Topic Modeling Analytics of Digital Economy Research: Trends and Insights

Umawadee Detthamrong[1], Lan Thi Nguyen[2,*], Yuttana Jaroenruen[3], Akkharawoot Takhom[4], Vispat Chaichuay[2], Knitchepon Chotchantarakun[5], Wirapong Chansanam[2,*]

[1]College of Local Administration, Khon Kaen University, Khon Kaen, THAILAND.
[2]Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen, THAILAND.
[3]Digital Content and Media Program, School of Informatics, Walailak University, Thai Buri, Nakhon Si Thammarat, THAILAND.
[4]Faculty of Engineering, Thammasat School of Engineering, Thammasat University, Pathum Thani, THAILAND.
[5]Department of Information Studies, Faculty of Humanities and Social Sciences, Burapha University, Chonburi, THAILAND.

ABSTRACT

This paper incorporates scholarly articles, conference papers, and published books, analyzing these sources to explore how topic modeling is used to uncover trends, identify research areas, and contribute to understanding the digital economy. The study utilizes bibliography data and Python libraries for topic modeling to examine 8,321 documents from the Scopus database. Through rigorous analysis, three distinct topics were identified, and their development trends were traced, providing insights into the shifting focus within the field. The study also validates the classification proposed by the coherence coefficient and contributes theoretically and methodologically by employing Latent Dirichlet Allocation (LDA) topic modeling technology within text analytics. The results of this study indicated that "Digital Transformation" emerged as the most popular, accounting for 56.6% of tokens. However, the "Digitalization" topic exhibited relatively lower popularity, representing only 21.2% of tokens. The findings help enhance the understanding of global research trends and offer a valuable framework for comprehending digital economy research. Furthermore, the study emphasizes the significance of analyzing digitalization, data governance, and digital transformation, highlighting the efficacy of LDA as a powerful tool for efficient and accurate text analytics. The research findings are especially pertinent for scholars in information science, data mining, econometrics, and bibliometrics. They offer a foundational understanding of topic modeling, facilitating further investigation and exploration in these fields. However, the study identifies two limitations, including the limited dataset extracted solely from Scopus and the restriction to abstracts rather than full texts. Future research should consider expanding data sources and incorporating full texts from authoritative articles in multiple languages to attain more comprehensive outcomes.

Keywords: Research trends, Topic development, Digital economy, Topic modeling, Data governance.

Correspondences:

Wirapong Chansanam
Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen-40002, THAILAND.
Email: wirach@kku.ac.th
ORCID ID:0000-0001-5546-8485

Lan Thi Nguyen
Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen-40002, THAILAND.
Email: nguyenth@kku.ac.th
ORCID ID: 0000-0002-8848-2168

Received: 30-05-2023;
Revised: 21-12-2023;
Accepted: 12-03-2024.

INTRODUCTION

The digital economy is growing rapidly, and it is estimated to account for a significant share of global economic output.[1] In 2016, the global digital economy was worth $11.5 trillion, and it is projected to reach $24.2 trillion by 2025. In the United States, the digital economy is responsible for 7.9 million jobs, and it is projected to create an additional 1.4 million jobs by 2025.[2-4] The digital economy is a term used to describe the economic activities that are enabled by digital technologies.[1] These activities include the production, distribution, and consumption of digital goods and services. The digital economy is having a major impact on the global economy. It is creating new jobs, new industries, and new ways of doing business. The digital economy is also changing the way we live and work. For example, we are now able to shop, bank, and communicate with friends and family online.[5] However, the digital economy also poses some challenges of the digital divide. The digital divide refers to the gap between those who have access to digital technologies and those who do not. The digital divide can have a negative impact on economic growth and social inclusion. Another challenge of the digital economy is the rise of cybercrime. Cybercrime is a growing problem, and it can have a significant financial impact on businesses and individuals.[6] The digital economy is still in its early stages, and it is difficult to predict what the future holds. However, it is clear that the digital economy will continue to grow and have a major impact on the global economy.[7]

DOI: 10.5530/jscires.13.2.35

Publishing Partner : EManuscript Tech. [www.emanuscript.in]

The digital economy has fundamentally transformed our way of life and professional endeavors. It has revolutionized how we shop, conduct financial transactions, and connect with loved ones. Moreover, this digital revolution has extended its influence to education, employment, and recreation.

In light of these challenges, examining notable trends within the digital economy is important. One such trend is the proliferation of mobile computing. Mobile devices have gained remarkable power and ubiquity, reshaping our means of accessing and utilizing digital services. Additionally, the ascent of cloud computing has gained significant traction. Cloud computing represents a paradigm for delivering computing services via the internet, and its popularity is growing steadily. Businesses are increasingly adopting cloud computing as it offers cost savings and enhanced operational efficiency. Another consequential development within the digital economy is the emergence of big data analytics. This field encompasses extracting valuable insights from extensive datasets and finds utility in improving organizations' decision-making, marketing strategies, and operational processes. Furthermore, Artificial Intelligence (AI) has experienced a meteoric rise, significantly impacting the digital economy. AI finds applications across various domains, including customer service, fraud detection, and product development. By understanding these trends and developments, we can navigate the complexities of the digital economy and address its challenges. Efforts to bridge the digital divide, along with leveraging mobile computing, cloud computing, big data analytics, and AI, can contribute to fostering a more inclusive and prosperous digital ecosystem.[8-10]

Previous studies explored different aspects of digital technology on digital economy. For instance, Brynjolfsson and McAfee extensively explore the impact of digital technologies on the economy and society, focusing on artificial intelligence, robotics, and machine learning's transformative effects on industries, employment, and productivity.[1] Manyika *et al.*[11] analyze big data's implications across sectors, emphasizing its potential for enhancing decision-making, fostering innovation, and stimulating economic growth. McAfee *et al.*[12] highlight big data's transformative power, revealing how organizations revolutionize operations and decision-making. They stress analytics' pivotal role in optimizing business processes and gaining a competitive edge. Porter and Heppelmann[13] and Joachimsthaler *et al.*[14] discuss smart, connected products' impact on business models and customer experiences. Tapscott and Williams[15] explore challenges and opportunities in the digital economy, guiding companies on innovating in a rapidly evolving landscape. The World Economic Forum[16] focuses on workforce strategies and policy recommendations, shaping a prosperous digital future. Together, these works offer diverse insights into the digital economy's transformative potential, aiding organizations and policymakers navigating this rapidly changing landscape. In addition, other studies also mentioned that digital transformation play important roles for developing digital economy, including fostering mindset and skills of human resources, applications of digital technologies in organizations, etc.[8,10,17-19] Moreover, in order to adapt to the trends of digital economy, it is necessary for businesses to use digital technologies in digitizing and governing their data.[20-24]

In understanding the digital economy and its research, topic modeling is a valuable tool to identify and extract latent themes or topics from a collection of documents. It is widely employed in various fields, including natural language processing, information retrieval, text mining, and computational linguistics. Researchers can gain insights into the underlying themes and patterns within a body of literature by applying topic modeling. These text analytics aims to provide an overview of the key concepts, methods, and applications of topic modeling. The study incorporates scholarly articles, conference papers, and published books by analyzing these sources to explore how topic modeling is used to uncover trends, identify research areas, and contribute to understanding the digital economy.

This influential paper introduces Latent Dirichlet Allocation (LDA), a widely-used probabilistic topic modeling algorithm.[25] It thoroughly explains LDA's generative process and inference techniques. Griffiths and Steyvers[26] demonstrate LDA's application in uncovering latent topics within scientific articles. Blei[27] offers an overview of probabilistic topic modeling, exploring LDA's extensions and discussing its challenges and future directions. Mimno *et al.*[28] propose a method to enhance topic models' coherence using pointwise mutual information. Sievert and Shirley[29] introduce LDAvis, a web-based tool for visualizing and interpreting topics. Ramage *et al.*[30] introduce Labeled LDA, incorporating label information into the model, showing its utility in credit attribution tasks. Wang *et al.*[31] adapt topic models for image classification and annotation. Meimaris *et al.*[32] present a survey paper discussing various topic modeling techniques' strengths and limitations across diverse domains. Collectively, these works provide deep insights into LDA, its applications, interpretability enhancements, visualization tools, and extensions beyond textual data analysis. Topic modeling, a statistical technique, uncovers latent topics within documents, aiding content understanding, trend identification, and predictive analysis in various domains like market research, product development, and customer segmentation.

The existing literature on the digital economy and topic modeling reveals potential research gaps necessitating further exploration. Current studies lack a specific focus on applying topic modeling techniques to analyze the digital economy, particularly in understanding digital platforms, e-commerce, and the impact of digital transformation. While studies exist, they often statically analyze data within a limited timeframe, overlooking temporal dynamics in the digital economy. Moreover, there's a lack of research evaluating topic modeling techniques' suitability for capturing digital economy-related topics. Bridging disciplinary gaps and encouraging interdisciplinary research is crucial to comprehensively understand the digital economy's relationship with topic modeling. Ethical and societal implications of topic modeling in the digital economy, such as privacy concerns and societal impacts, require more attention. Challenges include coping with the vast and diverse data sources in the digital economy. Despite these hurdles, topic modeling offers benefits like trend identification, data-driven insights for improving products and services, and predictive analysis. It stands as a powerful tool for understanding and benefiting businesses, governments, and individuals in the digital economy.

In addition to the challenges and benefits mentioned above, it is important to note that topic modeling is a complex technique that requires careful planning and execution. To achieve the desired results, it is important to select a suitable topic model, collect the correct data, and train the model correctly. Additionally, it is important to interpret topic modeling results carefully and avoid making over-confident predictions. This study answers questions as follows:

What are trend topics of digital economy based on extracted data?

What contributions of topic modelling do influence the research of digital economy?

By addressing these gaps in the existing literature,[33-37] we can significantly enrich our comprehension of the digital economy and the effective utilization of topic modeling techniques within this domain. By bridging these gaps, we will foster a more profound insight into the digital economy and how topic modeling can be harnessed to its full potential. Moreover, this endeavor can lay the groundwork for future research endeavors that delve into the interconnections between these two fields and their broader societal ramifications. Furthermore, such efforts will offer valuable insights to policymakers, businesses, and researchers, equipping them with a comprehensive understanding of the opportunities, challenges, and implications that the digital economy presents from a topic-modeling standpoint.

METHODOLOGY

The primary objective of this study was to conduct a quantitative analysis of research articles on the digital economy, with specific emphasis on the content found in their abstracts. To achieve this, a methodological framework consisting of three sub-processes was employed: data retrieval, preprocessing, and topic analysis. The study utilized Latent Dirichlet allocation (LDA) topic models, enabling the identification of crucial topics embedded within the texts. Nonetheless, the research encountered challenges along the way, including the necessity to adequately preprocess text collections, carefully select appropriate model parameters, evaluate the model's reliability, and interpret the resulting topics. The sequence of using Latent Dirichlet Allocation (LDA) techniques is shown in Figure 1. By following this sequence, researchers can effectively apply LDA and relate techniques to uncover latent topics within a collection of documents, evaluate the quality of the model, and visualize the results to facilitate understanding and interpretation. This approach ensures a systematic and comprehensive analysis of the digital economy research topics, thereby educating and engaging the reader's understanding.

Data retrieval and pre-processing

This section provides a comprehensive overview of the data retrieval and preprocessing procedures conducted for the text analytics study. To obtain high-quality literature, the researchers accessed the Scopus Core Collection from 1953 to 2023, employing a carefully constructed search string (e.g., TS = "digital economy") to ensure the collection of solely relevant papers. Subsequently, inclusion and exclusion criteria were applied to

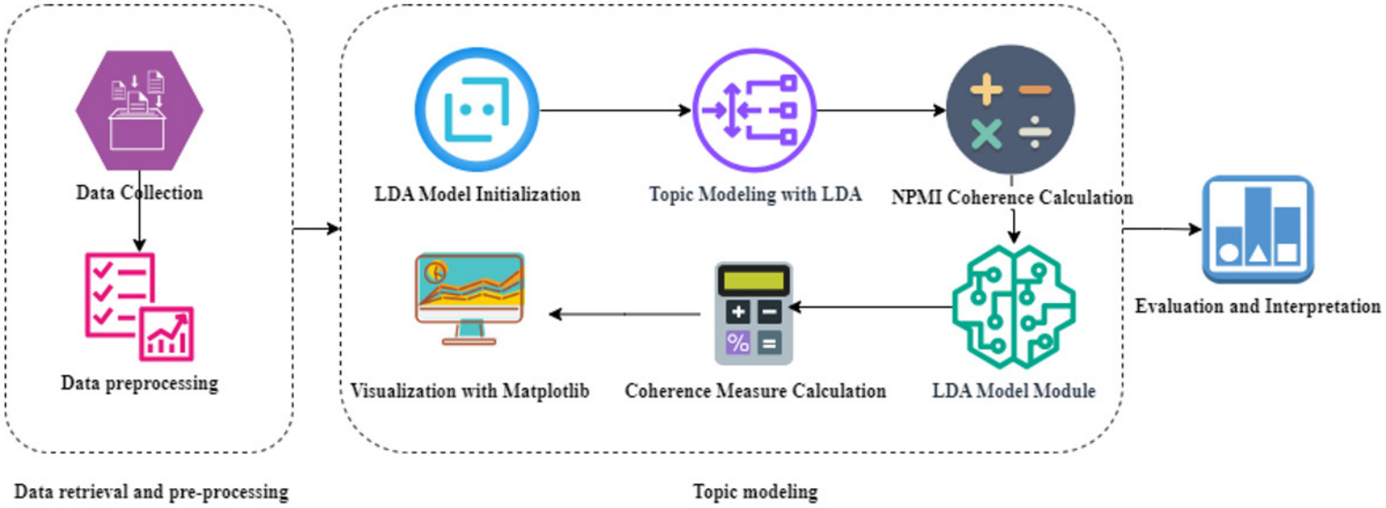

Data Collection
Data preprocessing
Data retrieval and pre-processing
LDA Model Initialization
Topic Modeling with LDA
NPMI Coherence Calculation
LDA Model Module
Coherence Measure Calculation
Visualization with Matplotlib
Topic modeling
Evaluation and Interpretation

Figure 1: The sequence of using Latent Dirichlet Allocation (LDA) techniques.

select 8,321 pertinent papers, which underwent analysis based on their titles, abstracts, year of publication, and journal titles. Scopus was deemed a suitable database for this study due to its broad coverage across various fields and its extensive utilization in bibliometric analysis.

In order to prepare the data for topic model mining, the researchers employed a range of pre-processing steps. Firstly, the text was segmented into tokens, with subsequent removal of punctuation, numbers, and unnecessary words. Additionally, the words were transformed into their base form utilizing Python programming language. The text pre-processing was carried out using PyCaret, an open-source and easily accessible tool renowned for its reliability.[38] A series of pre-processing steps were executed to extract representative keywords from the research articles. These steps encompassed the elimination of commonly used yet insignificant words, utilization of the unigram algorithm to identify common collocation phrases, and application of the TF-IDF algorithm to extract crucial keywords from the abstract section.[39,40]

Topic modeling

Topic modeling serves as a potent instrument for researchers to unveil concealed structures within document collections, facilitating data-driven decision-making and providing valuable insights into intricate subjects.[25] Nevertheless, the task of selecting the most suitable model presents challenges due to the distinct strengths and weaknesses exhibited by various models.[41] For instance, LDA is renowned for its proficiency in generating descriptive topics, while LSA excels in constructing a semantic representation of documents within a corpus.[42]

Despite its potential benefits, topic modeling can present challenges in understanding and interpretation.[43] In order to ensure the reliability and significance of the outcomes, researchers rely on metrics, such as perplexity and coherence to evaluate the modeling results. Perplexity quantifies the likelihood value of the model, while coherence is calculated using the Normalized Pointwise Mutual Information (NPMI) formula.[44] By assigning a high co-occurrence probability to word pairs, the NPMI formula yields modeling results that are highly comprehensible and interpretable, as in Equation (1).

$$NPMI\left(w_i, w_j\right) = \frac{log\frac{p\left(w_i,w_j\right)+\varepsilon}{p\left(w_i\right).p\left(w_j\right)}}{-log\left(p\left(w_i,w_j+\varepsilon\right)\right)} \qquad (1)$$

p(w) represents the probability that the word w exists in the provided document.

p($w_{i,}$ w_j) represents the probability that two words w_i, w_j appear together in the same context.

NPMI is a reliable measure for evaluating the quality of the model by assessing the association between word pairs. The topic modeling process in this study employed the LdaModel module from Gensim[45] and utilized the NPMI coherence indicator. We could determine the statistical significance of the obtained results by analyzing the probability of word co-occurrences within specific documents.[43,44]

It's essential to predefine crucial parameters like the number of topics (K), the apriori values governing topic distribution (α), and topic word distribution (η) in the LdaModel module of the Gensim library in Python. Then, the UCI (or CV) measure serves as an automatic coherence metric for evaluating topic coherence. The CoherenceModel class within the gensim package aids in identifying the optimal number of topics. Additionally, the C_v and C_umass algorithms are utilized to compute coherence scores, pinpointing the number of topics associated with the highest score. Following this, Latent Dirichlet Allocation (LDA) is commonly employed to categorize digital economy research into distinct topics. Subsequently, the pyLDAvis visualization tool is utilized to adjust word relevance accordingly.

RESULTS

The analysis results were displayed in three parts: Determination of optimal parameters, Topic naming and topic details, and Topic visualization.

Determination of optimal parameters

The Latent Dirichlet Allocation (LDA) model stands as one of the most widely utilized algorithms for topic modeling, having demonstrated promising outcomes in numerous studies.[46] It is important to highlight that the LDA model operates as an unsupervised machine-learning technique.

When utilizing the LdaModel module from the Gensim library in Python, it is crucial to predefine essential parameters, such as the number of topics (K), the apriori values of topic distribution (α), and topic word distribution (η). These parameters bear a significant influence on the effectiveness of topic mining.

The evaluation of topic model algorithms poses a formidable challenge, primarily due to the complexity and magnitude of the textual data involved. Human evaluation can be time-consuming and susceptible to biases, necessitating the development of theoretical evaluation models. Although these models cannot match the accuracy achieved through human-in-the-loop model evaluation,[43] they provide a valuable framework for assessing the quality of topic models.

One commonly employed approach for evaluating topic models is coherence, which assesses the level of semantic similarity among the most relevant words within a topic. This evaluation metric involves calculating pairwise scores for each topic's top n frequently occurring words. These scores are combined to derive the final coherence score,[47] as in Equation (2).

$$Coherence = \sum_{i<j} score(w_i, w_j) \quad (2)$$

The literature presents several coherence measures for evaluating topic models. For example, Newman *et al.*[48] introduced the UCI (or CV) measure as an automatic coherence metric for assessing topic understandability. This measurement compares word pairs and treats words as individual facts. Other researchers have also proposed coherence measures based on word statistics, as demonstrated by Stevens *et al.*,[28] Lau *et al.*,[42] and Mimno *et al.*[49]

To assess the effectiveness of topic mining, it is important to establish metrics such as perplexity and coherence. Evaluating the interpretability of the model is equally crucial for its practical application. Coherence, which focuses explicitly on the interpretability aspect, serves as a fundamental metric for evaluating topic mining. Higher coherence values indicate better model interpretability, making coherence the preferred metric for assessing topic mining. Cao *et al.*[50] suggested measuring the average cosine distance between each pair of topics to gain insights into the stability of the topic structure. This analysis can help identify any inconsistencies or overlaps in the topics, thereby enhancing the model's interpretability and effectiveness.

To determine the optimal number of topics for our study, we utilized the Coherence Model class in the gensim package, which calculates fitness scores for different topic numbers. Additionally, we employed the C_v and C_umass algorithms to compute coherence scores and identify the number of topics associated with the highest score. A higher C_v value signifies a better fit for the model. Subsequently, we created a line graph using Python's Matplotlib package to visually represent the results. We observed that the coherence score reached its peak at four topics (refer to Figure 1), and utilized the coherence parameter to streamline the model. Furthermore, we conducted experiments by varying the number of topics from 2 to 6, with a step size of 1, to identify the optimal number of topics. The findings indicated that the coherence parameter attained its highest value of -0.0641 when the number of topics was set to six (see Figure 1). Consequently, we selected six topics with automatically set values of α = 0.134 and η = 0.134.

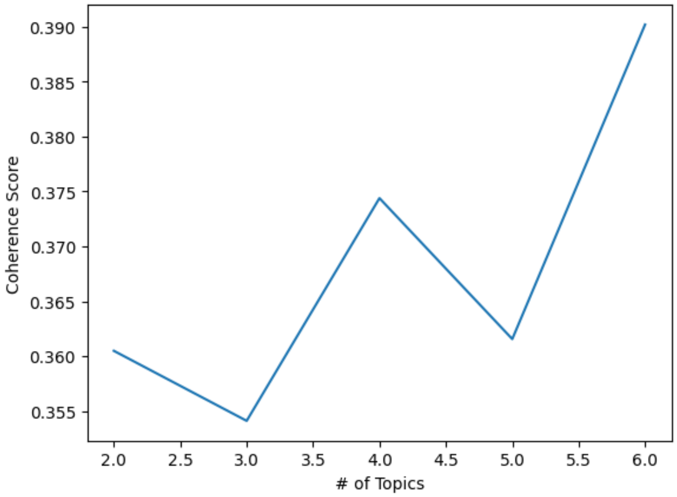


Figure 2: Topics number and coherence.

We refined the model by fine-tuning the auto-set value within a narrow range, and conducted tests on the coherence parameter to attain the best possible outcomes. The coherence score, measuring model performance, peaked at 0.4415 when α = 0.14 and η = 0.17, surpassing the previous score of 0.3590. These results were then scrutinized and validated by experts, who confirmed the clarity and interpretability of the six topics. Consequently, we determined the optimal parameters and obtained the optimal model results for the six topics, as illustrated in Figure 2.

It can be seen that the coherence score is high when there are four topics and decreases to five topics and then highest at six topics to be extracted. Hence, we categorized the collected articles into six distinct topics. The associated probability values shown in Figure 3 serve as a guide when selecting topics for inclusion using numerical markers. Upon presenting the topic visualization, as shown in Figure 4, we observed an overlap between Topic 2 and Topic 5. Therefore, we decided to adjust the number of topics to four and presented the revised topic visualization in Figure 5. However, we noticed that there remained an overlap between Topic 2 and Topic 3. As a result, we further adjusted the number of topics to three, then presented the updated topic visualization in Figure 6. The results revealed that there was no overlap among all three

Topic naming and topic details

In traditional topic modeling, topics are commonly assigned names based on the probability of topic words, with high-probability words being employed for labeling. Nevertheless, this approach often leads to a need for more specificity and discrimination in topic naming. Therefore, Carson Sievert[29] suggested a novel approach that involves incorporating keywords and their relationship to the topic for naming. Once the model is established, topic naming is carried out manually, which is crucial for ensuring the interpretability of the results. To achieve this, the correlation formula is utilized, which is outlined as in Equation (3).

$$\gamma(w, t|\lambda) = \lambda \log[p(w|t)] + (1-\lambda)\, log \left[\frac{p(w|t)}{p(w)}\right] \quad (3)$$

The value of λ, which determines the weight assigned to a topic word w in relation to its boost under topic t, was adjusted. Through the process of fine-tuning, we determined that setting λ to 0.6 yielded the optimal results. These findings are summarized in Table 1, offering a comprehensive overview of the resulting topics.

To assign names to each topic, we employed three distinct methods: (a) analyzing the top 10 words that exhibited the highest term-topic probability (βk) and frequency in the abstract, thus representing the topic most prominently; (b) generating a word cloud for each topic using the top 50 words, where the size of a term corresponds to its term-topic probability, aiding in the identification of the most representative terms within each topic; and (c) analyzing the top 20 articles (θd) with the highest proportion of words, enabling a deeper understanding of the narratives associated with each topic and facilitating the decision-making process for topic names. The resulting topics, along with their respective names, are presented in Table 1.

The Latent Dirichlet Allocation (LDA) method is widely utilized in the division of digital economy research into distinct topics. This process involves several essential steps, outlined as follows:

Determining the optimal number of topics by assessing the coherence score.

Assigning the determined number of topics to the respective digital economy research.

Employing a set of words to encapsulate the characteristics inherent to each topic.

Utilizing the resulting representation as a reference for the title of the digital economy research upon topic assignment.

Employing numerical markers (0, 1, and 2) as indicators for entry into the topic.

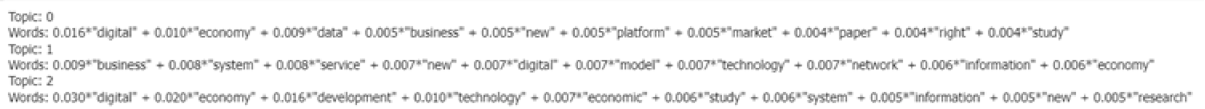
In order to ascertain the themes that encapsulate each formative topic, the process involves identifying the most prominent set of words associated with each topic. From this list of dominant words, a theme is determined, which serves as the representative name for the topic, often expressed through multiple words or sentences (as displayed in Table 1).

The information is summarized in Table 1, which includes the topic names, top 20 words, and a representative article for each topic. Regarding this, topic 0 (T0) was named "Digitalization." The articles in T0 represent the interrelationships among concepts in the digitalization of the digital economy, which refers to integrating advanced digital technologies and systems into various sectors and aspects of the economy. It involves the transformation of traditional business models, processes, and transactions into digital formats, enabling enhanced efficiency, innovation, and connectivity in economic activities. Digitalization encompasses a wide range of areas within the digital economy, including e-commerce, digital marketing, online platforms, financial technology (Fintech), artificial intelligence (AI), big data analytics, cloud computing, and Internet of Things (IoT). These technologies enable businesses to streamline operations, reach broader markets, offer personalized services, and gather valuable insights from data.

One of the key drivers of digitalization in the digital economy is the increasing accessibility and affordability of digital technologies and internet connectivity. This has led to a significant shift in consumer behavior, with a growing reliance on online platforms and digital services for various needs such as shopping, communication, entertainment, and financial transactions. The digitalization of the digital economy has profound implications for businesses, governments, and society as a whole. It presents opportunities for economic growth, job creation, and innovation, while also posing challenges such as cybersecurity risks, data privacy concerns, and the digital divide among different populations. Moreover, the digitalization of the digital economy has reshaped traditional industries, creating new business models and disrupting established market dynamics. It has enabled the rise of digital-native companies and platforms that have revolutionized industries such as retail, transportation, hospitality, and media. Overall, the digitalization of the digital economy represents a transformative shift in how economic activities are conducted, leveraging the power of digital technologies to drive economic development, improve productivity, and foster inclusive growth.[51]

Topic 1 (T1) was labeled "Data Governance." The articles in T1 examined the relationships between the co-words, which refer to the framework and processes to ensure the effective management, protection, and utilization of data assets within the digital realm. As the digital economy relies heavily on data as a valuable resource, proper governance practices are essential to maintain trust, privacy, and security while maximizing the potential benefits of data-driven activities. Data governance involves establishing policies, procedures, and guidelines to govern the collection, storage, access, sharing, and usage of data in the digital economy. It encompasses various dimensions, including data quality, data integrity, data privacy, data security, data ethics, and regulatory compliance. One of the primary objectives of data governance is to ensure data quality and reliability. This involves implementing processes to verify the accuracy, completeness, and consistency of data, enabling stakeholders to make informed decisions based on reliable information.

Data governance frameworks also define roles and responsibilities for data stewardship, outlining the individuals or teams responsible for data management and ensuring adherence to governance principles. Data privacy and security are crucial



```
Topic: 0
Words: 0.016*"digital" + 0.010*"economy" + 0.009*"data" + 0.005*"business" + 0.005*"new" + 0.005*"platform" + 0.005*"market" + 0.004*"paper" + 0.004*"right" + 0.004*"study"
Topic: 1
Words: 0.009*"business" + 0.008*"system" + 0.008*"service" + 0.007*"new" + 0.007*"digital" + 0.007*"model" + 0.007*"technology" + 0.007*"network" + 0.006*"information" + 0.006*"economy"
Topic: 2
Words: 0.030*"digital" + 0.020*"economy" + 0.016*"development" + 0.010*"technology" + 0.007*"economic" + 0.006*"study" + 0.006*"system" + 0.005*"information" + 0.005*"new" + 0.005*"research"
```

Figure 3: A list of unique word that represent a formed topic.

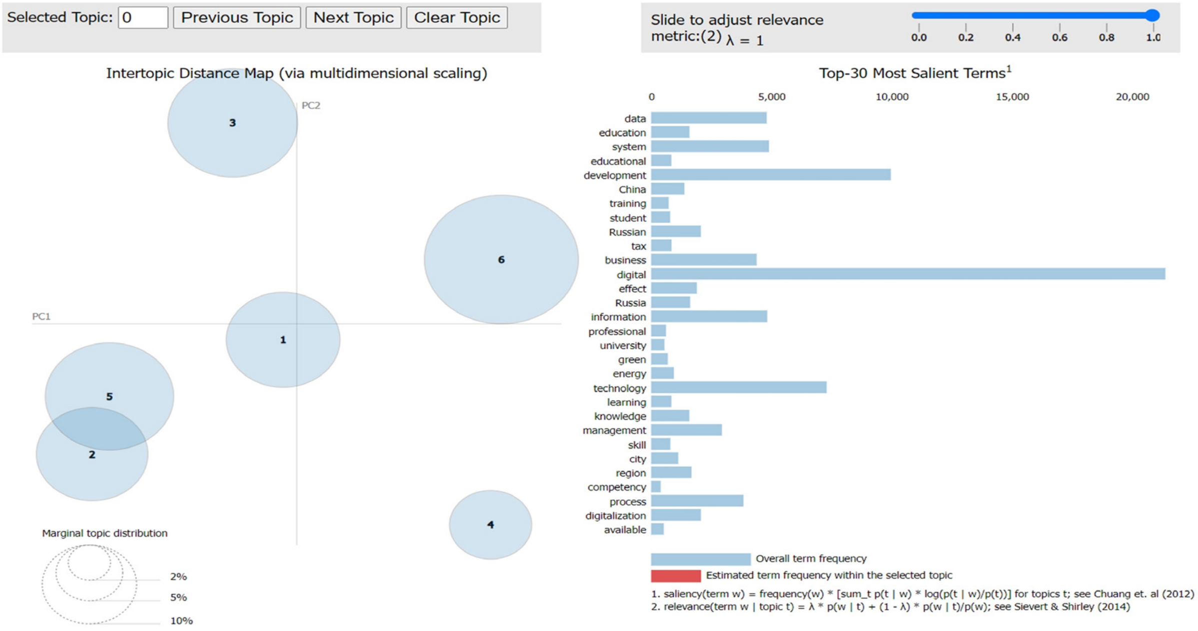
aspects of data governance in the digital economy. Organizations must establish robust measures to protect sensitive and personal data from unauthorized access, breaches, or misuse. This involves implementing encryption, access controls, authentication mechanisms, and data anonymization techniques to safeguard data privacy and mitigate potential risks. Ethical considerations are another critical component of data governance in the digital economy. It involves defining ethical guidelines for data collection, usage, and sharing, particularly when dealing with sensitive or personally identifiable information. Organizations need to adhere to ethical standards, ensuring transparency, fairness, and accountability in their data practices, while respecting the rights and expectations of individuals. Compliance with relevant regulations and legal frameworks is an essential aspect of data governance in the digital economy. Organizations must navigate various laws such as data protection regulations, industry-specific regulations, and international data transfer regulations.[21,23,24]

Table 1: Digital economy analysis topics.

Topic No.	Topic name	Representative unigram
0	Digitalization	Digital, economy, technology, business, new, service, model, information, paper, system, research, study, development, consumer, network, data, need, industry, Internet, application.
1	Data Governance	Digital, data, economy, new, market, technology, right, platform, paper, social, service, legal, tax, law, article, policy, research, economic.
2	Digital Transformation	Digital, economy, development, technology, economic, study, system, process, model, information, new, analysis, management, research, business, innovation, transformation, level, country, result.

Data governance frameworks provide guidance on compliance requirements, helping organizations meet their legal obligations and avoid penalties or reputational damage. Effective data governance in the digital economy enables organizations to harness the full potential of data. It facilitates data-driven decision-making, enhances operational efficiency, enables innovation, and fosters collaboration between stakeholders. By establishing clear data governance frameworks, organizations can build trust with customers, partners, and regulators, creating a foundation for sustainable growth and responsible data practices. In summary, data governance in the digital economy is a multidimensional approach to managing data assets effectively. It encompasses various aspects such as data quality, privacy, security, ethics, and compliance. By implementing robust data governance practices, organizations can navigate the complexities of the digital economy while ensuring responsible and beneficial use of data resources.[20,22]

Topic 2 (T2) was labeled as "Digital Transformation." Articles in T2 refer to the comprehensive and strategic integration of digital technologies, processes, and mindsets into various aspects of businesses and organizations. It involves reimagining traditional business models, operations, and customer experiences to leverage the potential of digital technologies and unlock new opportunities for growth, innovation, and competitiveness. At its core, digital transformation involves adopting and leveraging



Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4: The Interactive visualization of the LDA model has six topics. URL: https://www.huso-kku.org/de/lda_de6.html.

technologies such as cloud computing, Artificial Intelligence (AI), data analytics, Internet of Things (IoT), robotics, and automation. These technologies enable organizations to streamline operations, enhance efficiency, optimize resource utilization, and deliver personalized and seamless digital experiences to customers. Digital transformation encompasses multiple dimensions within the digital economy. It includes digitizing and automating existing processes to improve productivity, reduce costs, and eliminate inefficiencies. This may involve transitioning from manual or paper-based processes to digital workflows, enabling real-time data capture and analysis.

Additionally, digital transformation involves leveraging data as a strategic asset. Organizations harness data analytics and AI technologies to gain valuable insights from vast amounts of data, enabling data-driven decision-making, predictive analytics, and personalized customer experiences. This data-centric approach helps organizations identify trends, understand customer behaviors, and identify new business opportunities. Furthermore, digital transformation focuses on enhancing customer experiences

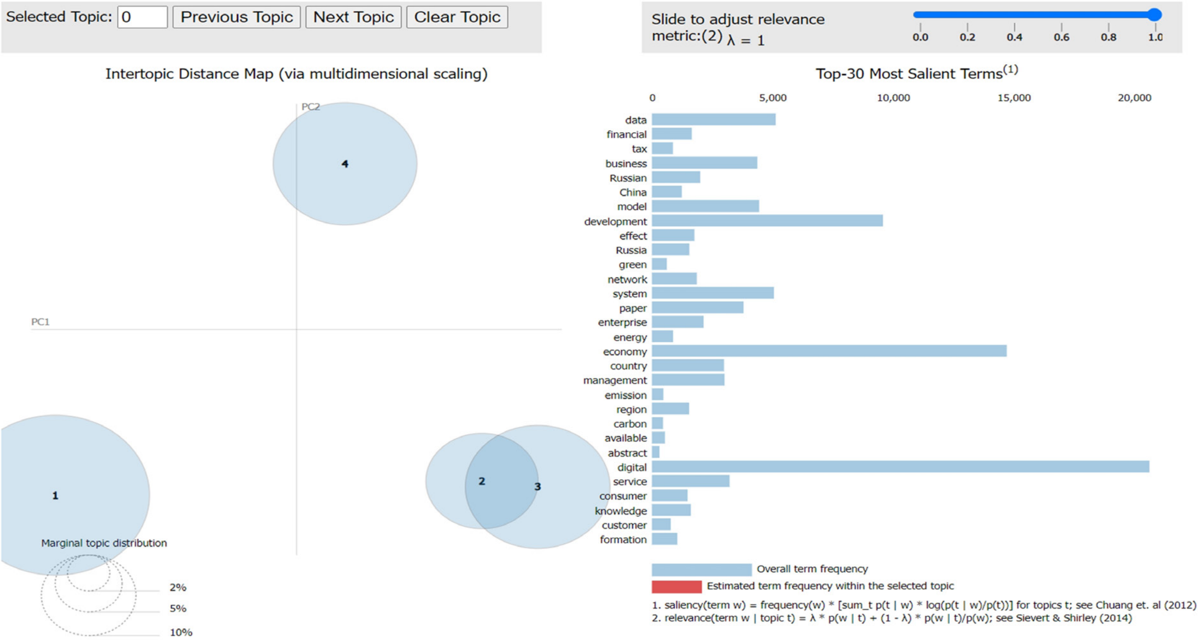

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2) λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms(1)
0
5,000
10,000
15,000
20,000
data
financial
tax
business
Russian
China
model
development
effect
Russia
green
network
system
paper
enterprise
energy
economy
country
management
emission
region
carbon
available
abstract
digital
service
consumer
knowledge
customer
formation
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5: The Interactive visualization of the LDA model has four topics. URL: https://www.huso-kku.org/de/lda_de4.html.

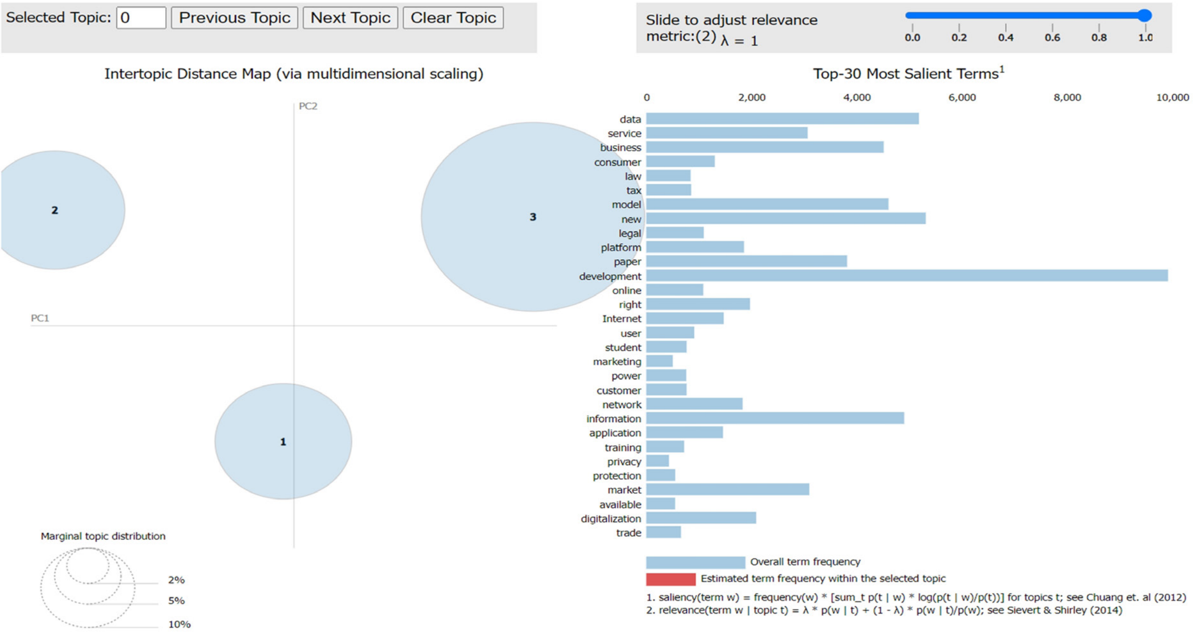

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2) λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
3
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0
2,000
4,000
6,000
8,000
10,000
data
service
business
consumer
law
tax
model
new
legal
platform
paper
development
online
right
Internet
user
student
marketing
power
customer
network
information
application
training
privacy
protection
market
available
digitalization
trade
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6: The Interactive visualization of the LDA model has three topics. URL: https://www.huso-kku.org/de/lda_de3.html.

by offering seamless and personalized interactions across digital channels. This includes developing user-friendly interfaces, optimizing websites and mobile applications, and implementing digital marketing strategies to reach and engage customers effectively. By embracing digital technologies, organizations can tailor their products, services, and communications to meet the evolving needs and expectations of customers. Digital transformation also encompasses changes in organizational culture and capabilities. It involves fostering a digital mindset, promoting innovation, and encouraging collaboration and agility. Organizations need to embrace a culture of continuous learning and adaptation to thrive in the digital economy.[8,10,17]

The benefits of digital transformation in the digital economy are manifold. It enables organizations to gain a competitive edge, adapt to changing market dynamics, and seize new business opportunities. By embracing digital technologies, organizations can enhance operational efficiency, optimize resource allocation, and drive revenue growth. However, digital transformation also poses challenges. It requires significant investments in technology infrastructure, talent acquisition and development, and change management. Organizations must address issues such as cybersecurity risks, data privacy concerns, and the need for ongoing training and upskilling. In conclusion, digital transformation in the digital economy represents a paradigm shift in how businesses and organizations operate, compete, and interact with customers. It involves harnessing digital technologies, data analytics, and a digital mindset to drive innovation, efficiency, and customer-centricity. By embracing digital transformation, organizations can navigate the complexities of the digital economy, remain agile, and capitalize on the immense opportunities presented by the digital era.[18,19]

Topic visualization

An interactive visual diagram proves highly effective in presenting the outcomes of a topic model. To this end, we employed the pyLDAVis package[29] to generate an interactive diagram showcasing the topics and their most representative words (Figure 6). The size of each bubble within the diagram corresponds to the topic's relevance within the corpus, while proximity between bubbles indicates their similarity. Notably, the pyLDAvis visualization method offers a significant advantage by allowing users to adjust the relevance of words within a topic using a slider.[29,52]

The clear differentiation between topics and their well-balanced popularity suggests the model's robustness and accuracy. Moreover, the multi-dimensional zoomed model view provides valuable insights into each topic's meaning, popularity, and relationships, facilitating interpretation and analysis. Each topic is succinctly summarized through a histogram displaying the top 30 relevant words, along with saliency and overall term frequency. Furthermore, the entire model is accessible via the World Wide Web, allowing readers to explore and utilize the topic model through an interactive interface. Ultimately, the pyLDAVis package introduces an innovative and effective approach for visualizing and interpreting topic models, empowering researchers to gain deeper insights into the structure and content of their data.

This tool provides a concise and user-friendly visualization that effectively demonstrates the relationships and significance of each topic. It accomplishes this by presenting the constituent words of each topic through a combination of a circle and a horizontal bar chart. The circle on the left panel offers a comprehensive overview of the entire model, enabling users to grasp the interconnections between topics and their strengths. Simultaneously, the right panel showcases a horizontal bar chart presenting the specific terms comprising each topic, furnishing users with a comprehensive understanding of the individual topics. Overall, this visualization tool offers a clear and intuitive representation, facilitating users to comprehend each topic's intricate relationships and relative importance.

The evident distinctiveness of the three identified topics, each pertaining to distinct research areas, is illustrated in Figure 6. Clicking on each topic circle within the tool generates a bar graph presenting the top 30 most relevant terms associated with that topic. This feature enables users to swiftly and succinctly grasp the topic's relevance by examining its most significant keywords. By conducting a lexical analysis of these keywords, it becomes feasible to categorize the three topics, as illustrated in Table 1.

DISCUSSION

This study analyzed bibliometric data on digital economy research, encompassing four distinct aspects: topic focus, characteristics, category, and key modes. To accomplish this analysis, we employed Python libraries for topic modeling and evaluated the performance of each model using metrics such as perplexity and coherence. Furthermore, we utilized Python libraries to create compelling visualizations that communicated our findings effectively. By harnessing the capabilities of these libraries, we successfully generated clear and informative visual representations that showcased the significant insights and trends derived from our analysis.

In this study, we conducted an in-depth analysis of global research trends and the contextual landscape of digital economy research. We used text analysis and topic modeling techniques to examine 8,321 documents from the Scopus database. Our rigorous analysis successfully identified three distinct topics and traced their development trends, shedding light on the shifting focus within topic development. Among the identified topics, "Digital Transformation" emerged as the most popular, accounting for 56.6% of tokens. On the other hand, the "Digitalization" topic exhibited relatively lower popularity, representing only 21.2% of tokens. The result of this study is different from previous study of Deng *et al*.,[36] who indicated that "digital economic innovation" are the common topic in extracted literature. Meanwhile, Lee *et al*.[37] focused on common topics of 'smart factory, sustainability and product-service systems, construction digital transformation, public infrastructure-centric digital transformation, techno-centric digital transformation, and business modelcentric digital transformation'. In addition, the study of Shan *et al*.[53] indicated that to advance the progression of AI and foster the expansion of the digital economy are outlined with a focus on industrial fusion, ecological frameworks, and self-driven innovation.

The study of Bukht and Heeks[54] indicated that the core of the digital economy lies within the IT/ICT sector, which is responsible for producing fundamental digital goods and services. Expanding beyond this core, the true digital economy encompasses emerging digital and platform services characterized by business models primarily based on digital goods or services. Lastly, the broadest scope, the digitalized economy, encompasses the pervasive use of ICTs across all economic fields. This classification aligns with previous research findings and provides a valuable framework for comprehending digital economy research.

Significantly, this study contributes both theoretically and methodologically by employing LDA topic modeling technology within text analytics. Through large-scale text analysis of 8,321 documents from the Scopus database, we identified research topics and their development trends for the first time. By utilizing machine learning training to derive core parameters, such as the number of topics, our study offers a valuable reference for future research endeavors in digital economy research. Overall, our study enhances the understanding of global research trends in digital economy research and makes substantial contributions to the academic community by employing advanced techniques to unravel topic development and providing a comprehensive framework for the field's exploration and analysis.

Topic modeling offers researchers a valuable opportunity to uncover emerging research areas and trends, providing insights into the impact of specific publications or authors on their respective fields.[55] Moreover, it addresses certain limitations of traditional bibliometric analysis, such as the excessive reliance on citation counts as a sole measure of significance.[55] This study makes a noteworthy contribution to the field by emphasizing the significance of analyzing digitalization, data governance, and digital transformation. It also highlights the efficacy of Latent Dirichlet Allocation (LDA) as a powerful tool for efficient and accurate text analytics. The findings of this research hold particular value for practitioners, policymakers, and scholars engaged in information science, data mining, econometrics, and bibliometrics, as they shed light on crucial aspects of these areas and provide a foundation for further investigation.

CONCLUSION

In recent years, topic modeling has emerged as a powerful technique for discovering hidden topics within large collections of texts. With the increase in online activities of academics, businesses, and the general public, there has been a surge in interest in this field. This article explores the major topic modeling algorithms and their applications, with a specific focus on the Python programming language libraries and tools that can assist researchers in implementing their ideas. By using topic modeling, individuals can update and develop their interests in different aspects, which could lead to new research questions and ideas. Thus, the findings of this paper are useful for policymaker to issue policies and strategies for developing digital economy; and organizations to apply appropriate technologies and plans to boost their orgainizations. Furthermore, the research findings are particularly support to scholars in relevant fields to offer a foundational understanding of topic modeling, facilitating further investigation and exploration in these fields.

This study has revealed two primary limitations. Firstly, the generalizability of the findings may be constrained by the utilization of a limited dataset extracted solely from Scopus. Thus, expanding the data sources to include repositories like the Web of Science core could enhance the broader applicability of the results. Secondly, the analysis was restricted to abstracts rather than full texts, potentially limiting the findings' level of detail and granularity. Therefore, future research should investigate authoritative articles in multiple languages, considering the full text, to attain more comprehensive and nuanced outcomes.

ACKNOWLEDGEMENT

We sincerely thank Khon Kaen University, Walailak University, Thammasat University, and Burapha University for their invaluable support throughout this research endeavor. Providing researchers from these esteemed institutions was pivotal in completing this project. We are profoundly grateful for their unwavering assistance, without which this work would not have been achievable.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Brynjolfsson A, McAfee A. The second machine age: Work, progress, and prosperity in a time of brilliant technologies.: W. W. Norton & Company; 2014.
2. OECD. OECD Digital Economy Outlook 2020: Organisation for Economic Co-operation and Development; 2020.
3. UNCTAD. Digital Economy Report 2021: Data for Inclusive Growth. In United Nations Conference on Trade and Development; 2021: United Nations Conference on Trade and Development.
4. World Bank. World Development Report 2019: The changing nature of work: World Bank; 2019.
5. Howkins J. The creative economy: How people make money from ideas.: 2002; Penguin UK.

6. Schwab K. The Fourth Industrial Revolution: What It Means and How to Respond?: Foreign Affairs; 2015.
7. Sala-i-Martín X. The global competitiveness report 2016–2017.: World Economic Forum; 2016.
8. Bresciani S, Ferraris A, M. R, G. S. Building a digital transformation strategy. Digital transformation management for agile organizations: A compass to sail the digital world. 2021:5-27.
9. Luca M, Richard D, Andrea Di S, Alvaro E. Digital transformation of production governance and assurance process for improving production efficiency. In International Petroleum Exhibition and Conference; 2021; Abu Dhabi: SPE.
10. Warner KS, Wäger M. Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal. Long range planning. 2019;52(3):326-349.
11. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, *et al.* Big data: The next frontier for innovation, competition, and productivity: McKinsey Global Institute; 2011.
12. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big data: the management revolution. Harvard business review. 2012;90(10):60-68.
13. Porter ME, Heppelmann JE. How smart, connected products are transforming competition. Harvard business review. 2014;92(11):64-88.
14. Joachimsthaler E, Chaudhuri A, Kalthoff M, Burgess-Webb A, Bharadwaj A. How smart, connected products are transforming competition. Harvard business review. 2015;93(1):4.
15. Tapscott D, Williams AD. Innovating the 21st-Century Enterprise. MIT Sloan Management Review. 2012;25(2):23-29.
16. World Economic Forum. The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution. In Global Challenge Insight Report.: World Economic Forum; 2016.
17. David R, Marinai L, Di Sarra A, Escorcia A, Akhtar MMJ, Al-Jefri A, *et al.* Digital transformation of production governance and assurance process for improving production efficiency. In Abu Dhabi International Petroleum Exhibition and Conference (p. D012S116R061); 2020; Abu Dhabi.
18. Schneider S, Kokshagina O. Digital transformation: What we have learned (thus far) and what is next. Creativity and Innovation Management. 2021;30:384-411.
19. Zaki M. Digital transformation: harnessing digital technologies for the next generation of services. Journal of Services Marketing. 2019;33(4):429-35.
20. Jouanjean M, Casalini F, Wiseman L, Gray E. Issues around data governance in the digital transformation of agriculture: The farmers' perspective. OECD Food, Agriculture and Fisheries Papers. 2020;(146).
21. Kitchin R. Getting smarter about smart cities: Improving data privacy and data security. [Online].; 2016. Available from: https://mural.maynoothuniversity.ie/7242/1/Smart.
22. Lei T, Tang Y. Digital Governance Model for Big Data Era-Based on Typical Practices in Singapore. Humanities and Social Sciences. 2019;7(2):76-82.
23. Shukla S, J. P. G, Tiwari K, Kureethara JV. Data Security. In Data Ethics and Challenges. Singapore: Springer Singapore; 2022. p. 41-59.
24. Wong DH, Maarop N, Samy GN. Data Governance and Data Stewardship: A Success Procedure. In 8th International Conference on Information Technology and Multimedia (ICIMU); 2020.
25. Blei DM, Andrew YN, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003:993-1022.
26. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences. 2004;101(Supplement 1):5228-35.
27. Blei DM. Probabilistic Topic Models. Communications of the ACM. 2012;55(4):77-84.
28. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In Proceedings of the 2011 conference on empirical methods in natural language processing; 2011. p. 262-272.
29. Carson S, Kenneth S. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces; 2014; Baltimore, Maryland, USA. p. 63-70.
30. Ramage D, Hall D, Nallapati R, Manning CD. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing; 2009; Singapore: Association for Computational Linguistics. p. 248-256.
31. Wang C, Blei DM, Li FF. Simultaneous Image Classification and Annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2009.
32. Meimaris M, Papadopoulos S, Y. M. A Survey on Topic Modeling in Text Mining. WIREs Data Mining and Knowledge Discovery. 2014;4(276-292):4.
33. Chen H, Zhang Y, Jin Q, Wang X. Exploring Patterns of Academic-Industrial Collaboration for Digital Transformation Research: A Bibliometric-Enhanced Topic Modeling Method. In 2022 Portland International Conference on Management of Engineering and Technology (PICMET); 2022: IEEE. p. 1-9.
34. Cheng X, Zhang S, Fu S, Liu W, Guan C, Mou J, *et al.* Exploring the metaverse in the digital economy: an overview and research framework. Journal of Electronic Business & Digital Economics. 2022; 1(1/2):206-24.
35. Dana LP, Crocco E, Culasso F, Giacosa E. Mapping the field of digital entrepreneurship: a topic modeling approach. International Entrepreneurship and Management Journal. 2023:1-35.
36. Deng G, Shen Y, Xu C. Research on the Topic Evolution of Digital Economy Based on LDA. In Proceedings of the 5th International Conference on Information Management and Management Science; 2022. p. 252-256.
37. Lee CH, Liu CL, Trappey AJ, Mo JP, Desouza KC. Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery. Advanced Engineering Informatics. 2021;50:101428.
38. Ali M. PyCaret: An open source, low-code machine learning library in Python.: PyCaret version 2; 2020.
39. Bettina G, Kurt H. Topic models: an R package for fitting topic models. Journal of Statistical Software. 2022;40(13):1-30.
40. Chowdhury CR, Bhuyan P. Information retrieval using fuzzy c-means clustering and modified vector space model. In 2010 3rd International Conference on Computer Science and Information Technology; 2010: IEEE. p. 696-700.
41. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, *et al.* Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications. 2019;78:15169-211.
42. Stevens K, Kegelmeyer P, Andrzejewski D, D. B. Exploring topic coherence over many models and many topics. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning; 2012. p. 952-961.
43. Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D. Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems. 2009;22:1-9.
44. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining. 2015:399-408.
45. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. [Online].; 2010. Available from: https://www.fi.muni.cz/usr/sojka/papers/lrec2010-rehurek-sojka.pdf.
46. Zhu B, Zheng X, Liu H, Li J, Wang P. Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. Chaos, Solitons & Fractals. 2020;140:110123.
47. Bovens L, Hartmann S. Solving the riddle of coherence. Mind. 2003;112(448):601-33.
48. Newman D, Lau JH, Grieser K, Baldwin T. Automatic evaluation of topic coherence. In The 2010 annual conference of the North American chapter of the association for computational linguistics; 2010. p. 100-108.
49. Lau JH, Newman D, Baldwin T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics; 2014. p. 530-539.
50. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. Neurocomputing. 2009;72(7-9):1775-1781.
51. Lu J, Zhou S, Xiao X, Zhong M, Zhao Y. The Dynamic Evolution of the Digital Economy and Its Impact on the Urban Green Innovation Development from the Perspective of Driving Force—Taking China's Yangtze River Economic Belt Cities as an Example. Sustainability. 2023;15(8):6989.
52. Chuang J, Manning CD, Heer J. Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces; 2012. p. 74-77.
53. Shan C, Wang J, Zhu Y. The Evolution of Artificial Intelligence in the Digital Economy: An Application of the Potential Dirichlet Allocation Model. Sustainability. 2023;15(2):1360.
54. Bukht R, Heeks R. Defining, conceptualising and measuring the digital economy. Development Informatics working paper. 2017;68:1-24.
55. Ninkov A, Frank JR, Maggio LA. Bibliometrics: Methods for studying academic publishing. Perspectives on medical education. 2022;11(3):173-6.

Cite this article: Detthamrong U, Nguyen LT, Jaroenruen Y, Takhom A, Chaichuay V, Chotchantarakun K, et al.Topic Modeling Analytics of Digital Economy Research: Trends and Insights. J Scientometric Res. 2024;13(2):448-58.