data

MDPI

Article

Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling

Kuzma Kukushkin, Yury Ryabov * and Alexey Borovkov

Computer-Aided Engineering Centre of Excellence (CompMechLab®), World-Class Research Center for Advanced Digital Technologies, Peter the Great St. Petersburg Polytechnic University, 195251 St. Petersburg, Russia

* Correspondence: ryabov_yua@spbstu.ru

Abstract: The digital twin has recently become a popular topic in research related to manufacturing, such as Industry 4.0, the industrial internet of things, and cyber-physical systems. In addition, digital twins are the focus of several research areas: construction, urban management, digital transformation of the economy, medicine, virtual reality, software testing, and others. The concept is not yet fully defined, its scope seems unlimited, and the topic is relatively new; all this can present a barrier to research. The main goal of this paper is to develop a proper methodology for visualizing the digital-twin science landscape using modern bibliometric tools, text-mining and topic-modeling, based on machine learning models—Latent Dirichlet Allocation (LDA) and BERTopic (Bidirectional Encoder Representations from Transformers). The scope of the study includes 8693 publications on the topic selected from the Scopus database, published between January 1993 and September 2022. Keyword co-occurrence analysis and topic-modeling indicate that studies on digital twins are still in the early stage of development. At the same time, the core of the topic is growing, and some topic clusters are emerging. More than 100 topics can be identified; the most popular and fastest-growing topic is 'digital twins of industrial robots, production lines and objects.' Further efforts are needed to verify the proposed methodology, which can be achieved by analyzing other research fields.

Keywords: digital twin; topic-modeling; systematic literature review; data analysis; bibliometrics; machine learning; BERTopic; LDA model

Citation: Kukushkin, K.; Ryabov, Y.; Borovkov, A. Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling. *Data* **2022**, *7*, 173. https://doi.org/10.3390/data7120173

Academic Editor: Florentino Fdez-Riverola

Received: 7 October 2022
Accepted: 22 November 2022
Published: 30 November 2022

1. Introduction

The digital twin first emerged nearly 20 years ago, as a concept to respond to the challenges of the modern manufacturing industry [1]. Michael Grieves introduced the concept initially and presented the digital twin as a model for Product Lifecycle Management (PLM). Since then, the digital twin has become an important topic in various academic and non-academic publications.

The concept of a digital twin is based on the theory that digital information about a physical system can be created as a separate entity. This virtual entity can be a 'twin' of the information embedded in the physical object or system. Furthermore, throughout the lifecycle of a physical object, there is a link between virtual and physical entities [1]. As defined by Michael Grieves and John Vickers, a digital twin 'is a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level'. There are two types of digital twins: digital twin prototype ('prototypical physical artifact') and digital twin instance ('corresponding physical product'). Digital twin prototype and digital twin instance exist in the digital twin environment ('multi-domain physics application space for operating on digital twins') [1].

The Google Books Ngram Viewer, a service to analyze the frequency of phrases (co-occurrence of two words) mentioned in English texts, based on a corpus of different texts, shows that the popularity of the digital twin has been increasing since 2014 (see Figure 1) [2].

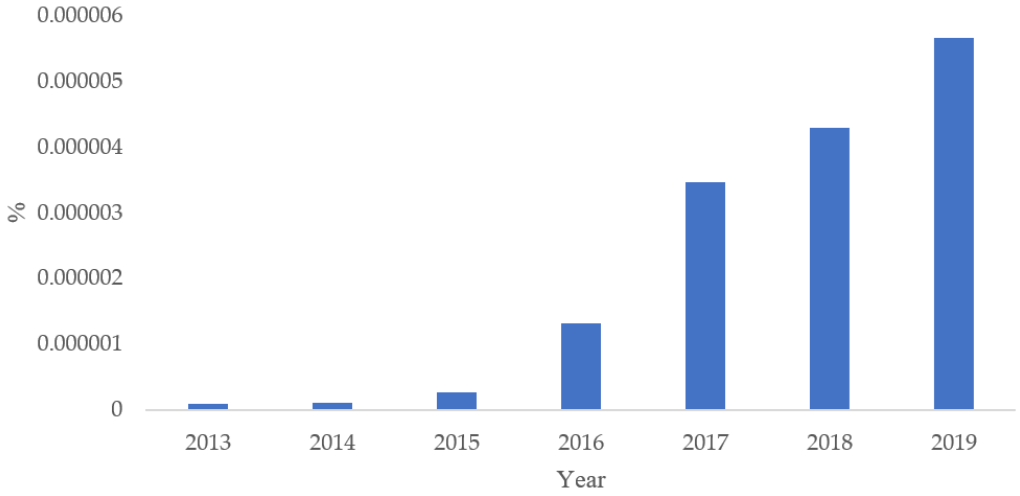


Figure 1. The term 'digital twin' occurred from 2013 to 2019, % (data taken from Google Books Ngram Viewer).

Digital twins are often considered as tools to overcome problems in a variety of fields, from manufacturing to medicine. This multimodal nature of the concept leads to a variety of approaches to this topic. In addition, the number of these approaches and new interpretations is growing with the rapid increase in thematic articles.

Standardizing definitions is a crucial task. First, it can help define what a digital twin is and what it is not. Second, it can help develop more advanced, effective, and universal methods for developing digital twins.

Another important task is to develop a clear typology of digital twins for different applications. This can help determine what a digital twin is for a given application, what approaches and methods are available for its development, and how to measure its effectiveness.

To accomplish this task, all information from different studies should be generalized. This can help develop clear definitions and typologies, and define the main applications of the concept.

These tasks are achieved today with the help of article reviews. Various researchers began to publish systematic reviews of work on digital twins.

Review authors are adding more and more publications to their research datasets. With the increasing number of publications, it becomes more and more challenging to perform reviews on this topic using traditional methods. Using bibliometric tools is one of the methods to overcome this problem. At the same time, there are still many problems with bibliometric studies on this topic. The results of clustering and created co-occurrence networks are often challenging to interpret.

One possible solution is to combine different methods to perform topic modeling on large datasets of research papers. In this paper, an attempt is made to propose such a combination of methods.

Our research objectives can be summarized in a short list of questions:

1. Is the digital twin a mature topic with established 'schools of thought' and approaches, or is it still a new topic and a kind of marketing label with a vague background?
2. The analysis has revealed some ambiguity in the co-occurrence of keywords, i.e., this co-occurrence seems random in some cases. Are there any clear mainstream themes on this topic?
3. Some researchers think machine learning is the future trend for digital-twin studies. What are the future trends in the development of the topic?

The main goal of this paper is to propose a combination of methods for bibliometric analysis that can be used to study large datasets of articles.

For this purpose, the paper is organized as follows. Section 2 examines 'traditional' literature reviews and bibliometric studies on the topic of digital twins. Section 3 describes the dataset, introduces data analysis (text mining) and machine-learning methods, and outlines the three stages of our research. Sections 4–6 report the results of each stage of the study. Section 7 concludes the paper with answers to the three research questions formulated in Section 1. Finally, Section 8 provides avenues for future research to verify the proposed methodology.

2. Related Literature

2.1. Digital Twin Literature Reviews

The growing interest in the topic led to attempts at classification and the definition of the typology of digital twins.

One of the most cited papers is 'Digital twin in manufacturing: a categorical literature review and classification' by Kritzinger et al., in 2018. This paper has been cited 736 times [3]. The authors developed a typology for digital-twin studies. According to the paper, more than half of the studies focused on the concept. In addition, the authors distinguished between digital twins in different papers. Most publications consider the concept of a digital model (no automated dataflow between physical and virtual counterparts) and a digital shadow (automated one-way dataflow from physical to virtual counterparts). Very few papers address the 'true' digital twin, which, according to the authors, has an automated dataflow in both directions [3]. The second most-cited review paper is 'A Review of the Roles of Digital Twin in CPS-based Production Systems" by Negri, Fumagalli, and Macchi, published in 2017, and cited 634 times [4]. In this paper, a literature review based on Scopus data is proposed. In 2017, 16 articles attempted to define the digital twin. In total, 26 articles were published on this topic between 2012 and 2017.

The authors claim that the scientific literature on this topic is ' . . . still in its infancy'. They categorize the publications into three topic groups:

1. Health Analysis and maintenance activities (deformation, anomalies, fatigue, etc.).
2. Digital mirrors of the life of the physical entity.
3. Decision support through engineering and statistical analysis [4].

2.2. Digital Twin Literature Meta-Reviews

Several modern reviews often provide new approaches to analysis.

The paper 'Digital Twins: A Meta-Review on Their Conceptualization, Application, and Reference Architecture' by Rossman and Hertweck (2022) provides an overview of 14 systematic literature-reviews on digital twins published between 2018 and 2021 [5]. Based on the analysis, the authors develop an architecture of the digital twin that includes nine layers:

1. Physical entities, physical twin.
2. Data generation.
3. Network, connectivity.
4. Data storage and integration.
5. Data preparation and representation.
6. Data model, algorithms, a virtual entity, virtual twin.
7. Micro-services, deployment.
8. System security, data privacy.
9. Business model, processes [5].

Another meta-review was published by Kuehner, Scheer, and Strassburger in 2021, entitled 'Digital Twin: Finding Common Ground—A Meta-Review' [6]. In this study, 24 papers on the digital twin were examined. The authors created a detailed classification of the reviews considered. It was found that most authors (83%) agree that the concept of the digital twin is at an early stage of development. The paper focuses on the benefits and challenges identified in the reviews. Condition monitoring and tracking, system prediction,

system analysis, system prescription, and data management are among the key benefits of using digital twins.

Key issues considered in the reviews include:

1. Data infrastructure.
2. Modeling and simulation.
3. Implementation.
4. Privacy, security, and legal issues.
5. Concept standardization.
6. Clarification of benefits.
7. Digital twin and human interaction [6].

In summary, the appearance of meta-reviews is evidence of the increasing number of publications and, at the same time, the increasing complexity of the generalizations presented in these reviews. The analysis of individual publications is becoming more complicated, so other methods are constantly being sought to increase the dataset size and facilitate the analysis of datasets containing hundreds or even thousands of publications.

2.3. Bibliometric Tools in Digital Twin Literature Research

Bibliometric tools offer a simple way to solve these problems. Early bibliometric studies typically focused on citation and co-citation analysis. These methods are used to assess the impact of publications and authors. As noted by Mejia et al., trend analysis of publication topics has been one of the directions in bibliometrics since 2013 [7]. With bibliometric trend-analysis and new machine-learning methods, significant publications datasets can be examined, and fairly accurate conclusions can be drawn. Very few reviews of digital twins have relied on bibliometric tools thus far.

One of the first attempts to apply bibliometric tools to digital-twin analysis was made in the monograph 'Digital Twins in the High-Technology Manufacturing Industry' by Borovkov, Gamzikova, Kukushkin, and Ryabov, published in 2019. The monograph examines the digital twins' landscape from a bibliometric perspective. The study also includes an analysis of co-authorship, co-citation networks, and co-occurrence of keywords [8].

The authors created keyword networks and defined the main publication clusters:

1. Industry 4.0, smart factory, big data, industrial internet of things, artificial intelligence.
2. Cyber-physical systems, machine learning, simulation, virtual factory.
3. Digital thread, virtualization, Product Lifecycle Management, modeling.
4. Internet of things, AR/VR, digital shadow [8].

The analysis results suggest that a simple clustering of keywords can help identify the most popular words; on the other hand, the results seem somewhat ambiguous. It is difficult to determine whether the co-occurrence of keywords in clusters is random or intentional. Some authors may be trying to avoid unordered topic clustering by making more precise queries to the citation systems. In 2021, Warke et al. published a paper entitled 'Sustainable Development of Smart Manufacturing Driven by the Digital Twin Framework: A Statistical Analysis' [9].

The authors present an interpretation of the evolution of the digital twin. They define four evolutionary stages:

1. Information monitoring model (1985–2002).
2. Digital simulation (2003–2014).
3. Implementation of IoT devices (2014–2016).
4. Use of decision-making tools (2017–present).

The authors consider seven research papers with a bibliometric analysis of digital-twin applications in smart manufacturing. It is suggested that the main challenge for the 'traditional' approach is to examine all types of papers found on this topic. Finally, a specific approach is proposed, to provide a clearer focus on smart manufacturing and Industry 4.0. The Proknow-C method chosen by the researchers is based on the use of 'Master Keywords', 'Primary Keywords' and 'Secondary Keywords' when searching for articles on a topic.

The further study is based on 509 articles found through a query using the Proknow-C method. The authors considered the co-occurrence of keywords and formed co-authorship and co-citation networks.

The authors formulated the following conclusions:

1. Existing literature discusses using digital twins for the entire process or plant.
2. There is no literature on multi-domain models describing numerical and mathematical modeling for system monitoring and optimization.
3. While some studies apply machine-learning algorithms, they are often not validated by simulations and mathematical models [9].

The literature analysis above shows a transition of research methods from definition typology to complicated meta-reviews with different levels of search queries. At the same time, there is still a lack of publications that deal with topic modeling and examine a larger corpus of text rather than small samples of articles. This conclusion is very important for the objectives of this study.

3. Materials and Methods

3.1. Materials

The information on research papers for the study was obtained from the Scopus database [10]. For the purposes of the study, we used the 'traditional' approach [9] and collected the entire corpus of text on the topic of digital twins.

Papers were selected using the Scopus database query for the approximate search 'Digital Twin' in the publication's Title, Abstract, and Keywords fields.

The quotation marks "" were used to find articles in which the phrase 'Digital Twin' appears in a fixed order.

The selection was made in September 2022. The dataset contains 8693 articles matching the specified search query, and published between 1993 and 2022 (the first eight months).

The search dataset was exported from Scopus to a *.csv file in six steps, since only 2000 complete records can be downloaded at once.

Publications on digital twins have increased dramatically over the past six years (see Figure 2). In total, 8693 articles had been published on this topic by September 2022. While only 29 articles were published in 2016, the number had increased to 2997 by 2021. The Digital Twin dataset contains 8693 rows of complete citation information for the articles. One row in the table represents information about one article. There are also 54 columns. Each of these columns represents a piece of citation information from an article.

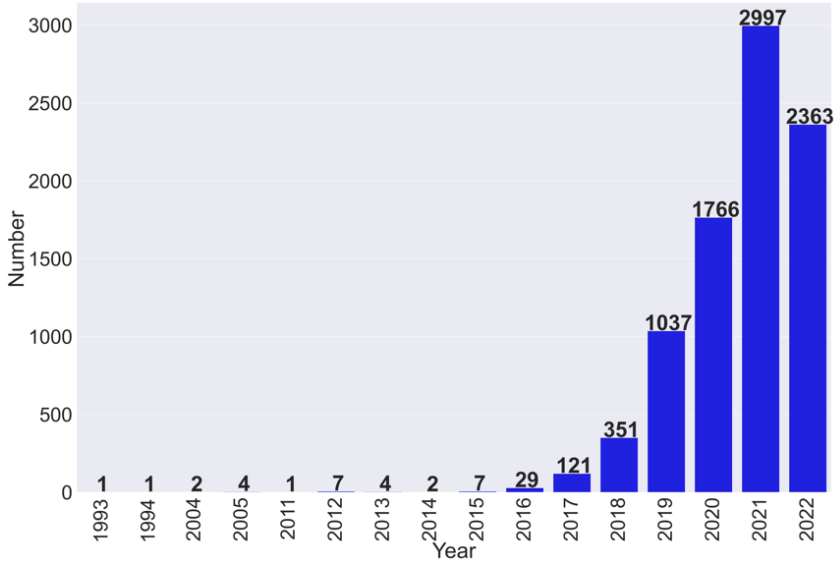


Figure 2. The number of publications on digital twins (data taken from Scopus, records from January 2016 to March 2022).

The most important columns for this study are shown in Table 1 below:

Table 1. The most important columns of the Digital Twin dataset.

№ of Column	Column	Content
0	Authors	Names of the authors
2	Title	Title of the publication
3	Year	Year of publication
12	Cited by	Number of citations
17	Abstract	Abstract of the article
18	Author Keywords	Keywords used by authors of an article

3.2. Methods

3.2.1. Text Mining

This study uses data analysis (text mining) and machine-learning methods to explore a corpus of articles on digital twins to achieve the research objectives.

Text-mining technology emerged in the 1990s, when it became possible to upload and process large, unstructured text-datasets. This type of analysis makes it possible to extract information from a small corpus of text that is unknown or difficult to detect [11–13].

In general, the text-mining process can be divided into several phases:

1. Information is collected from unstructured data.
2. Information is transformed into structured data (text is cleaned, words are tokenized, and stop words are filtered out).
3. Patterns/themes are identified in the structured text.
4. The pattern/topic is analyzed.
5. Valuable information is extracted (e.g., visualized or stored in the database) [11].

Text mining can be used to analyze any document set. In recent years, this approach has been increasingly used in bibliometric studies involving the mathematical and statistical analysis of monographs, research papers, and other types of publications.

The text mining and data processing in this study were performed in the Jupyter Notebook client, which provides tools for data analysis using the Python programming language.

In the Jupyter Notebook environment, libraries such as Counter (counting items on a list), Pandas (analysis of data frames and tables), Matplotlib, Seaborn, and Plotly (data visualization) were used.

We also used the following libraries for text markup and tokenization: Natural Language Toolkit (NLTK), a package of libraries and programs for symbolic and statistical natural-language processing.

3.2.2. Topic Analysis: Machine Learning

Machine learning was used to identify the main topics of the articles.

The topics were analyzed using Latent Dirichlet Allocation (LDA), a popular topic-modeling algorithm implemented notably in the Gensim library [14].

The LDA model was introduced in 2003 [15], and is one of the most widely used methods for publication topic-modeling.

This model is based on a generative approach, where classes do not need to be pre-labeled. Instead, the algorithm generates a probabilistic model that is used to define topic clusters. The model can be used to classify both existing and new documents [16].

The BERT library and BERTopic were also used to identify key topics. The BERT library was developed by Google in 2018 for natural-language processing, including the classification, generation, and summarization of text. The BERTopic model allows for the obtaining of vector representations of the text or embeddings [17,18].

While the LDA model is based on statistical calculations, the BERT model considers the context of the words, i.e., it assumes that the words appearing in similar contexts are likely to have similar meanings [19].

There is a debate in the literature about which models are most effective for text analysis. For example, a key drawback of the LDA model is the possible overlap of topics. The problem with BERTopic is that each document can only be assigned to a single topic [20].

From this point of view, BERTopic was considered more suitable for the purposes of this study. However, we also created an LDA model for a more comprehensive analysis of articles on digital twins.

3.2.3. Stages of the Study

The study consisted of three stages involving stepwise collection and data processing.

1. Dataset was uploaded to the Jupyter Notebook; the number of publications per year was calculated, the co-authors were analyzed, and the most-cited authors and the most-cited publications were identified.
2. Analysis of the keywords of the publications:
    - The author's keywords were normalized. This step helps avoid double counting of a keyword in different spellings. For example, 'digital twin' can be referred to as 'dt', 'digital twins', 'digital-twin', etc. Normalization means that all spellings of a keyword are renamed to the spelling that will be used for the analysis.
    - The occurrences of the author's keywords were counted.
    - The co-occurrences of the author's keywords were counted.
    - A graph of the co-occurrence of keywords was constructed.
3. Analysis of abstracts of publications:
    - Stop words and punctuation marks were removed from the text.
    - The abstract text was tokenized/decomposed into bigrams.
    - The occurrences of individual words in the abstracts were counted.
    - The most frequently occurring bigrams were identified.
    - The GenSim model was trained for LDA analysis, and the key topics were identified in the abstracts.
    - The BERTopic model was trained, and the obtained results were visualized.

Indeed, the proposed steps include the procedure for analyzing large datasets of research papers. Moreover, the machine-learning tools are intended to be applied to open-access abstracts containing short summaries of articles.

4. Stage I

At the first stage, the information was uploaded to Jupyter Notebook for subsequent data analysis in the Python environment. The Pandas library (tabular processing-data) and the Matplotlib and Seaborn libraries (visualization) were used for uploading.

The list of downloaded datasets was compiled in a table. The articles were sorted by the number of citations, in descending order (see Table 2).

Table 2. Most-cited publications on digital twins (data taken from Scopus, records from January 1993 to September 2022).

№	Article Name	Number of Citations
1	Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, Fangyuan Sui 'Digital twin-driven product design, manufacturing and service with big data' [21]	1147
2	Dmitry Ivanov 'Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case' [22]	772

Table 2. *Cont.*

№	Article Name	Number of Citations
3	Werner Kritzinger, Matthias Karner, Georg Traar, Jan Henjes, Wilfried Sihn 'Digital twin in manufacturing: a categorical literature review and classification' [3]	736
4	Fei Tao, He Zhang, Ang Liu, Andrew Yeh-Ching Nee 'Digital Twin in Industry: State-of-the-Art' [23]	728
5	Michael Grieves, John Vickers 'Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems' [1]	708

The top five most popular articles on the topic of digital twins were cited 4091 times.

The most-cited papers fall into three categories: reviews that present a typology of digital twins [3,23], conceptual papers that provide an understanding of the topic [1,21], and a case study of the successful practical application of the concept [22].

Next, Vosviewer (a data-visualization platform) [24] was used to create a co-authorship graph containing co-authors who collaborated on 15 or more articles (see Figure 3). The result contains a graph with 137 authors and their co-authorship connections.

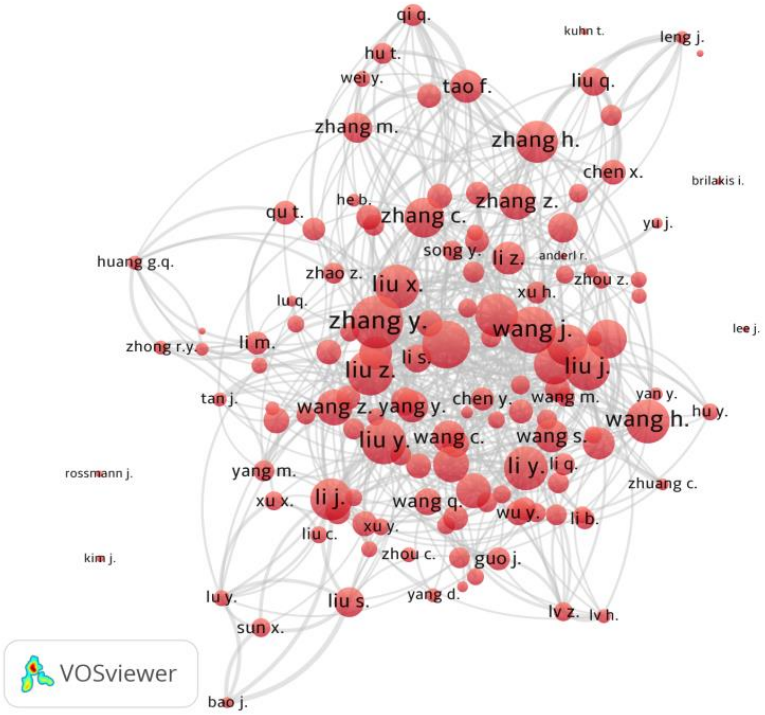


Figure 3. Co-authorship graph on digital twins (created with Vosviewer).

The biggest digital twin 'school of thought' is represented by Fei Tao, professor at the School of Automation Science and Electrical Engineering and vice-dean of the Institute of Science and Technology of Beihang University, and his colleagues.

Researchers in this topic cluster, which is shown in red in Figure 3, focus on the theoretical approaches to integrating digital twins into the manufacturing process at three levels: an industrial unit (equipment), a system, and a system of systems. Beihang University's group of researchers has also been working on the theory of a five-dimensional model for digital twins, since 2018.

According to Tao and his colleagues, the established three-component definition of a digital twin, which includes a real object, a virtual model, and the relationship between them, should be supplemented with two more dimensions: data and services. In addition to the direct interaction between the 'physical' and the 'virtual' object, researchers have

also considered the potential free exchange of digital-twin data between the virtual and real environments. Such functions of the digital twin as monitoring, evaluation and prediction, can be generated as a set of 'services' to help decision-makers in overall production management [25].

The analysis shows that the red cluster also represent different 'schools of thought'. For example, in the article by L. Yangguang (Dalian University of Technology) et al., 'An IoT-enabled simulation approach for process planning and analysis: a case from engine re-manufacturing industry', the authors consider a new simulation approach based on a case study of an engine re-manufacturing plant [26]. Another article by J. Liu (Jiangsu University of Science and Technology) et al., 'Dynamic Evaluation Method of Machining Process Planning Based on Digital Twin' proposes a new method for evaluating manufacturing processes based on digital-twin technology [27]. Y. Wang (University of New South Wales) et al,. in their article 'Digital Twin-Driven Analysis of Design Constraints', investigate the role of digital twins in the analysis of design constraints, using the example of a robot vacuum-cleaner [28].

Next, Vosviewer (a data-visualization platform) [24] was used to create a co-citation graph. Figure 4 contains the citation network of authors who were co-cited more than 200 times.

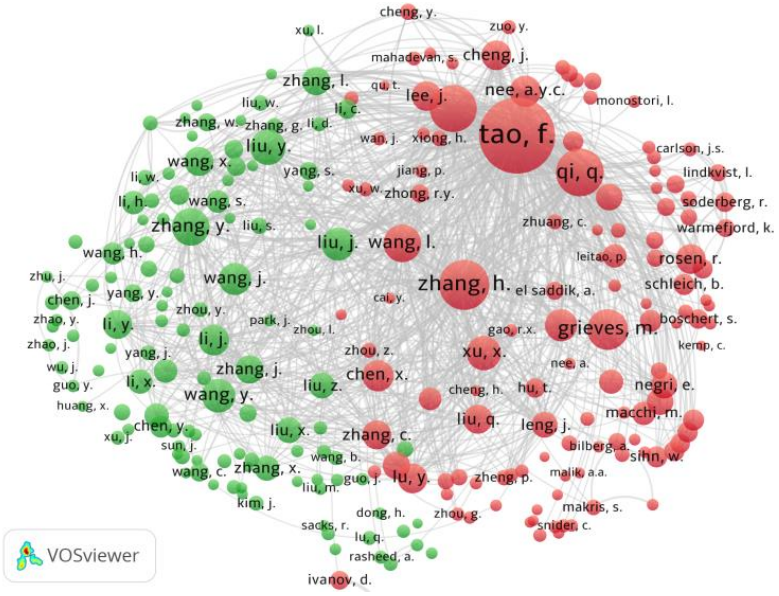


Figure 4. Co-Citation graph for digital twins (created with Vosviewer).

The clustering of the graph is based on the principle of co-citation. This means that the closer the dots are to each other, the more often the authors are mentioned together in the bibliography of scientific articles.

The graph is divided into two clusters: red dots (125 authors) and green dots (119 authors).

The red cluster co-citation leaders are F. Tao, H. Zhang and M. Grieves. This cluster clearly represents the most-cited and popular publications and authors on the concept of the digital twin [1,3,4,25,29–32]. The red cluster also includes the key co-authors of Fei Tao. The co-citation leaders of the green cluster are Y. Zhang, Y. Liu and J. Wang and others [26–28]. These researchers represent 'schools of thought' in mostly Chinese universities that focus on digital-twin studies.

To sum it up, co-authorship and co-citation analysis can identify key researchers and their groups. Key research-teams may focus on different topics, so the clusters of authors do not match the clusters of topics.

5. Stage II

To analyze trends in digital-twin publications, we calculated the number of mentions of keywords by publication authors, and assessed the frequency of co-occurrence of keywords (see Figure 5).

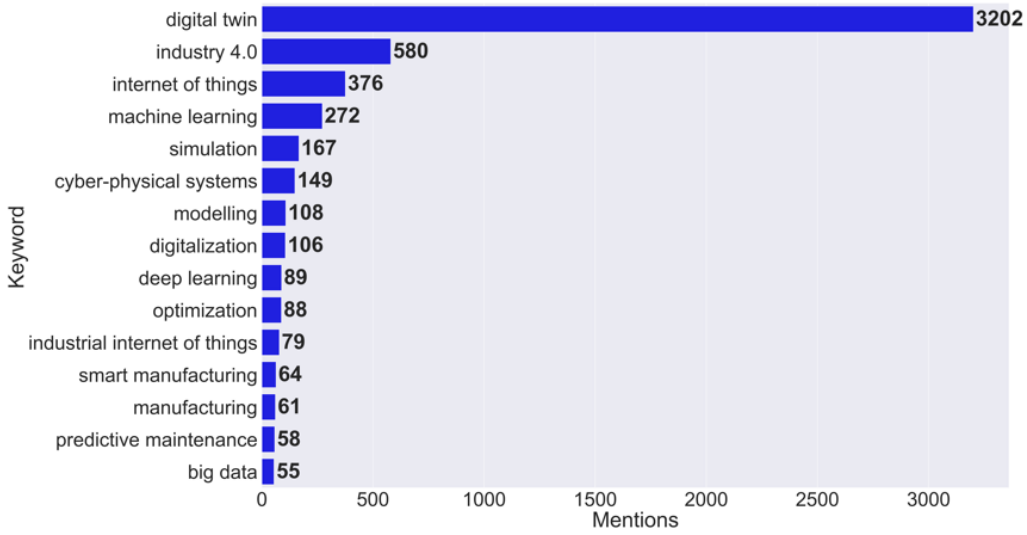


Figure 5. The number of mentions of keywords by publication authors.

A thesaurus was used to calculate the number of keyword mentions and analyze their co-occurrence (for example, the keywords 'dt', 'dts', 'digital twins', and 'digital-twin' were replaced by the keyword 'digital twin'). In addition, all words were converted to lowercase, to avoid double counting.

The analysis shows that the most frequently mentioned terms in the articles on digital twins are 'Industry 4.0' (580 occurrences), 'Internet of things' (376 occurrences), 'machine learning' (272 occurrences), 'simulation' (167 occurrences), and 'cyber-physical systems' (149 occurrences).

It should be noted that the keywords in the considered articles are very diverse. Even the most popular keywords, such as 'Internet of things', are mentioned only 376 times in 8693 publications, representing only 4.3% of all publications.

We also created a graph of the co-occurrence of keywords, limiting ourselves to the first 113 keywords so that the dataset contains the keywords mentioned 20 times or more (see Figure 6).

Three clusters can be identified: the red cluster (63 keywords), the green cluster (27 keywords), and the blue cluster (21 keywords). The red cluster contains keywords related to Industry 4.0, including cyber-physical systems, smart and digital factories, and model-based systems engineering. The green cluster contains terms related to the Internet of things, including sensors, big data, and cybersecurity. The blue cluster contains keywords: BIM (building information modeling), virtual reality, augmented reality, mixed reality, etc.

The keyword analysis shows the cross-cutting nature of digital-twin technology and the proximity of the digital twin to the stack of manufacturing technologies (Industry 4.0, industrial internet of things, cyber-physical systems).

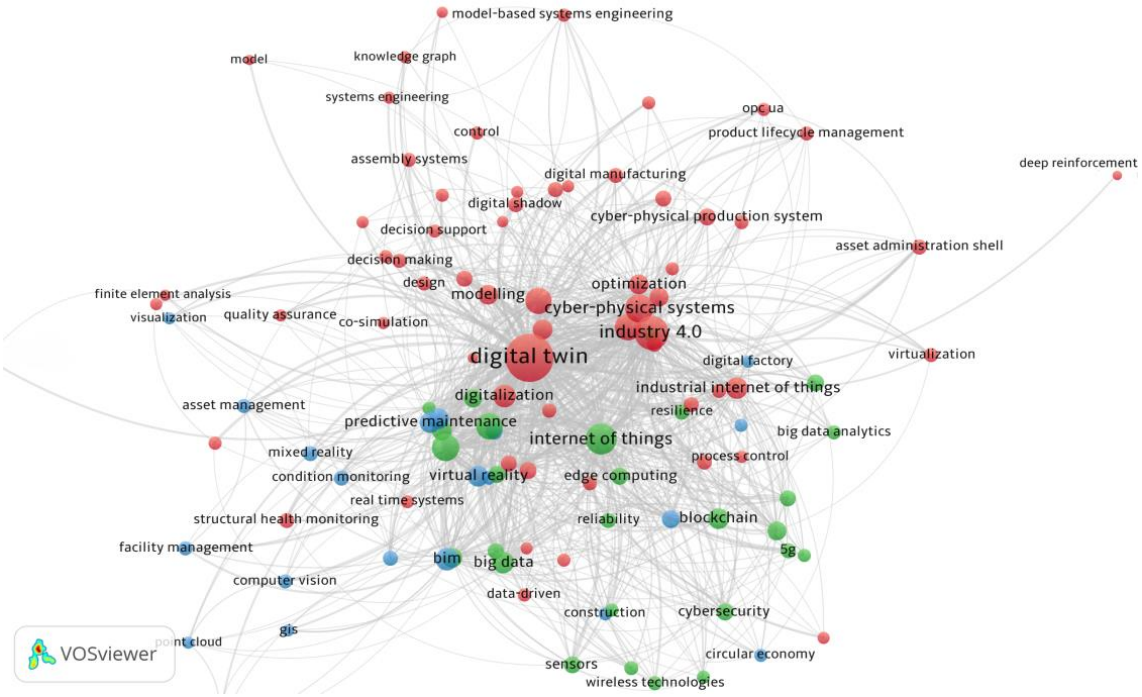


Figure 6. Co-occurrence of keywords (limited to 113 keywords) (created with Vosviewer).

6. Stage III

6.1. Comparison of LDA and BERTopic Models

There are some key differences between the two models that have a major impact on the research results. The LDA model represents the text as a bag-of-words that does not take into account the context of the words. Therefore, the results of the LDA analysis are sometimes unable to represent the topics in the texts. This situation has led to the increasing popularity of text embedding. These techniques produce contextual word and sentence representations. As a result, similar texts can be close to each other in the vector space [19].

The LDA model has several advantages. For example, it generates mixed topics that sometimes better fit the content of the texts. In addition, the number of topics is smaller than in topic-embedding models, which is important for easier text interpretation. The disadvantage is the neglect of word correlation, and the results may generate overlapping topics [20].

The advantages of BERTopic include good performance, flexibility and the ability to present topics as a distribution of words. The first disadvantage of the BERTopic model is the strict correspondence between a topic and a document. In addition, the model only shows the meaning of the words in a topic, so it is often difficult to interpret the results and describe a topic correctly [19].

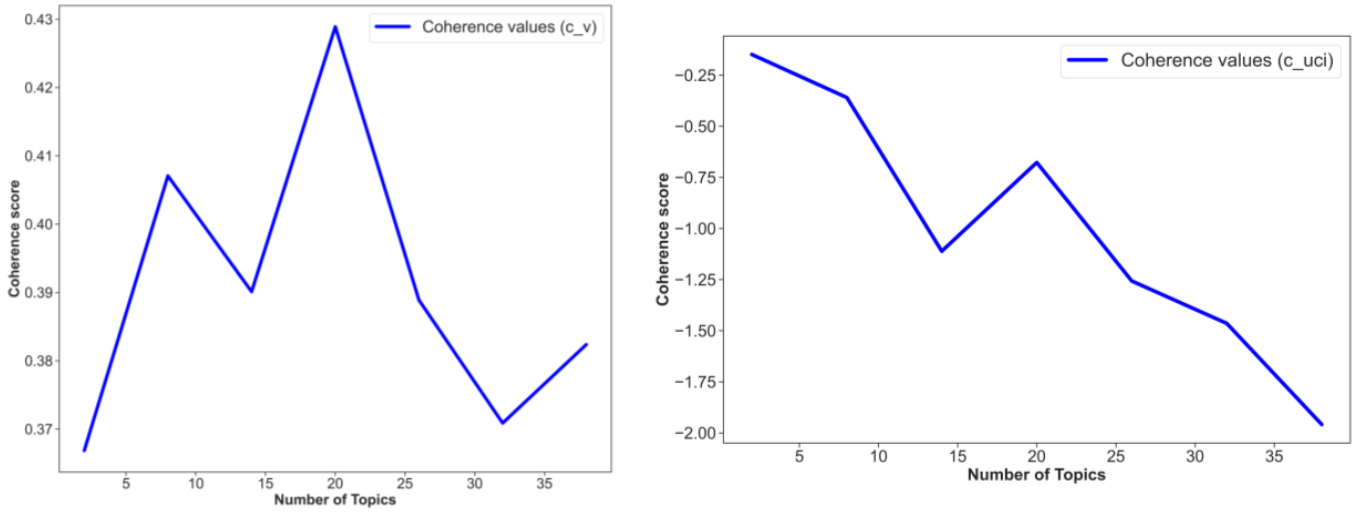
6.2. The LDA Model

To prepare the texts of the abstracts for analysis, the data (8693 rows) were first cleaned of all rows that did not contain texts of the abstracts. As a result, 8487 rows (abstracts) were selected. Then, the texts of the abstracts were cleaned of stop words (e.g., punctuation marks, prepositions, articles) and collected in a list.

To avoid false or uninterpretable results, the raw text should be prepared for analysis. This means that only the words that are important for topic modeling should remain in the texts of the abstracts. Therefore, the abstracts were first cleaned of stop words, e.g., punctuation marks, prepositions, and articles. Stop words that frequently appear in the articles also include the type of publication (e.g., review, article, etc.), the name of the publisher, the year of publication, and the name of the organization. Finally, all the texts of the abstracts were collected in a long list.

During the training of the LDA model, the researcher should manually enter the number of topics as a model parameter. Coherence measures are commonly used to determine the optimal number of topics for analysis. These measures evaluate the degree of semantic similarity of the most popular words within a topic [33]. In this study, two indicators are used to determine the optimal number of topics for analysis: 'c_v' and 'c_uci'. The 'c_uci' indicator is based on the pointwise mutual information (PMI), while the 'c_v' indicator is based on the normalized pointwise mutual information (NPMI). PMI helps to evaluate the probability of co-occurrence of two words, taking into account the fact that this co-occurrence can be caused by the frequency of the words [34].

The results of the analysis show peaks of coherence for 20 topics (see Figure 7). This means that 20 topics are probably optimal for the analysis.



(**a**) (**b**)

Figure 7. LDA coherence score: (**a**) 'c_v' and (**b**) 'c_uci' indicators.

Another method, called PyLDAvis, is used to check the optimal number of topics for analysis. This method is also used to reveal overlapping topics (see Figure 8).

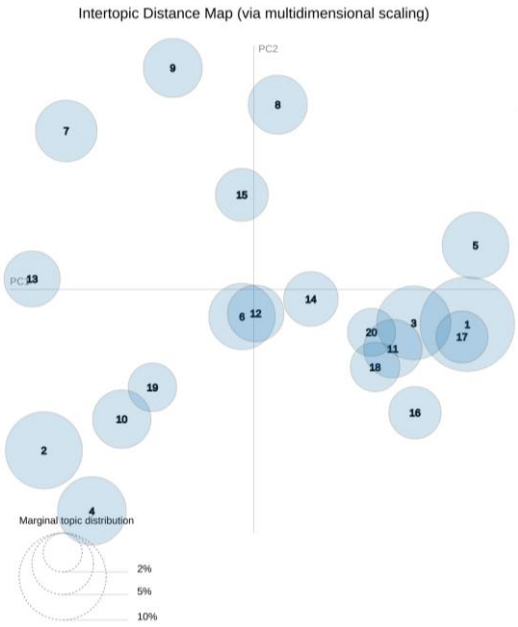


Figure 8. Intertopic Distance Map (PyLDAvis).

The results presented in Figure 8 show that only a few topics have overlapping words when 20 topics are selected for the LDA model. Thus, 20 seems to be the right number of topics for modeling.

An LDA model was trained to evaluate the article topics on digital twins. The model was trained for the task of generating 20 key topics.

First, the articles were assigned to the topic that was most heavily weighted in the article abstracts. We estimated the number of articles for each topic by assigning the article to the topic that had the greatest weight in that article. For each topic, we selected the most popular words to describe the topic. The 20 topics are listed in descending order in Table 3.

Table 3. Number of articles by dominant topic (data taken from Scopus, records from January 1993 to September 2022).

Topic Name and Number	Number of Articles by Dominant Topic
Topic 10. Transformation, intelligence, companies, solutions, technological	1354
Topic 2. Human, construction, BIM, safety, robot	690
Topic 15. Thermal, temperature, power, flow, experimental	689
Topic 6. Test, planning, methodology, solution, flexible	531
Topic 14. IoT, cloud, CPS, knowledge, services	488
Topic 9. Material, properties, measurements, materials, element	468
Topic 17. Equipment, assembly, fault, line, workshop	463
Topic 8. Power, construction, vehicle, safety, traffic	368
Topic 13. Networks, edge, communication, computing, security	368
Topic 20. Mining, heritage, infrastructure, railway, equipment	363
Topic 18. Security, business, drilling, literature, DTs	343
Topic 11. Structural, damage, health, sensor, bridge	324
Topic 5. Machining, engine, prediction, tool, crack	308
Topic 1. Energy, city, urban, scheduling, battery	304
Topic 7. Robot, reality, robots, robotic, space	300
Topic 4. Construction, students, education, modeling, knowledge	252
Topic 16. Supply, chain, patient, health, medical	246
Topic 12. Value, service, logistics, lifecycle, context	231
Topic 3. Reality, metaverse, VR, augmented, commissioning	198
Topic 19. Structure, structural, point, construction, wind	198

Some topics can be considered mixed: Topic 2. Human, construction, BIM, safety, robot (690 articles), Topic 8. Power, construction, vehicle, safety, traffic (368 articles), Topic 20. Mining, heritage, infrastructure, railway, equipment (363 articles), Topic 18. Security, business, drilling, literature, DTs (343 articles), Topic 4. Construction, students, education, modeling, knowledge (252 articles), Topic 16. Supply, chain, patient, health, medical' (246 articles), Topic 19. Structure, structural, point, construction, wind (198 articles).

At the same time, there are 13 topics that can be interpreted unambiguously:

1. Topic 10. Business digital transformation (1354 articles).
2. Topic 15. Physical processes (689 articles).
3. Topic 6. Tests and software testing (531 articles).
4. Topic 14. Internet of things and cloud services (488 articles).
5. Topic 9. New materials and modeling (468 articles).
6. Topic 17. Manufacturing processes (463 articles).
7. Topic 13. Networks, communication and computing (368 articles).
8. Topic 11. Structural-damage analysis and monitoring (324 articles).
9. Topic 5. Engine- and machining-failure prediction (308 articles).
10. Topic 1. Cities and infrastructure (304 articles).
11. Topic 7. Robots (300 articles).
12. Topic 12. Logistics and service (231 articles).
13. Topic 3. Metaverse and virtual reality (198 articles).

As mentioned earlier, the results of the LDA model are characterized by overlap, while conversely, the identified subtopics can be considered together in scientific publications.

6.3. The BERTopic Model

When analyzing publications on the topic of digital twins using the BERTopic model, we found that the existing 'topics' are very diverse. In total, over 104 topics were identified. This means that there are approximately 81.6 publications per topic. This value shows that the research areas on the digital twin are incredibly diverse. The BERTopic model did not automatically assign 3683 articles to any topic. This means that only 4804 of the 8488 articles were sampled and divided into topics.

The BERTopic analysis makes it possible to identify the most important topics in the publications on the digital twin. The most popular 10 topics comprise 1536 articles or 18% of the total abstract-dataset (see Table 4).

Table 4. Number of articles by topic (data taken from Scopus, records from January 1993 to September 2022).

Topic Name and Number	Number of Articles on Topic
1_Robot_robots_robotic_human	283
2_Digital_twins_digital twins_twin	257
3_Construction_BIM_building_information	239
4_Power_grid_energy_power grid	175
5_Machining_cutting_tool_process	161
6_Patients_healthcare_health_medicine	149
7_Ship_vessel_ships_marine	139
8_Fatigue_crack_damage_structural	133
9_Teaching_students_education_learning	120
10_Ontology_knowledge_semantic_ontologies	107
11_Logistics_supply_supply chain_chain	106
12_Maintenance_predictive maintenance_prediction	90
13_City_urban_cities_smart	87
14_Systems_MBSE_engineering_systems engineering	87
15_Blockchain_sharing_data_decentralized	84
16_Security_attack_cyber_attacks	80
17_Bridge_structural_bridges_monitoring	79
18_Fault_diagnosis_fault diagnosis_faults	78
19_Battery_batteries_lithium-ion_charging	70
20_Driving_vehicle_vehicles_traffic	61

As the topic clusters are generated for the publications, the user can choose the option for the BERTopic model to suggest sample abstracts that most closely match the concept described by an individual topic cluster. This analysis was performed for the eight clusters.

1. Robots and virtual reality (keywords: robot, robots, robotic, human, assembly). 283 articles. Many publications deal with digital twins of assembly lines or factory floors in the virtual- and augmented-reality format. Examples of publications on this topic include: using digital-twin technologies to improve the quality and efficiency of robotic assemblies [35], training the digital twin of a robot in a virtual environment to reduce the training time of the real robot [36], and using the digital twin of an agricultural robot that receives information from the physical twin [37].
2. Digital twin (keywords: digit3al, twins, digital twins, twin, digital twin). 257 articles. This topic is devoted to the conceptual study of digital twins. In particular, this category includes reviews that deal with the analysis of existing concepts and definitions of digital twins, the applications of digital twins in the context of the sustainable-production paradigm [38], and the framework and analysis of successful practices to implement digital twins [33].
3. Construction and BIM technologies (keywords: construction, BIM, building, information, management). 239 articles. Creating digital twins of construction objects

in the design phase and digital shadows of existing buildings makes it possible to optimize financial costs and shorten the construction cycle. Publications on this topic address possible approaches to integrating physical and digital twins throughout the construction process by including the third component, i.e., the social network, in these models [34]; an analysis is presented for the cases where a digital twin is used to automate construction processes [39].

4. Energy and power (keywords: power, grid, energy, power grid, distribution). 175 articles. Digital twins can be used to model their processes and objects and digital shadows. Articles on digital twins in power grids and energy distribution are devoted to the digital twins of energy facilities, e.g., the digital twin of a transformer [40] and the digital twin of electric grids [41].
5. Machining (keywords: machining, cutting, tool, process, grinding). 161 articles. Publications on digital twins of manufacturing equipment used for machining parts. Publications on this topic deal with digital twins of CNC cutting technologies [42], digital twins for grinding with abrasive belts [43], and digital twins for CNC plunge-cut grinding [44].
6. Medical applications (keywords: patients, healthcare, health, medicine, patient). 149 articles. Articles on this topic cover digital twins as decision-support instruments [45], digital twins of humans, including ethical aspects [46] and perspectives of digital-twin technology in medicine [47].
7. Ships and shipbuilding (keywords: ship, vessels, machine, vessels). 139 articles. Key topics in this cluster include marine engines (development of "the generic procedure for the creation and usage of a complete system simulation for propulsion systems of ships with a focus on complex hybrid systems" [48]), the role of digital twins in shipbuilding and the benefits of this technology [49], the use of sensors in a research vessel to obtain data on its operation ("ship as an ideal platform from which to explore a definitive trend in the future marine industry: digital twin technology. This is a digital real-time in-context operational mime of an asset, which connects the digital and real word representations towards actionable insights" [50]).
8. Structural damage and fatigue testing (keywords: fatigue, crack, damage, structural, crack growth). 133 articles. Key topics in this cluster include structural-health-monitoring systems using digital-twin technologies [51], damage prediction and modeling under creep conditions [52], development of a crack-growth algorithm using an airframe digital twin [53].

An important advantage of the considered model is that the popularity of different topics can be represented on a time scale. The figure shows the frequency of occurrence of the eight most important topics in the last six years. The reason for the decrease in 2022 is that only articles published in the first eight months of the year were included in the dataset (see Figure 9). The vertical axis plots the number of articles devoted to a particular topic. It can be seen that even the most popular topic was covered in fewer than 100 articles in 2021, which is less than 3.3% of all articles published that year. This means that a consensus on which topics are central to the field has yet to be reached, i.e., the most popular topic in 2021 could potentially be replaced by another one in 2022.

Analyzing information from previous years, we can see that interest in topics such as the use of digital twins on assembly lines and factory floors (0_robot_robots_robotic_human has steadily increased since 2016. By 2021, topic 0 remained the most popular, while BIM technologies (2_construction_bim_building_information) took second place.

Topic 3 (3_power_grid_energy_power grid) and topic 4 (4_machining_cutting_tool_process) rank third and fifth respectively. Publications on the conceptual study of the digital twin (1_digital_twins_digital twins_twin) rank fourth. The fifth topic (5_patients_healthcare_health_medicine) and the sixth topic (6_ship_vessel_ships_marine) rank fifth and sixth, respectively.

Topic 7 (7_fatigue_crack_damage_structural) has existed since 2012, and ranks seventh.

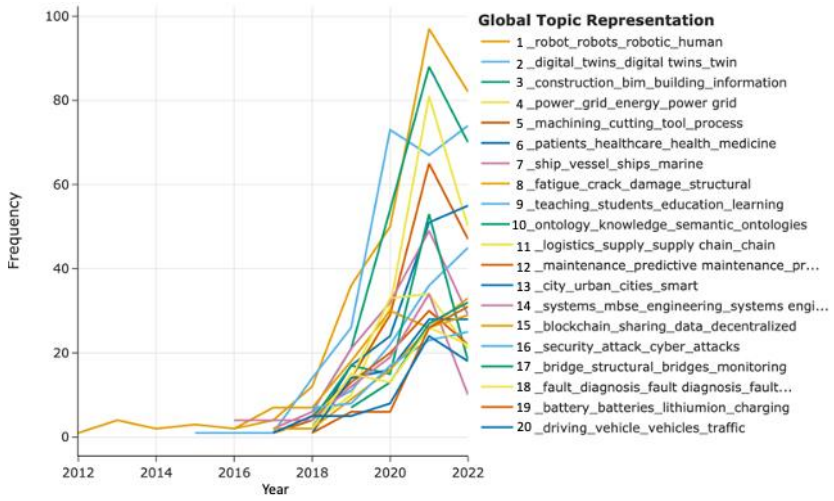


Figure 9. Change in the frequency of occurrence of topics on digital twins (from January 2012 to September 2022).

6.4. Comparison of LDA and BERTopic Analysis Results

6.4.1. Comparison of Model Performance Based on Unsupervised Clustering

There are several articles comparing the results of LDA and BERTopic modeling in different scientific fields. Most authors agree that the BERTopic model often shows better results than the LDA model [20,54]. The BERTopic not only shows more interpretable results, but also does not require preprocessing of data, and consumes fewer computer resources [54].

Unsupervised clustering of topics followed by dimensionality reduction was used to compare the models. It should be noted that these methods were used in this study only to determine which model generally performed better.

Clustering based on k-means is used to group unlabeled data. Clustering results can be improved by dimensionality reduction. This means that topic-modeling methods use a smaller number of terms [55]. There are three basic methods of dimensionality reduction:

- Uniform manifold approximation and projection, or UMAP, which creates a high-dimensional graph and then converts it to a low-dimensional one that should be structurally similar to the first one [56].
- Principal component analysis, or PCA, which reduces a large set of variables to a smaller set.
- The t-distributed stochastic neighbor embedding, or the t-SNE technique, transforms high-dimensional vectors into lower dimensional vectors, and preserves the relative similarity to the original [55].

The LDA and BERTopic models based on the digital-twin articles-dataset were clustered, and then dimensionality reduction was performed. The clustering results were compared, using the silhouette score. This method helps evaluate which clustering method performs better and which model is more accurate. The higher the score, the better the clusters are separated from each other (see Table 5).

Table 5. BERTopic and LDA topic-modeling silhouette scores compared (data taken from Scopus, records from January 1993 to September 2022).

Clustering Type	LDA	BERTopic
PCA	−0.00777	0.33081
UMAP	0.06249	0.34919
t-SNE	0.09098	0.36905

The clustering results show that the BERTopic model is more accurate than the LDA, and that the t-SNE method is the most accurate for clustering.

6.4.2. Comparison of the Models Based on Topics Interpretation

The LDA analysis does not initially exclude any abstracts from the sample. The model is built on 8488 abstracts of the articles. One of the most important parameters of this model is its ability to find multiple topics in an article. This means that when comparing LDA and BERTopic models, it may be useful to optimize the LDA model and select only one dominant topic for each abstract in a sample.

The other features of the LDA model are overlapping topics and topics that have subtopics. Thus, all articles in a sample have one of 13 topics that can be interpreted and 7 topics that contain two or more subtopics: 2460 abstracts can be assigned to mixed topics and 6028 to unique topics.

The BERTopic model excluded 3683 abstracts as mixed or uninterpretable, and focused on 4804 articles. One of the most important features of the result is a strong fragmentation of topics, or approximately 81.6 articles per topic. However, at the same time, all the topics found show strong differences, and can be interpreted.

The topics received after the analysis are listed in descending order in Table 6, below.

Table 6. BERTopic and LDA topic-modeling results comparison (data taken from Scopus, records from January 1993 to September 2022).

No	Topics (BERTopic)	Topics (LDA)
1.	1_Robot_robots_robotic_human	Topic 10. Transformation, intelligence, companies, solutions, technological
2.	2_Digital_twins_digital twins_twin	Topic 2. Human, construction, bim, safety, robot
3.	3_Construction_bim_building_information	Topic 15. Thermal, temperature, power, flow, experimental
4.	4_Power_grid_energy_power grid	Topic 6. Test, planning, methodology, solution, flexible
5.	5_Machining_cutting_tool_process	Topic 14. IoT, cloud, CPS, knowledge, services
6.	6_Patients_healthcare_health_medicine	Topic 9. Material, properties, measurements, materials, element
7.	7_Ship_vessel_ships_marine	Topic 17. Equipment, assembly, fault, line, workshop
8.	8_Fatigue_crack_damage_structural	Topic 8. Power, construction, vehicle, safety, traffic
9.	9_Teaching_students_education_learning	Topic 13. Networks, edge, communication, computing, security
10.	10_Ontology_knowledge_semantic_ontologies	Topic 20. Mining, heritage, infrastructure, railway, equipment
11.	11_Logistics_supply_supply chain_chain	Topic 18. Security, business, drilling, literature, DTs
12.	12_Maintenance_predictive maintenance_prediction	Topic 11. Structural, damage, health, sensor, bridge
13.	13_City_urban_cities_smart	Topic 5. Machining, engine, prediction, tool, crack
14.	14_Systems_mbse_engineering_systems engineering	Topic 1. Energy, city, urban, scheduling, battery
15.	15_Blockchain_sharing_data_decentralized	Topic 7. Robot, reality, robots, robotic, space
16.	16_Security_attack_cyber_attacks	Topic 4. Construction, students, education, modeling, knowledge
17.	17_Bridge_structural_bridges_monitoring	Topic 16. Supply, chain, patient, health, medical
18.	18_Fault_diagnosis_fault diagnosis_faults	Topic 12. Value, service, logistics, lifecycle, context
19.	19_Battery_batteries_lithiumion_charging	Topic 3. Reality, metaverse, VR, augmented, commissioning
20.	20_Driving_vehicle_vehicles_traffic	Topic 19. Structure, structural, point, construction, wind

There are only four topics that were received by both the LDA and BERTopic models:

1. Robotics and robots (1_Robot_robots_robotic_human (BERTopic)/Topic 7. Robot, reality, robots, robotic, space (LDA).
2. Cities and infrastructure (13_City_urban_cities_smart (BERTopic)/Topic 1. Energy, city, urban, scheduling, battery (LDA)).
3. Machining and engines (5_Machining_cutting_tool_process (BERTopic)/Topic 5. Machining, engine, prediction, tool, crack (LDA)).
4. Bridges and structural damage (17_Bridge_structural_bridges_monitoring (BERTopic)/Topic 11. Structural, damage, health, sensor, bridge (LDA)).

These topics are most clearly defined in the corpus of abstracts, as both models can find them. The other results show the wide variety of publication trends on digital twins,

from blockchain to education, from cyberattacks to metaverse, and from blockchain to supply chain.

7. Conclusions

We formulated three research questions at the beginning of the study. The first was whether the digital twin is a mature topic with different 'schools of thought' and approaches, or whether it is still a new topic and a kind of marketing label with a vague background.

When analyzing the keywords, we noticed that a cloud of terms has accumulated around the term 'digital twin', which are mentioned in publications exclusively in connection with this term. This could indicate a wide variety of keywords and a specific field of research that is developing around the digital-twin concept. The answer to the first question is that studies on this topic are still in the early stages, but the various approaches and theories are now coalescing into a distinct field of research.

The second question was whether it is possible to identify clearly defined research topics related to the digital twin. The LDA model identified several key topics, but the interpretation remains unclear. The BERTopic model has identified more than 100 topics, eight of which are considered the most important.

The BERTopic model has shown that researchers are most interested in the digital twins of manufacturing equipment. Thus, the answer to the second question is that well-defined topics can be recognized in publications, including those related to digital twins.

The third question asked which topics might become most prominent in the coming years. Analysis of the graph shows that the number of publications on the following topics is increasing: (1) digital twins of assembly lines and robots; (2) BIM technology and digital twins in construction; (3) digital twins of power grids and energy distribution. This trend is likely to continue in the coming years. It is noted that interest in conceptual studies of the digital twin and in works evaluating the digital twin's role in Industry 4.0 is gradually decreasing.

8. Future Research

While we have generally answered the questions raised by the analysis, an open challenge is to develop a new research methodology that can contribute to better accuracy in processing large text-datasets. This can hypothetically be achieved by combining tools for automated analysis, text mining, and machine-learning models.

However, we report mixed results on this topic, and plan to address this by answering the following questions in our future research:

1. Do the results of topic modeling performed on a dataset of publication abstracts on the topic of digital twins match those of a dataset of full-text publications? Clearly, the larger the dataset used to train the machine model, the more accurate the analysis results, including predictions for the topics of new publications.
2. To what extent does the automation of topic analysis in publications enable the preservation of the nuances of the text, i.e., the detection of all possible topics mentioned? As for the LDA model, in general, it is impossible to clearly separate the topics of the texts (one text can be assigned to several topics); moreover, it is difficult to interpret the results. The advantage of the BERTopic model is that the text can be assigned to only one topic; other topics of the source text are lost in this case, but the results are easier to interpret.
3. Are there better topic-modeling techniques that can produce more accurate results? For example, several authors use a combination of the LDA model and BERTopic embeddings. According to their analysis, this mixed method is better than the LDA or BERTopic used individually [55,57,58]. Although finding the best method for topic modeling was not one of the goals of this paper, finding the best method for topic modeling can be a goal for future research.

Regardless of the answers to these questions, text-mining tools should be used to train a machine model properly. The texts should be cleaned of stop words that can significantly distort the study results, and then be marked up for further analysis.

Thus, in addition to the operations required to train the machine model, the text-mining phase may include the generation of graphs of co-authorships and citations, as well as the generation of a network for the co-occurrence of keywords, which generally enhances the understanding of the main research trends within the field under consideration.

The research methodology can be tested on various topics of scientific publications, e.g., Industry 4.0, industrial internet of things, cyber-physical systems, etc. The issues raised above could also be explored by analyzing these topics.

Author Contributions: Conceptualization and methodology, K.K., Y.R. and A.B.; data curation and visualization, K.K.; writing—original draft preparation, K.K.; writing—review and editing, Y.R. and K.K.; project administration, Y.R.; supervision and funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: The research is funded by the Ministry of Science and Higher Education of the Russian Federation under the strategic academic leadership program "Priority 2030" (Agreement 075-15-2021-1333 dated 30 September 2021).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from https://zenodo.org/record/7377950#.Y4ZGdHZBxPY.

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

References

1. Grieves, M.; Vickers, J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In *Transdisciplinary Perspectives on Complex Systems*; Kahlen, F.-J., Flumerfelt, S., Alves, A., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 85–113, ISBN 978-3-319-38754-3.
2. Google Books Ngram Viewer. Available online: https://books.google.com/ngrams/graph?content=digital+twin&year_start=2010&year_end=2019&corpus=26&smoothing=3&direct_url=t1%3B%2Cdigital%20twin%3B%2Cc0 (accessed on 27 July 2022).
3. Kritzinger, W.; Karner, M.; Traar, G.; Henjes, J.; Sihn, W. Digital Twin in Manufacturing: A Categorical Literature Review and Classification. *IFAC-Pap.* **2018**, *51*, 1016–1022. [CrossRef]
4. Negri, E.; Fumagalli, L.; Macchi, M. A Review of the Roles of Digital Twin in CPS-Based Production Systems. *Procedia Manuf.* **2017**, *11*, 939–948. [CrossRef]
5. Rossmann, A.; Hertweck, D. Digital Twins: A Meta-Review on Their Conceptualization, Application, and Reference Architecture. In Proceedings of the 55th Hawaii International Conference on System Sciences, Maui, HI, USA, 4 January 2022.
6. Kuehner, K.; Scheer, R.; Straßburger, S. Digital Twin: Finding Common Ground—A Meta-Review. *Procedia CIRP* **2021**, *104*, 1227–1232. [CrossRef]
7. Mejia, C.; Wu, M.; Zhang, Y.; Kajikawa, Y. Exploring Topics in Bibliometric Research Through Citation Networks and Semantic Analysis. *Front. Res. Metr. Anal.* **2021**, *6*, 742311. [CrossRef] [PubMed]
8. Borovkov, A.I.; Gamzikova, A.A.; Kukushkin, K.V.; Ryabov, Y.A. *Digital Twins in the High-Technology Manufacturing Industry. A Preliminary Research Report (September 2019)*; POLYTECH-PRESS: St. Petersburg, Russia, 2019; ISBN 978-5-7422-6922-9. (In Russian)
9. Warke, V.; Kumar, S.; Bongale, A.; Kotecha, K. Sustainable Development of Smart Manufacturing Driven by the Digital Twin Framework: A Statistical Analysis. *Sustainability* **2021**, *13*, 10139. [CrossRef]
10. Scopus—Document Search | Signed in. Available online: https://www.scopus.com/search/form.uri?display=basic#basic (accessed on 6 July 2022).
11. Dang, S. Text Mining: Techniques and Its Application. *Int. J. Eng. Technol. Innnovation* **2014**, *1*, 22–25.
12. Talib, R.; Kashif, M.; Ayesha, S.; Fatima, F. Text Mining: Techniques, Applications and Issues. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 20–25. [CrossRef]
13. Zdonek, I. The Role of Word and N-Gram Frequency Analysis 2 in Inference of the Content of Scientific Publication. *Sci. Pap. Silesian Univ. Technol. Organ. Manag. Ser.* **2020**, *2020*, 21–31. [CrossRef]
14. Gensim. Topic Modelling for Humans. Available online: https://radimrehurek.com/gensim/ (accessed on 28 July 2022).
15. Campbell, J.C.; Hindle, A.; Stroulia, E. Latent Dirichlet Allocation. In *The Art and Science of Analyzing Software Data*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 139–159, ISBN 978-0-12-411519-4.

16. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimed. Tools Appl.* **2017**, *78*, 15169–15211. [CrossRef]
17. Using BERT embeddings for Text Modeling (in Russian). Available online: https://habr.com/ru/post/653443/ (accessed on 6 July 2022).
18. Grootendorst, M. BERTopic. Available online: https://github.com/MaartenGr/BERTopic (accessed on 6 July 2022).
19. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794. [CrossRef]
20. Egger, R.; Yu, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef] [PubMed]
21. Tao, F.; Cheng, J.; Qi, Q.; Zhang, M.; Zhang, H.; Sui, F. Digital Twin-Driven Product Design, Manufacturing and Service with Big Data. *Int. J. Adv. Manuf. Technol.* **2018**, *94*, 3563–3576. [CrossRef]
22. Ivanov, D. Predicting the Impacts of Epidemic Outbreaks on Global Supply Chains: A Simulation-Based Analysis on the Coronavirus Outbreak (COVID-19/SARS-CoV-2) Case. *Transp. Res. Part E Logist. Transp. Rev.* **2020**, *136*, 101922. [CrossRef] [PubMed]
23. Tao, F.; Zhang, H.; Liu, A.; Nee, A.Y.C. Digital Twin in Industry: State-of-the-Art. *IEEE Trans. Ind. Inform.* **2019**, *15*, 2405–2415. [CrossRef]
24. Perianes-Rodriguez, A.; Waltman, L.; van Eck, N.J. Constructing Bibliometric Networks: A Comparison between Full and Fractional Counting. *J. Informetr.* **2016**, *10*, 1178–1195. [CrossRef]
25. Tao, F.; Zhang, M.; Nee, A.Y.C. *Digital Twin Driven Smart Manufacturing*; Academic Press: Cambridge, MA, USA, 2019; ISBN 978-0-12-817631-3.
26. Lu, Y.; Min, Q.; Liu, Z.; Wang, Y. An IoT-Enabled Simulation Approach for Process Planning and Analysis: A Case from Engine Re-Manufacturing Industry. *Int. J. Comput. Integr. Manuf.* **2019**, *32*, 413–429. [CrossRef]
27. Liu, J.; Zhou, H.; Liu, X.; Tian, G.; Wu, M.; Cao, L.; Wang, W. Dynamic Evaluation Method of Machining Process Planning Based on Digital Twin. *IEEE Access* **2019**, *7*, 19312–19323. [CrossRef]
28. Wang, Y.; Wang, X.; Liu, A. Digital Twin-Driven Analysis of Design Constraints. *Procedia CIRP* **2020**, *91*, 716–721. [CrossRef]
29. Rosen, R.; von Wichert, G.; Lo, G.; Bettenhausen, K.D. About The Importance of Autonomy and Digital Twins for the Future of Manufacturing. *IFAC-Pap.* **2015**, *48*, 567–572. [CrossRef]
30. Söderberg, R.; Wärmefjord, K.; Carlson, J.S.; Lindkvist, L. Toward a Digital Twin for Real-Time Geometry Assurance in Individualized Production. *CIRP Ann.* **2017**, *66*, 137–140. [CrossRef]
31. Schleich, B.; Anwer, N.; Mathieu, L.; Wartzack, S. Shaping the Digital Twin for Design and Production Engineering. *CIRP Ann.* **2017**, *66*, 141–144. [CrossRef]
32. Tuegel, E.J.; Ingraffea, A.R.; Eason, T.G.; Spottswood, S.M. Reengineering Aircraft Structural Life Prediction Using a Digital Twin. *Int. J. Aerosp. Eng.* **2011**, *2011*, 154798. [CrossRef]
33. Sivarethinamohan, R.; Sujatha, S. Reimagining the Digital Twin: Powerful Use Cases for Industry 4.0. In *Advances in Mechanical Engineering*; Manik, G., Kalia, S., Sahoo, S.K., Sharma, T.K., Verma, O.P., Eds.; Lecture Notes in Mechanical Engineering; Springer: Singapore, 2021; pp. 175–182, ISBN 9789811609411.
34. Turk, Ž.; Klinc, R. A Social–Product–Process Framework for Construction. *Build. Res. Inf.* **2020**, *48*, 747–762. [CrossRef]
35. Li, X.; He, B.; Zhou, Y.; Li, G. Multisource Model-Driven Digital Twin System of Robotic Assembly. *IEEE Syst. J.* **2021**, *15*, 114–123. [CrossRef]
36. Hassel, T.; Hofmann, O. Reinforcement Learning of Robot Behavior Based on a Digital Twin. In Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020), Valletta, Malta, 4 July 2022; pp. 381–386.
37. Lumer-Klabbers, G.; Hausted, J.O.; Kvistgaard, J.L.; Macedo, H.D.; Frasheri, M.; Larsen, P.G. Towards a Digital Twin Framework for Autonomous Robots. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021; pp. 1254–1259.
38. Ball, P.; Badakhshan, E. Sustainable Manufacturing Digital Twins: A Review of Development and Application. In *Sustainable Design and Manufacturing*; Scholz, S.G., Howlett, R.J., Setchi, R., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 262, pp. 159–168, ISBN 9789811661273.
39. Al-Saeed, Y.; Edwards, D.J.; Scaysbrook, S. Automating Construction Manufacturing Procedures Using BIM Digital Objects (BDOs): Case Study of Knowledge Transfer Partnership Project in UK. *Constr. Innov.* **2020**, *20*, 345–377. [CrossRef]
40. Delong, Z.; Zhijun, Y.; Huipeng, C.; Peng, Z.; Jiliang, L. Research on Digital Twin Model and Visualization of Power Transformer. In Proceedings of the 2021 IEEE International Conference on Networking, Sensing and Control (ICNSC), Xiamen, China, 3–5 December 2021; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2021.
41. Zolin, D.S.; Ryzhkova, E.N. Digital Twins for Electric Grids. In Proceedings of the 2020 International Russian Automation Conference (RusAutoCon), Sochi, Russia, 6–12 September 2021; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; pp. 175–180.
42. Ward, R.; Sun, C.; Dominguez-Caballero, J.; Ojo, S.; Ayvar-Soberanis, S.; Curtis, D.; Ozturk, E. Machining Digital Twin Using Real-Time Model-Based Simulations and Lookahead Function for Closed Loop Machining Control. *Int. J. Adv. Manuf. Technol.* **2021**, *117*, 3615–3629. [CrossRef]

43. Wang, Y.-H.; Lo, Y.-C.; Lin, P.-C. A Normal Force Estimation Model for a Robotic Belt-Grinding System. In Proceedings of the 2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Boston, MA, USA, 6–9 July 2020; IEEE: Boston, MA, USA, 2020; pp. 1922–1928.
44. Akintseva, A.V.; Pereverzev, P.P.; Omel'chenko, S.V.; Kopyrkin, A.A. Digital Twins and Multifactorial Visualization of Shaping in CNC Plunge-Cut Grinding. *Russ. Eng. Res.* **2021**, *41*, 671–675. [CrossRef]
45. Gaebel, J.; Keller, J.; Schneider, D.; Lindenmeyer, A.; Neumuth, T.; Franke, S. The Digital Twin: Modular Model-Based Approach to Personalized Medicine. *Curr. Dir. Biomed. Eng.* **2021**, *7*, 223–226. [CrossRef]
46. Boulos, M.K.; Zhang, P. Digital Twins: From Personalised Medicine to Precision Public Health. *J. Pers. Med.* **2021**, *11*, 745. [CrossRef]
47. De Maeyer, C.; Markopoulos, P. Future Outlook on the Materialisation, Expectations and Implementation of Digital Twins in Healthcare. In Proceedings of the 34th British HCI Conference, London, UK, 20–21 July 2021; BCS Learning and Development Ltd.: Swindon, UK, 2021; pp. 180–191.
48. Jannsen, L.-E. Development of a Simulation Environment for Hybrid Propulsion Drive Trains: Utilization of a Holistic Approach to Predict the Dynamic Behavior in the Early Design Stage. In Proceedings of the Volume 1: Offshore Technology, Online, 3–7 August 2020; American Society of Mechanical Engineers: New York, NY, USA, 2020; p. V001T01A027.
49. Morais, D.; Goulanian, G.; Danese, N. The Future Reality of the Digital Twin as a Cross-Enterprise Marine Asset. In Proceedings of the 19th International Conference on Computer Applications in Shipbuilding 2019, Rotterdam, The Netherlands, 24–26 September 2019; The Royal Institution of Naval Architects: London, UK, 2019; Volume 2.
50. Bekker, A. Exploring the Blue Skies Potential of Digital Twin Technology for a Polar Supply and Research Vessel. In Proceedings of the 13th International Marine Design Conference Marine Design XIII (IMDC 2018), Helsinki, Finland, 10–14 June 2018; CRC Press: Boca Raton, FL, USA, 2018; Volume 1, pp. 135–146.
51. Loghin, A.; Ismonov, S. Assessment of Crack Path Uncertainly Using 3d Fea and Response Surface Modeling. In Proceedings of the AIAA Scitech Forum, Orlando, FL, USA, 6–10 January 2020; American Institute of Aeronautics and Astronautics Inc.: Reston, VA, USA, 2020; Volume 1. Part F.
52. Wang, K.; Wang, X.; Wen, J.; Zhang, X.; Gong, J.; Tu, S. Creep Rupture: From Physical Failure Mechanisms to Lifetime Prediction of Structures. *Jixie Gongcheng Xuebao J. Mech. Eng.* **2021**, *57*, 132–152. [CrossRef]
53. Ocampo, J.; Millwater, H.; Crosby, N.; Gamble, B.; Hurst, C.; Reyer, M.; Mottaghi, S.; Nuss, M. An Ultrafast Crack Growth Lifing Model to Support Digital Twin, Virtual Testing, and Probabilistic Damage Tolerance Applications. In *ICAF 2019—Structural Integrity in the Age of Additive Manufacturing*; Niepokolczycki, A., Komorowski, J., Eds.; Lecture Notes in Mechanical Engineering; Springer International Publishing: Cham, Switzerland, 2020; pp. 145–158, ISBN 978-3-030-21502-6.
54. Robinson, D. Is LDA Topic Modeling Dead? Available online: https://towardsdatascience.com/is-lda-topic-modeling-dead-9543c18488fa (accessed on 20 November 2022).
55. George, L.; Sumathy, P. An Integrated Clustering and BERT Framework for Improved Topic Modeling. *Res. Sq.* **2022**. [CrossRef]
56. Understanding UMAP. Available online: https://pair-code.github.io/understanding-umap/ (accessed on 20 November 2022).
57. Atagun, E.; Hartoka, B.; Albayrak, A. Topic Modeling Using LDA and BERT Techniques: Teknofest Example. In Proceedings of the 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 15–17 September 2021; IEEE: Ankara, Turkey, 2021; pp. 660–664.
58. Shao, S. Contextual Topic Identification. Available online: https://blog.insightdatascience.com/contextual-topic-identification-4291d256a032 (accessed on 20 November 2022).