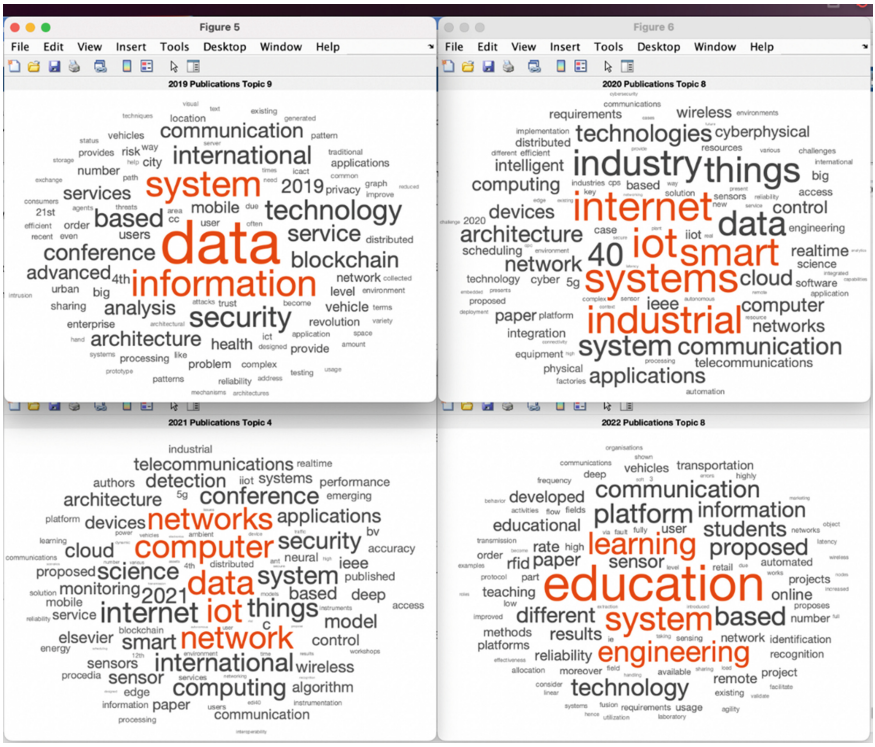
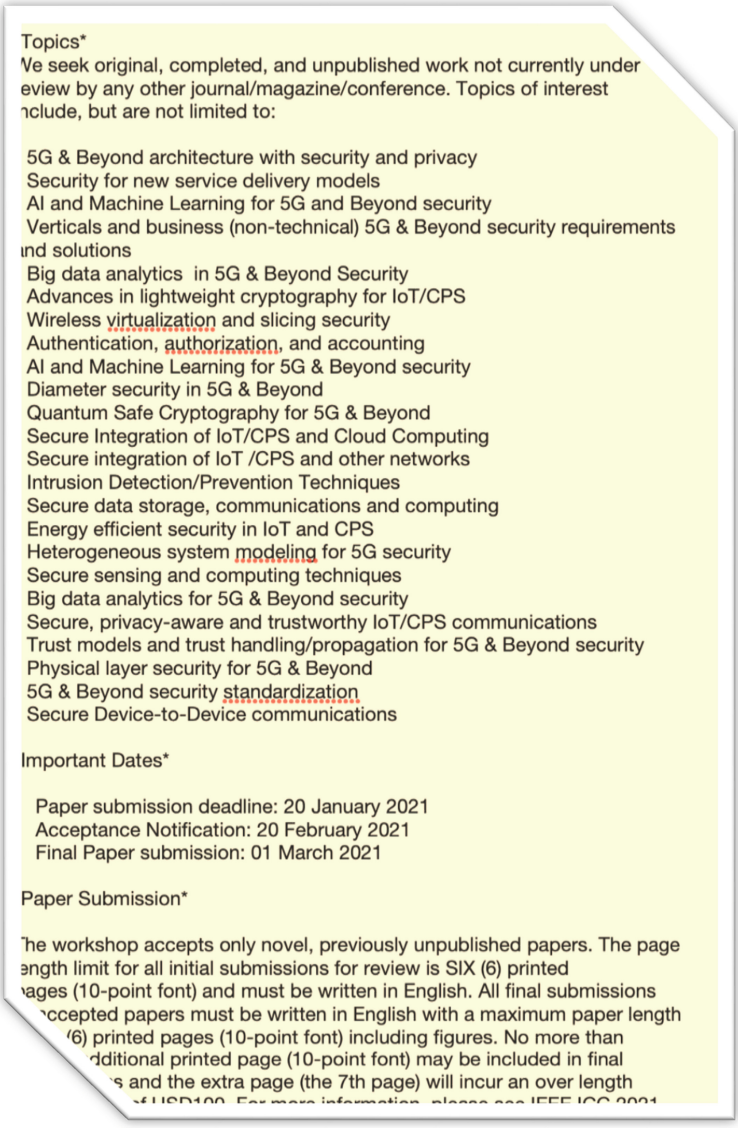


Fig. 11. Sample results of predicting Tccc content using LDA topic models for WoS publications


Topics*
Ve seek original, completed, and unpublished work not currently under
eview by any other journal/magazine/conference. Topics of interest
nclude, but are not limited to:

5G & Beyond architecture with security and privacy
Security for new service delivery models
AI and Machine Learning for 5G and Beyond security
Verticals and business (non-technical) 5G & Beyond security requirements
nd solutions
Big data analytics in 5G & Beyond Security
Advances in lightweight cryptography for IoT/CPS
Wireless virtualization and slicing security
Authentication, authorization, and accounting
AI and Machine Learning for 5G & Beyond security
Diameter security in 5G & Beyond
Quantum Safe Cryptography for 5G & Beyond
Secure Integration of IoT/CPS and Cloud Computing
Secure integration of IoT /CPS and other networks
Intrusion Detection/Prevention Techniques
Secure data storage, communications and computing
Energy efficient security in IoT and CPS
Heterogeneous system modeling for 5G security
Secure sensing and computing techniques
Big data analytics for 5G & Beyond security
Secure, privacy-aware and trustworthy IoT/CPS communications
Trust models and trust handling/propagation for 5G & Beyond security
Physical layer security for 5G & Beyond
5G & Beyond security standardization
Secure Device-to-Device communications

Important Dates*

Paper submission deadline: 20 January 2021
Acceptance Notification: 20 February 2021
Final Paper submission: 01 March 2021

Paper Submission*

he workshop accepts only novel, previously unpublished papers. The page
ength limit for all initial submissions for review is SIX (6) printed
ages (10-point font) and must be written in English. All final submissions
ccepted papers must be written in English with a maximum paper length
(6) printed pages (10-point font) including figures. No more than
dditional printed page (10-point font) may be included in final
s and the extra page (the 7th page) will incur an over length

Fig. 12. Sample extract of Tccc email sent in early 2021

4 Discussion and Conclusion

4.1 LDA Topic Models for Decision Support

Numerous studies have used published articles to analyse the trends of the technologies associated with the fourth industrial revolution (4IR). However, most of these studies take static and quantitative approaches such as bibliometric analysis [2, 42] and literature reviews [3, 7, 8, 11, 19, 21]. In this paper, we build on this approach through the

introduction of text mining using Latent Dirichlet Allocation (LDA) topic modelling. We apply these models to identify trends over time, 2019 to 2022 in this case. We further introduce aspects of dynamisms by predicting possible trends in 4IR technologies. By combining the power of bibliometric analysis with LDA topic modelling, we provide researchers, practitioners and policymakers with a variety of visual and dynamic tools that can aid in determining the current and the near-future direction (such as popularity or decline) of selected 4IR technologies.

Through the identification of topic mixtures and correlations among these topics for instance, it is possible to identify the popularity of various 4IR technologies. For example, our LDA models revealed the popularity of the concept of sustainability in relation to smart manufacturing. Also, by plotting heat maps for the 2021 Web of Science publications, it is clear that big data and related technologies had a very strong (0.6603) correlation with a topic cluster made up of: digital transformation, productive systems and additive manufacturing. For an additive manufacturing business looking into investing in big data technologies, this would be a confirmation that resources (including tools and human resources) for such an investment are available. For a company looking to identity skills set for its employees, this information could also be used in reaching informed decisions. With this, we demonstrate our solution's ability to aid in decision making.

4.2 Further Work

In this paper, very easy to follow steps on how the 4IR technologies trends identification tool was developed, are clearly presented. However, the raw-code nature of this makes it difficult for policymakers to adopt it. Further work in form of automating these steps (in form batch processing for instance) is recommended. Besides, some non-technical and very commonly used (in publications) terms ended up being assigned very high probabilities in the LDA topic models, hence introducing a lot noise into the models. For example, Topic 6 in Fig. 9, has much higher probability than Topic 8 but the earlier has non-4IR-technical words such as 'different', 'results', 'approach', 'paper', 'problem' and 'case'. We propose the adoption of more enhanced tokenization algorithms to improve the models' performance. Finally, given that it is possible to beforehand, identify the keywords to represent the 4IR trends, other variations of LDA Topic Modelling that put into consideration pre-defined keywords, could fasten the identification and training processes of the proposed tool. In this case, the implementation of Dynamic ETM (D-ETM), as described in [31], is recommended.

References

1. Lu, Y.: Industry 4.0: a survey on technologies, applications and open research issues. J. Ind. Inf. Integr. **6**, 1–10 (2017)
2. Muhuri, P.K., Shukla, A.K., Abraham, A.: Industry 4.0: a bibliometric analysis and detailed overview. Eng. Appl. Artif. Intell. **78**, 218–235 (2019)
3. Oztemel, E., Gursev, S.: Literature review of Industry 4.0 and related technologies. J. Intell. Manuf. **31**(1), 127–182 (2018)

4. Da Xu, L., Xu, E.L., Li, L.: Industry 4.0: state of the art and future trends. Int. J. Prod. Res. **56**(8), 2941–2962 (2018)
5. Tao, F.: PM10 - Digital Twin shop-floor: a new shop-floor paradigm towards smart manufacturing. Robot. Comput. Integr. Manuf. **61**, 10 (2017)
6. Kang, H.S., et al.: Smart manufacturing: past research, present findings, and future directions. Int. J. Precis. Eng. Manuf. Green Technol. **3**(1), 111–128 (2016)
7. Zhong, R.Y., Xu, X., Klotz, E., Newman, S.T.: Intelligent Manufacturing in the Context of Industry 4.0: a review. Engineering **3**(5), 616–630 (2017)
8. Monostori, L., et al.: Cyber-physical systems in manufacturing. CIRP Ann. **65**(2), 621–641 (2016)
9. Hofmann, E., Rüsch, M.: Industry 4.0 and the current status as well as future prospects on logistics. Comput. Ind. **89**, 23–34 (2017)
10. Ivanov, D., Dolgui, A., Sokolov, B.: The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. Int. J. Prod. Res. **57**(3), 829–846 (2019)
11. Ben-Daya, M., Hassini, E., Bahroun, Z.: Internet of things and supply chain management: a literature review. Int. J. Prod. Res. **57**(15–16), 4719–4742 (2019)
12. Maquera, G., Costa, B.B.F., Mendoza, Ó., Salinas, A., Haddad, A.N.: Intelligent Digital Platform for Community-Based Rural Tourism — A Novel Concept Development in Peru, pp. 1–18 (2022)
13. da Rocha, A.B.T., de Oliveira, K.B., Espuny, M., da M. Reis, J.S., Oliveira, O.J.: Business transformation through sustainability based on Industry 4.0', Heliyon, vol. 8, no. July, p. e10015 (2022)
14. Kitsantas, T.: 'Exploring blockchain technology and enterprise resource planning system: business and technical aspects, current problems, and future perspectives. Sustainability **14**(13), 7633 (2022)
15. Kumar, S., Raut, R.D., Agrawal, N., Cheikhrouhou, N., Sharma, M., Daim, T.: Technovation Integrated Blockchain and Internet of Things in the food supply chain: adoption barriers. Technovation **118**, 102589 (2022)
16. Yetis, H., Karakose, M., Baygin, N.: Blockchain-based mass customization framework using optimized production management for Industry 4.0 applications. Eng. Sci. Technol. Int. J. **36**, 101151 (2022)
17. Aivaliotis, S., et al.: An augmented reality software suite enabling seamless human robot interaction. Int. J. Comput. Integr. Manuf. **00**(00), 1–27 (2022)
18. Omerali, M., Kaya, T.: Augmented reality application selection framework using spherical fuzzy COPRAS multi criteria decision making. Cogent Eng. **9**(1), 1–38 (2022)
19. Tortorella, G.L., Prashar, A., Saurin, T.A., Fogliatto, F.S., Antony, J., Junior, G.C.: Impact of Industry 4.0 adoption on workload demands in contact centers. Hum. Factors Ergon. Manuf. **32**, 406–418 (2022)
20. Nasir, A., Zakaria, N., Yusoff, R.Z.: Cogent business & management the influence of transformational leadership on organizational sustainability in the context of Industry 4.0: mediating role of innovative performance the influence of transformational leadership on organizational sustainab. Cogent Bus. Manag. **9**(1), 0–31 (2022)
21. Grabowska, S., Saniuk, S., Gajdzik, B.: Industry 5.0: improving humanization and sustainability of Industry 4.0. Scientometrics **127**(6), 3117–3144 (2022)
22. Masinde, M., Phoobane, P., Brown, J.: Mkulima platform: an inclusive business platform ecosystem that integrates African Small-Scale Farmers into Agricultural Value Chain'. In: Sheikh, Y.H., Rai, I.A., Bakar, A.D. (eds.) e-Infrastructure and e-Services for Developing Countries. AFRICOMM 2021. Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST, vol. 443 LNICST, pp. 397–419 (2022) https://doi.org/10.1007/978-3-031-06374-9_26

23. Xie, X., Han, Y., Anderson, A., Ribeiro-Navarrete, S.: Digital platforms and SMEs' business model innovation: exploring the mediating mechanisms of capability reconfiguration. Int. J. Inf. Manage. **65**, 102513 (2022)
24. Rodrigues Dias, V.M., Jugend, D., de Camargo Fiorini, P., do A. Razzino, C., Paula Pinheiro, M.A.: Possibilities for applying the circular economy in the aerospace industry: Practices, opportunities and challenges. J. Air Transp. Manage. **102**, 102227 (2022)
25. Cheah, C.G., Chia, W.Y., Lai, S.F., Chew, K.W., Chia, S.R., Show, P.L.: Innovation designs of Industry 4.0 based solid waste management: machinery and digital circular economy. Environ. Res. **213**, 113619 (2022)
26. Pauna, V.H., Picone, F., Le Guyader, G., Buonocore, E., Franzese, P.P.: The scientific research on ecosystem services: a bibliometric analysis. Ecol. Quest. **29**(3), 53–62 (2018)
27. Phoobane, P., Masinde, M., Mabhaudhi, T.: Predicting infectious diseases: a bibliometric review on Africa. Int. J. Environ. Res. Public Health **19**(3), 1893 (2022)
28. Shuttleworth, L., Schmitz, S., Beier, G.: Impacts of Industry 4.0 on industrial employment in Germany: a comparison of industrial workers ' expectations and experiences from two surveys in 2014 and 2020. Prod. Manuf. Res. **10**(1), 583–605 (2022)
29. Tian, T., Fang, Z.: Attention-based autoencoder topic model for short texts. Procedia Comput. Sci. **151**, 1134–1139 (2019)
30. Aly, M., Khomh, F., Yacout, S.: What do practitioners discuss about IoT and Industry 4.0 related technologies? Characterization and identification of IoT and Industry 4.0 categories in stack overflow discussions. Internet of Things (Netherlands) **14**, 100364 (2021)
31. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: The Dynamic Embedded Topic Model, pp. 1–17 (2019)
32. Jelodar, H., et al.: Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools Appl. **78**(11), 15169–15211 (2018)
33. Chang, J., Blei, D.M.: Relational topic models for document networks. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, pp. 81–88 (2009)
34. Zhu, J., Ahmed, A., Xing, E.P.: MedLDA: maximum margin supervised topic models for regression and classification. In: Proceedings of the 26th International Conference on Machine Learning, pp. 1–8 (2009)
35. Zhu, J., Xing, E.P.: Maximum entropy discrimination Markov networks. J. Mach. Learn. Res. **10**, 2531–2569 (2009)
36. Lin, T., Lee, C.: Latent Dirichlet Allocation For Text And Image Topic Modeling, Na, pp. 1–6 (2013)
37. Goedecke, P.J.: Comparison of Methods for Choosing an Appropriate Number of Topics in an LDA Model, University of Memphis (2017)
38. Harandizadeh, B., Priniski, J.H. Morstatter, F.: Keyword assisted embedded topic model. In: WSDM 2022, pp. 372–380 (2021)
39. Hartung, A.F.: Computer Communications, Comput. (Long. Beach. Calif.) **6**(2), 13 (1973)
40. van Eck, N.J., Waltman, L.: Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics **111**(2), 1053–1070 (2017). https://doi.org/10.1007/s11192-017-2300-7
41. van Eck, N.J., Waltman, L.: VOSviewer Manual. Univeristeit Leiden, Leiden (2013)
42. Thai, L., Trung, D., Van Le, H., Ngoc, G.: A bibliometric analysis of Research on Education 4.0 during the 2017–2021 period (2022)