IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1863~1875
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1863-1875

Research themes and trends in the field of blockchain engineering: a topic modelling analysis

Dinara Zhaisanova[1,2], Madina Mansurova[1]
[1]Department of Artificial Intelligence and Big Data, Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]King's College of London, King's-Bolashaq Bespoke Programme, London, United Kingdom

Article Info

Article history:

Received May 28, 2025
Revised Jan 18, 2026
Accepted Feb 6, 2026

Keywords:

Blockchain development
Blockchain engineering
Latent Dirichlet allocation
Natural language processing
Topic modeling

ABSTRACT

This study employed topic modeling to identify key research themes in blockchain engineering and examined how these themes have evolved over time. The dataset of collected abstracts from 3,665 relevant papers of Web of Science (WoS) core collection for the period from 2019 to 2024 was analyzed with latent Dirichlet allocation (LDA) approach. Based on the results of the topic development trends analysis, the topics collectively highlight the evolving landscape of technologies such as blockchain, smart contracts, the internet of things (IoT), and edge computing, focusing on their integration and impact across sectors like finance, healthcare, supply chain management, and energy systems. It offers valuable insights and implications for research related to blockchain engineering. Latent semantic indexing (LSI) provided further understanding by highlighting strong connections between specific topics, such as energy trading, supply chains, and medical applications. A comparison of LDA and LSI topics revealed overlapping themes, which supports the reliability of the topic structure identified by LDA.

Corresponding Author:

Dinara Zhaisanova
Department of Artificial Intelligence and Big Data, Faculty of Information Technology
Al-Farabi Kazakh National University
71 Al-Farabi ave., Bostandyk District, Almaty-050040, Kazakhstan
Email: zhaisanova15@gmail.com

1. INTRODUCTION

Blockchain technology has emerged as a revolutionary force across various industries, offering decentralized, secure, and transparent solutions for complex problems. Across multiple industries, blockchain supports peer-to-peer exchange of digital assets—such as currencies, securities, votes, shares, and commodities—enables transparent tracking of data related to financial assets and physical goods, and automates the execution and administration of various types of contracts, including insurance agreements and programmable payment systems [1]–[4]. As the field of blockchain engineering continues to evolve rapidly, it has become increasingly important to understand the underlying trends and latent topics driving its development. Identifying these trends not only helps in tracking the progression of the technology but also in anticipating future directions and challenges in the domain.

Latent Dirichlet allocation (LDA) topic modelling has demonstrated its effectiveness as a tool for uncovering hidden structures within large text corpora. Topic modelling [5] is a method used in text classification to identify the underlying topics within a text corpus, with LDA being the most widely used technique. Recent studies have demonstrated the versatility of LDA in diverse fields, including agriculture and food industry [6], pedagogy and knowledge [7], transportation [8], and AI application [9].

In the context of blockchain engineering, where the volume of academic publications, technical reports, and industry white papers is rapidly expanding, LDA offers a methodologically sound approach to distilling complex information into coherent topics. This is particularly relevant as the field of blockchain evolves rapidly, requiring systematic methods to track its multifaceted development. Recent applications of LDA in technology domains have shown its potential for identifying trends and knowledge gaps, making it an ideal tool for analyzing blockchain literature [10], [11]. In this way, Sharma *et al.* [12] identified current trends in blockchain technology to guide future research by collecting data from IEEE, Springer, ACM, and other digital databases, applying LDA for topic modeling, and analyzing the model's outcomes through key terms and documents associated with each topic. Lee *et al.* [13] analyzed scholarly publications on blockchain to forecast emerging industries with a high potential for blockchain adoption. Their study employed LDA and dynamic topic modeling to process large-scale textual data, effectively reducing dimensionality and extracting insights from the underlying knowledge structure of the literature.

This paper aims to apply LDA topic modelling to a comprehensive dataset of blockchain engineering literature to identify and analyze the latent topics and emerging trends within the field. By doing so, the search was made for contribution to a deeper comprehension of the present condition and prospective development of blockchain engineering, providing a valuable resource for both researchers and practitioners. Such insights are crucial not only for advancing academic research but also for guiding industry innovation in blockchain technologies.

Building upon past research, this study examines a total of 3,665 publications pertaining to blockchain engineering from the period of 2019 to 2023, as indexed in the Web of Science (WoS). By employing a combination of topic modelling techniques and topic evolution analysis, the abstract texts of these papers are analyzed to identify the key research topics and trace the developmental trends in blockchain research from a global perspective. The study is guided by two core research questions:

Q1: which topics represent the most significant focus of current research?

Q2: what emerging trends are projected to shape the future development of these topics?

To address the aforementioned research questions, the abstracts of 3,665 documents were analyzed using term frequency–inverse document frequency (TF-IDF)–based identification of keywords to construct a collection of extracted keywords. Topic evolution was assessed using the bibliometric package in R, while LDA topic modelling was employed to identify nine distinct topics within the corpus. The properties and advancement trajectories of each identified topic was subsequently examined. While several prior studies [10], [12], [13] have applied LDA topic modelling to related domains, this study is among the first to extensively apply LDA to blockchain engineering research. Notably, key parameters, including the number of topics, were determined by means of training machine learning models using perplexity and coherence metrics to optimize the model's performance. The analysis encompassed 3,665 documents from the WoS, covering the period from 2019 to 2023. By leveraging a large-scale textual dataset, this study extracted research topics and developmental trends in blockchain engineering, thereby offering both theoretical insights and methodological frameworks for future research in the domain.

Blockchain engineering is a multidisciplinary field that combines principles from computer science, cryptography, distributed systems, and economics to develop decentralized systems that ensure data integrity, security, and transparency. Since the inception of blockchain technology with Bitcoin in 2008 [14], the field has seen rapid expansion into various sectors covering finance, supply chain management, healthcare, and other application areas. Topic modelling, particularly LDA, has gained prominence as an effective method for text mining and natural language processing (NLP) tasks. LDA has been commonly employed used in various domains to uncover hidden topics within large text corpora, facilitating the understanding of thematic structures and trends over time [15], [16]. In academic research, LDA has been applied to analyze topics in fields as diverse as blended learning [17], mining analysis of educational scientific research projects [18], and systematic technology analysis and roadmap formulation within the blockchain domain [19]. These applications demonstrate the broad applicability of LDA in handling large datasets and its utility in revealing insights that are not immediately apparent.

Zimmermann *et al.* [20] addressed existing shortcomings by proposing new approaches that enhance topic coherence without increasing computational complexity and provide an objective method for selecting the optimal number of topics in a text corpus. Their approach includes the development of a more refined stop word list, a new dimensionality-reduction heuristic that assesses word importance, and an eigenvalue technique for topic determination. These methods are integrated into the Zimm approach, which demonstrates superior performance to LDA by correctly identifying the number of topics in 7 out of 9 subsets of the 20-newsgroup dataset, compared to LDA's accuracy in none of the subsets. Luo *et al.* [21] presented the innovative graph contrastive neural topic model (GCTM), which utilizes a graph-based sampling technique that leverages detailed correlations and irrelevancies between documents and words. The model conceptualizes an input document as a bipartite graph linking documents and words, constructing networks

linking both positively and negatively co-occurring words to capture sophisticated semantic associations between terms. Ozkara *et al.* [22] sought to demonstrate the effectiveness of NLP and topic modeling approaches for organizing and interpreting large-scale scholarly literature in stroke research. Their findings show that these methods can streamline literature reviews, uncover hidden thematic structures, and monitor emerging research directions, highlighting the prominence of animal models, the growing focus on rehabilitation studies, and the crucial role of reperfusion therapy. Akar and Yörük [23] used the LDA algorithm, a topic-modeling method within the field of text mining, to elucidate the principal research themes associated with psychological contract breaches and violations, emphasizing the shifts in focus over time and particularly during the COVID-19 pandemic. Alkamli and Alabduljabbar [24] adopted a data-driven methodology that combined Twitter data analysis with a user survey to investigate privacy issues associated with ChatGPT, a widely used generative AI model. By applying LDA topic modeling and data classification techniques, the study identified major privacy-related concerns, contributing to a deeper understanding of user perceptions and providing insights to support policy formulation and guide future research on privacy in generative AI.

Saqib *et al.* [25] analyzed LDA and nonnegative matrix factorization (NMF) using latent semantic indexing (LSI) to evaluate their effectiveness in opinion and text mining, concluding that while both models perform well in topic detection, NMF demonstrates a slight advantage over LDA. Horasan [26] proposed a hybrid model for collaborative recommendation systems (CRS) based on LSI, demonstrating that LSI-based user, item, and hybrid models outperform traditional Pearson correlation coefficient (PCC) models in prediction accuracy while achieving lower computational complexity through dimensionality reduction, with the hybrid model yielding the most precise recommendations. Roy *et al.* [27] investigated machine learning approaches for uncovering social and behavioral determinants of health (SBDH) from unstructured electronic health record (EHR) notes, demonstrating that LSI, applied to 2,083,180 clinical notes from the medical information mart for intensive care III (MIMIC-III) dataset, performed comparably to GPT-3.5 and GPT-4 while offering advantages such as robustness, determinism, and scalability without document-size limitations or cost constraints, making it well-suited for real-world healthcare applications.

Despite the growing interest in blockchain technology, the application of LDA topic modeling within this domain remains relatively underexplored. A few studies have begun to employ LDA to analyze blockchain-related content, such as research papers, patents, and social media discussions, to identify key trends and emerging areas of interest [28], [29]. These studies have contributed initial knowledge regarding the thematic progression of blockchain research, yet they also highlight the need for more comprehensive analyses that consider a wider range of sources and a more extensive temporal scope.

While existing research has laid the groundwork for understanding the thematic development of blockchain engineering, several gaps remain. First, comprehensive studies are currently lacking that apply LDA to a broad dataset of blockchain engineering literature, including both academic and industry sources. Second, previous studies have often focused on specific aspects of blockchain, such as security or scalability, without providing a holistic view of the field's evolution. Lastly, there is a need for analyses that can track the temporal dynamics of topics within blockchain engineering, helping to identify not only current trends but also potential future directions. This paper aims to address the gaps identified by applying LDA to a comprehensive dataset of blockchain engineering literature, providing a more detailed and dynamic understanding of the latent topics and trends shaping the field.

2. METHOD

This research performed a quantitative examination of a large body of blockchain engineering publications, focusing on abstract-level content analysis. Figure 1 presents the workflow for dataset collection and analytical procedures. The methodology was structured into three main stages: data collection and preprocessing, topic extraction using LDA, and validation of the resulting topic structure.

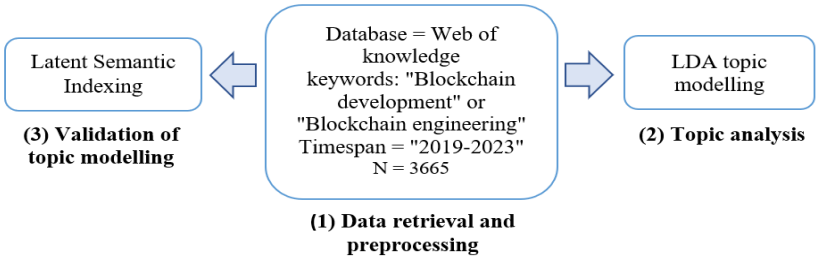

Figure 1. The flowchart depicting the methodology for dataset acquisition and analysis

2.1. Data collection and preparation

In this study, the widely recognized literature database WoS, which provides access to journals indexed in the social sciences citation index (SSCI) and the science citation index (SCI), was chosen as a comprehensive database that offers high-quality data for conducting research evaluations [30]–[32]. Ensuring high-quality literature content was vital to achieving reliable research results. The data was queried on July 31, 2024, to include all publications in the WoS database till 2023. The descriptive analysis revealed key information, including a total of 10,079 authors and an average publication age of 3.92 years. Each paper received a mean of 19.25 citations, with an average of 3.29 citations per document per year. Additionally, the percentage of international co-authorships was 33.62%.

The literature search was limited to documents categorized as "article" within the WoS. The search strategy was developed based on an understanding of blockchain development and on search terms adopted from Zhaisanova and Mansurova [33]. Specifically, the advanced search function of the WoS database was used to apply the criteria: "Topic=('Blockchain development') OR Topic=('Blockchain engineering')". Here, topic searches encompassed titles, abstracts, and author keywords relevant to blockchain engineering research published between 2019 and 2023. All publications deemed relevant to addressing the research questions were collected. Following the exclusion of non-English publications, an initial set of 3,665 articles was obtained. For each article, bibliographic information such as the title, abstract, year of publication, and journal name was extracted. The inclusion criteria comprised peer-reviewed journal articles published in English between 2019 and 2023, with a focus on blockchain development, blockchain engineering, and related applications. The exclusion criteria involved omitting non-English publications, document types other than journal articles (including conference papers, reviews, and book chapters), as well as studies not directly related to blockchain engineering.

As illustrated in Figure 2, the number of downloads for relevant literature showed a general upward trend. Figure 2 shows that research related to blockchain development expanded rapidly from 234 articles to 1,023 until 2022, reaching a peak at 2023 with 1,052 articles. This study conducted topic model mining using abstract texts as the research focus. Current research employs LDA to identify conceptual trends and emerging themes in blockchain engineering, as illustrated in Figure 3. By applying the LDA technique to a dataset comprising 3,665 articles, the research aims to extract prevailing trends and patterns within the field of blockchain engineering.

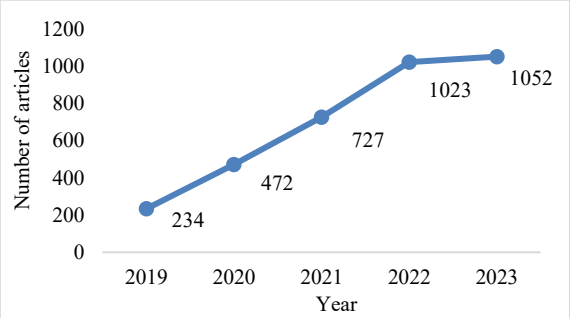

Figure 2. Annual scientific production per year (2019–2023)

It was revealed the main steps of LDA model according to the Figure 3. Prior to this, data preprocessing was required. Preprocessing included removing missing values, tokenizing the text into words, converting to lowercase, removing non-alphabetic characters and stopwords. Creating a processed version of the text data enabling diverse NLP applications, including topic modeling, sentiment analysis, and text classification. The preprocessing steps included the following:

i) Handling missing values: initially, all records with missing abstract texts were removed to maintain data integrity. This resulted in an initial selection corpus of X articles.
ii) Tokenization: the text data was tokenized into individual words using the natural language toolkit (NLTK), enabling further linguistic processing.
iii) Case normalization: all text was converted to lowercase to ensure uniformity and avoid duplication of terms due to case differences.
iv) Removal of non-alphabetic characters: punctuation marks, numerals and special symbols were eliminated, retaining only alphabetic words relevant for analysis.

v) Stopword processing: common English words excluded from analysis (e.g., "the," "is," "and") were eliminated using NLTK's predefined stopword list to enhance the focus on meaningful terms.

Following preprocessing, a refined corpus was created, consisting of Y articles after filtering and text cleaning. This corpus was then transformed into a structured format suitable for topic modeling. A dictionary was constructed using the processed tokens, mapping unique words to their respective numerical IDs. Finally, a bag-of-words representation was generated for each document, forming the final corpus for analysis. The LDA model in the provided code uses several hyperparameters: corpus, which represents the dataset in a bag-of-words format; id2word a mapping of words to unique identifiers; num_topics, setting the number of topics to extract; random_state, ensuring reproducibility; alpha, set to 'auto', allowing the model to dynamically adjust the document-topic distribution; and per_word_topics, set to true, which enables the assignment of topic distributions to individual words.

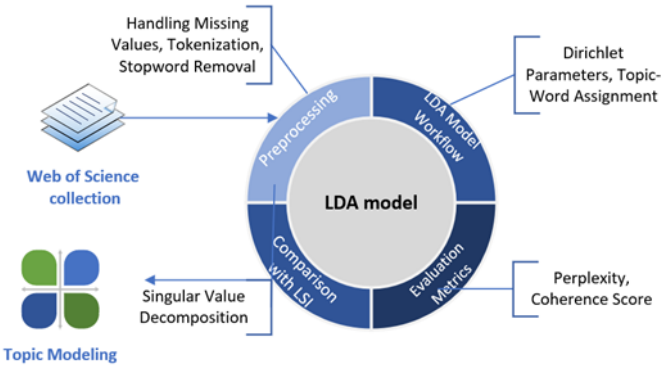

Figure 3. LDA model

2.2. Topic modeling

LDA is a probabilistic approach for identifying latent topics within a set of documents. As a generative model, it assumes that each document arises from a combination of topics, with each topic represented as a probability distribution over words.

- Let K be the number of topics.
- θ_i is the topic distribution for document i, which follows a Dirichlet distribution:
  θ_i~Dirichlet(α), where α is the hyperparameter controlling the distribution over topics.
- ϕ_k is the word distribution for topic k, which also follows a Dirichlet distribution:
  ϕ_k~Dirichlet(β), where β is the hyperparameter controlling the distribution over words for each topic.
- Document generation process: for each word w in document distribution:
  Choose a topic $z \sim$ Multinomial(θi)
  Choose a word w from p ($w|z$, ϕ)

LDA model fitting, objective function: the goal is to maximize the likelihood of the data given the parameters θ and ϕ as in (1).

$$L = p(W \mid \theta, \phi) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \sum_{k=1}^{K} p\,(w_{ij}|z_{ij} - k, \varphi k) p(k|\theta_i) \quad (1)$$

Variational Bayes or Gibbs sampling is used to estimate θ and ϕ for optimization. Metrics commonly used in topic modeling include perplexity and coherence. Perplexity assesses how well the model assigns words to specific topics by evaluating the model's likelihood value. Coherence score indicates how closely the words in a topic are related in meaning. The coherence score $C(t)$ for topic t is computed using the pairwise co-occurrence of the top words in the topic as in (2).

$$C(t) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(w_m,\ w_l)+1}{D(w_l)} \quad (2)$$

Where, $D(w_m,\ w_l)$ denotes the number of documents in which both terms w_m and w_l appear, and $D(w_l)$ refers to the count of documents that include the term w_l. LSI is another topic modeling approach that applies singular value decomposition (SVD) to achieve dimensionality reduction of the term–document matrix.

Term–document matrix: a matrix A is constructed such that each element A_{ij} corresponds to the weight (for example, a TF–IDF value) of term i in document j. SVD: the matrix A is factorized into three component matrices as in (3).

$$A=U\sum V^T \tag{3}$$

Where U is term-topic matrix, Σ is a diagonal matrix containing singular values, V is document-topic matrix.

2.3. Topic representation with LSI

The rows of U correspond to the "topic" vectors, and the columns of V correspond to the representation of documents in the reduced-dimensional space. Retain the top K singular values to approximate A with lower-rank matrices, effectively reducing the noise and focusing on the most significant "topics". Heatmap of document similarities was created by computation the cosine similarity matrix S between documents as in (4).

$$Sij = \frac{A_i * A_j}{||A_i||\,||A_j||} \tag{4}$$

The study applied the LDA model module from the well-established Gensim Python library to perform topic modeling and used a coherence metric to identify the optimal number of topics.

3. RESULTS AND DISCUSSION

Thematic evolution plot shown in Figure 4 has been generated using bibliometric, an R-based tool. This visualization is based on the "abstract" field and uses N-grams with "trigrams" to identify key research themes. The visualization, displayed as a Sankey diagram, illustrates the evolution of blockchain-related research themes from 2019 to 2024. Colored blocks represent dominant topics in each period, while connecting flows show their continuity, transformation, or decline over time.

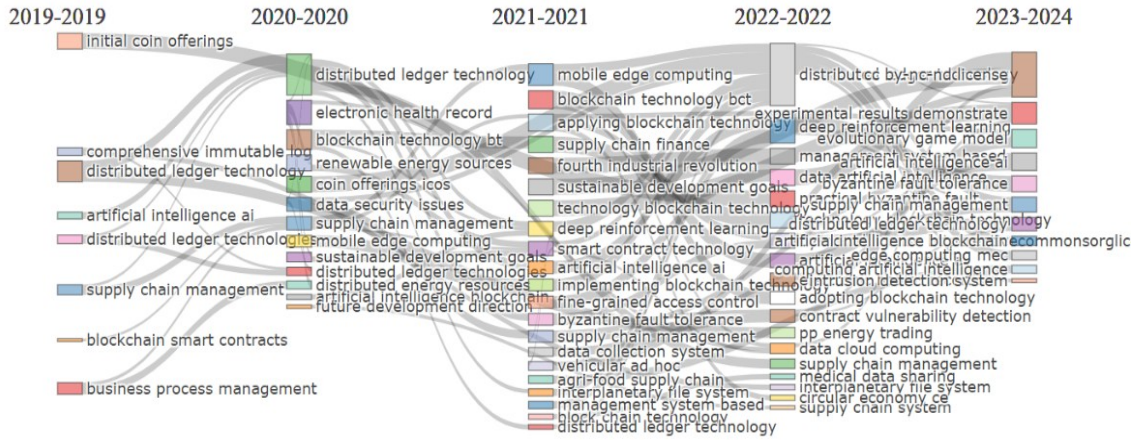

Figure 4. Thematic evolution of blockchain engineering area

In 2019, research focused on foundational themes such as initial coin offerings (ICOs), distributed ledger technology, AI, smart contracts, business process management, and supply chain applications. By 2020, the scope expanded to applied domains including EHRs, renewable energy, data security, and mobile edge computing, reflecting increased real-world adoption of blockchain. In 2021, more advanced applications emerged, including blockchain in finance, fine-grained access control, self-sovereign identity, reinforcement learning, and trust in distributed systems. The 2022 period marked a stronger emphasis on smart contracts, industrial adoption, supply chain resilience, and sustainability. In 2023-2024, blockchain research reached a mature stage, highlighting resiliency-by-design, blockchain-AI integration, threat intelligence, contract vulnerability detection, edge computing, and environmental impact assessment.

This Figure 4 effectively illustrated the dynamic evolution of blockchain-related research topics over time. The transition from basic concepts to more complex and interdisciplinary applications highlighted the increasing maturity of blockchain technology and its integration with various industries. The continuous flow of topics across different years indicates sustained interest and ongoing advancements, with emerging trends suggesting a future focus on security, AI integration, and sustainable blockchain solutions.

3.1. Determination of optimal parameters

LDA is a highly regarded topic modeling method known for it is empirical performance [30]. This unsupervised method requires predefined parameters—such as number of topics (K), topic distribution prior (α), and topic–word distribution prior (η)—which play a crucial role in determining topic quality. Topic model performance is commonly evaluated by employing measures like perplexity and coherence, with coherence emphasizing interpretability. Given the importance of interpretability for meaningful analysis and practical use, coherence was chosen as the primary evaluation criterion. The priors α and η were set automatically, and the optimal number of topics was determined based on coherence metrics. As shown in Figure 5, the model achieved its highest coherence value (0.41) when K was set to nine topics.

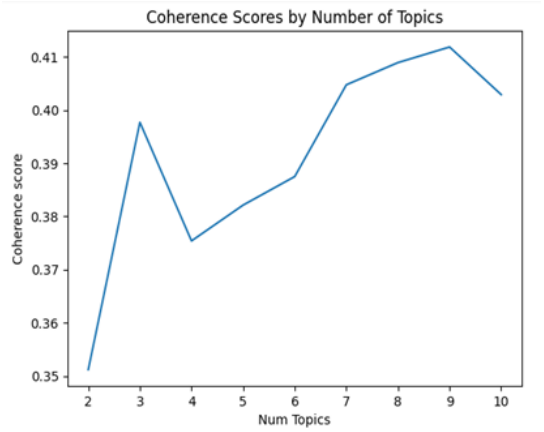

Figure 5. Topics number and coherence

Once the model was established, the topics were named manually based on the results, which significantly enhanced the interpretability of the findings. Traditionally, naming is primarily guided by the topic-word probability $p(w|t)$, with keywords exhibiting high $p(w|t)$ values chosen for topic naming. The correlation formula used in this process involves a parameter λ, which specifies the weight of a topic word w based on its significance within topic t. After adjusting λ, the analysis revealed that setting λ=0.6 provided the most relevant and prominent terms across topics. An overview of the topics identified using LSI for the 2019-2024 dataset is provided in Table 1, showing the final topics determined through this process.

LSI identified nine distinct topics in the dataset, each represented by a set of terms and a proportional share of the corpus, reflecting its prevalence across the analyzed documents.

- Topic 1 focuses on the internet of things (IoT), particularly security and smart devices, highlighting blockchain as a solution for enhancing data privacy, device authentication, and protection against cyber threats. This integration is especially relevant for smart homes, industrial automation, and healthcare applications.
- Topic 2 centers on blockchain technology and its role in information management systems. It emphasizes blockchain's decentralized architecture for securing digital transactions, improving data handling, and enhancing transparency and trust across sectors such as finance, supply chain management, and governance.
- Topic 3 addresses supply chain management and traceability, especially in the food sector. Blockchain improves transparency, product authentication, and regulatory compliance by enabling immutable and real-time tracking across production and distribution processes.
- Topic 4 highlights data security and privacy, focusing on blockchain-based schemes for protecting sensitive information. These solutions enhance system resilience and integrity in domains such as finance, healthcare, and cloud storage by eliminating centralized points of failure.

- Topic 5 examines blockchain applications in energy markets, particularly peer-to-peer electricity trading and smart grid management. Blockchain supports efficient, transparent energy transactions and facilitates the integration of distributed renewable energy resources.
- Topic 6 explores the broader research landscape of blockchain, including challenges related to scalability, interoperability, and governance. Advances in consensus mechanisms and cryptographic techniques are driving blockchain adoption and digital transformation across industries.
- Topic 7 discusses blockchain-based system and software design, emphasizing architectural considerations for decentralized applications. Such systems enable secure, scalable, and autonomous digital services in areas including finance, e-governance, and cloud computing.
- Topic 8 focuses on healthcare, emphasizing blockchain's role in securing medical data, improving interoperability, and ensuring the integrity of patient records. These applications are critical for EHRs, clinical trials, and pharmaceutical supply chains.
- Topic 9 addresses smart contracts and blockchain-based financial technologies. Smart contracts automate and secure financial transactions, supporting decentralized finance (DeFi) platforms and improving efficiency in banking, insurance, and investment services.

Table 2 presents the results obtained from the LDA model, including the representative terms and their proportions within the corpus. The identified topics cover a broad range of themes, such as IoT security, blockchain applications in healthcare, smart contracts, and energy systems. These themes reflect the diversity of research directions captured by the model.

- Topic 1 focuses on academic research in AI and knowledge management, emphasizing research methodologies, applications, and future trends. These studies contribute to intelligent systems that support automation, data-driven decision-making, and innovation across domains such as healthcare, finance, and education.
- Topic 2 addresses digital transformation in industries, particularly in supply chain management. The topic highlights how emerging technologies enhance efficiency, transparency, and adaptability through automation, real-time data access, and improved coordination among stakeholders.
- Topic 3 relates to digital healthcare platforms, including patient data management and blockchain-based solutions. These technologies enhance data security, interoperability, and service efficiency, supporting telemedicine, real-time monitoring, and secure information sharing.
- Topic 4 examines IoT technologies with a prioritization of data security and privacy, and secure communication. It highlights mechanisms for protecting IoT systems from cyber threats, which are critical for applications in smart cities, industrial automation, and healthcare.
- Topic 5 centers on smart contracts, blockchain, and machine learning models. The integration of these technologies enables automated transactions, fraud detection, and intelligent decision-making across decentralized networks.
- Topic 6 focuses on information management and data sharing in supply chains. Secure access control and efficient information exchange enhance transparency, coordination, and operational reliability among supply chain participants.
- Topic 7 highlights security and privacy issues in IoT and healthcare environments, emphasizing protective measures for cloud-based data processing and storage to ensure regulatory compliance and trust in digital services.
- Topic 8 addresses blockchain network security and consensus mechanisms. It examines algorithms that ensure transaction validity, system integrity, and resistance to attacks in decentralized applications.
- Topic 9 focuses on energy markets and smart grids, emphasizing decentralized energy trading using blockchain technology and efficient resource management. These technologies support sustainable, resilient, and decentralized energy infrastructures.

As can be seen from Tables 1 and 2, it was defined predominant research topics in the field of blockchain engineering according to research question 1. The heatmap as shown in Figure 6 visualizes the differences between topics in a topic modeling analysis, specifically using the Jaccard distance as the measure of difference. Here's a detailed description of what the plot represents.

The x- and y-axes represent nine topics (0–8) identified by the model, while the color scale ranges from 0 to 1, indicating Jaccard distance between topic pairs. Dark blue values represent low distance and high similarity, whereas dark red values indicate high distance and minimal overlap. The red diagonal reflects self-comparisons of topics, which is expected, while off-diagonal cells show pairwise topic similarities.

Lighter blue cells suggest moderate overlap between topics, indicating related themes or shared vocabulary, whereas darker blue cells reflect stronger similarity. For instance, topics 0 and 2 show moderate overlap, while topics such as 4 and 6 appear more distinct, reflecting unique thematic content. Overall, the

visualization demonstrates that the topic model effectively differentiates between themes while revealing areas of thematic overlap.

Additional insights from LSI highlight strong associations among topics related to energy trading, supply chains, and healthcare. The consistency between LDA and LSI results reinforces the robustness of the identified topic structure. Topic relationships were visualized using pyLDAvis, with Figure 7 summarizing the intertopic similarities.

Table 1. Blockchain engineering topics (2019-2024) for LSI

No	Topic name (proportion in the whole corpus)	Representative terms
1	IoT and security (15%)	"iot"(0.022), "security" (0.017), "smart" (0.014), "blockchain" (0.011), "network" (0.011), "learning" (0.010), "devices" (0.009), "internet" (0.008), "data" (0.008), "proposed" (0.007)
2	Blockchain technology and systems (18%)	"blockchain" (0.026), "technology" (0.023), "development" (0.018), "model" (0.012), "information" (0.012), "management" (0.011), "system" (0.011), "based" (0.010), "data" (0.010), "paper" (0.009)
3	Blockchain in supply chain management (12%)	"supply" (0.042), "chain" (0.041), "blockchain" (0.018), "technology" (0.012), "food" (0.012), "model" (0.009), "product" (0.008), "chains" (0.008), "study" (0.007), "traceability" (0.007)
4	Blockchain security systems (14%)	"data" (0.032), "blockchain" (0.020), "security" (0.013), "system" (0.012), "proposed" (0.010), "scheme" (0.008), "development" (0.008), "privacy" (0.007), "based" (0.007), "technology" (0.007)
5	Blockchain in energy trading (9,0%)	"energy" (0.056), "power" (0.017), "trading" (0.015), "blockchain" (0.011), "market" (0.010), "electricity" (0.009), "grid" (0.008), "development" (0.008), "system" (0.008), "smart" (0.007)
6	Blockchain research and technological advancements (11%)	"blockchain" (0.021), "research" (0.017), "technology" (0.013), "technologies" (0.012), "study" (0.011), "digital" (0.011), "development" (0.010), "industry" (0.007), "paper" (0.006), "challenges" (0.006)
7	Blockchain software and system design (10%)	"blockchain" (0.016), "data" (0.014), "system" (0.010), "software" (0.008), "design" (0.007), "information" (0.007), "use" (0.006), "digital" (0.006), "technology" (0.006), "decentralized" (0.006)
8	Blockchain in healthcare (13%)	"data" (0.029), "medical" (0.023), "health" (0.021), "healthcare" (0.020), "system" (0.013), "blockchain" (0.012), "information" (0.009), "technology" (0.008), "patient" (0.007), "patients" (0.007)
9	Smart contracts and blockchain in financial applications (12%)	"smart" (0.020), "blockchain" (0.016), "contracts" (0.015), "financial" (0.011), "development" (0.011), "contract" (0.010), "digital" (0.008), "technology" (0.008), "transactions" (0.006), "cryptocurrency" (0.005)

Table 2. Blockchain engineering topics (2019-2024) for LDA

No	Topic name (proportion in the whole corpus)	Representative terms
1	General research in AI and knowledge management (10.5%)	"research" (0.025), "paper" (0.010), "analysis" (0.009), "field" (0.008), "application" (0.008), "study" (0.008), "intelligence" (0.007), "knowledge" (0.007), "ai" (0.007), "future" (0.006)
2	Digital transformation and supply chain management (9.0%)	"study" (0.013), "digital" (0.010), "research" (0.009), "industry" (0.008), "chain" (/), "supply" (0.007), "paper" (0.006), "business" (0.006), "new" (0.005), "management" (0.005)
3	Digital health platforms and blockchain in healthcare (8.0%)	"health" (0.018), "platform" (0.012), "medical" (0.012), "patient" (0.009), "record" (0.008), "sharing" (0.006), "bitcoin" (0.006), "care" (0.006), "security" (0.006), "nft" (0.006)
4	IoT security, privacy, and communication protocols (11.0%)	"scheme" (0.020), "storage" (0.011), "vehicle" (0.010), "security" (0.009), "proposed" (0.008), "internet" (0.008), "node" (0.007), "iot" (0.007), "communication" (0.007), "privacy" (0.007)
5	Smart contracts, blockchain technology, and machine learning (12.5%)	"contract" (0.032), "smart" (0.030), "model" (0.017), "method" (0.010), "proposed" (0.008), "detection" (0.008), "network" (0.007), "learning" (0.007), "result" (0.006), "code" (0.006)
6	Information management and user access in supply chains (7.5%)	"chain" (0.015), "information" (0.013), "management" (0.011), "supply" (0.010), "user" (0.010), "model" (0.010), "sharing" (0.009), "problem" (0.009), "mechanism" (0.008), "access" (0.008)
7	Privacy and security in IoT and Cloud services (9.5%)	"security" (0.022), "iot" (0.020), "privacy" (0.014), "device" (0.010), "healthcare" (0.009), "proposed" (0.009), "internet" (0.009), "cloud" (0.008), "service" (0.007), "thing" (0.007)
8	Network security and consensus algorithms in blockchain (10.0%)	"network" (0.024), "security" (0.018), "consensus" (0.011), "application" (0.011), "smart" (0.009), "algorithm" (0.008), "communication" (0.007), "proposed" (0.007), "transaction" (0.007), "mechanism" (0.007)
9	Energy trading, smart grids, and resource management (12.0%)	"energy" (0.039), "power" (0.012), "trading" (0.012), "smart" (0.011), "transaction" (0.010), "model" (0.009), "market" (0.008), "proposed" (0.008), "grid" (0.007), "resource" (0.007)

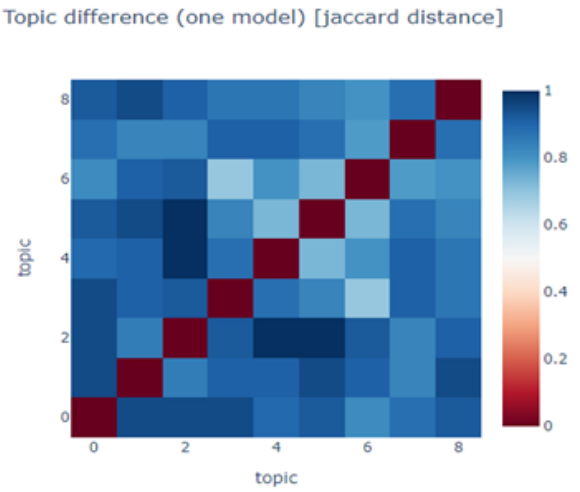

Figure 6. Heatmap of document similarities

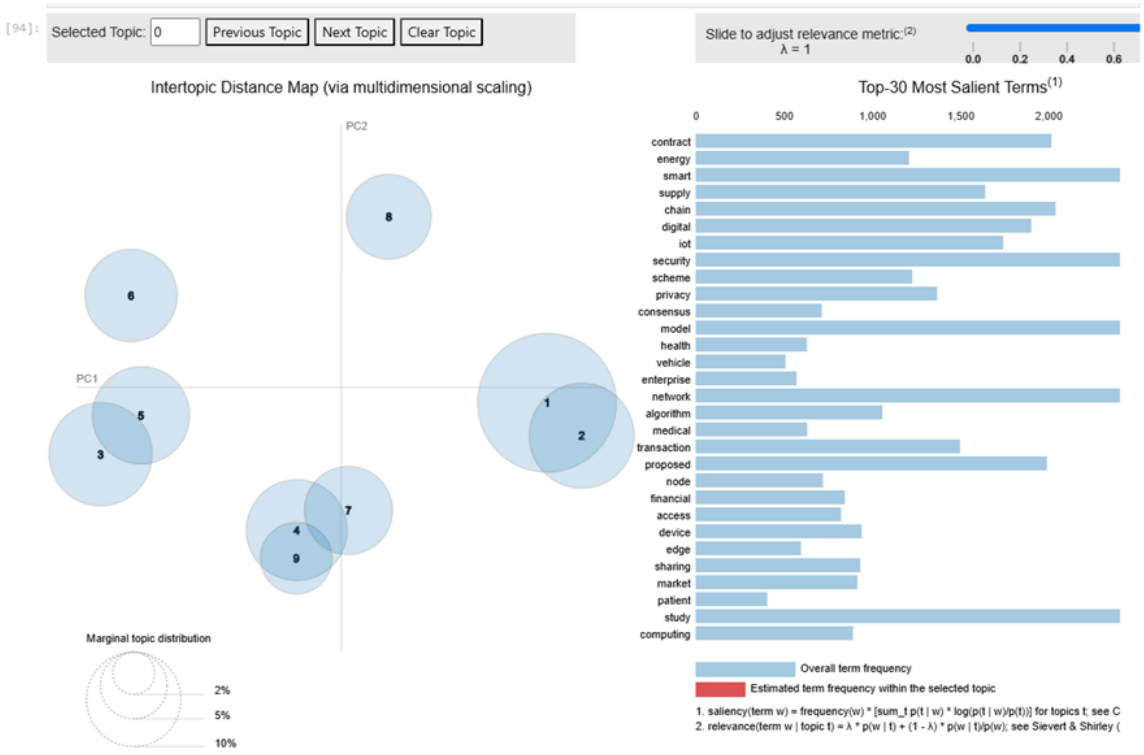

Figure 7. Visualization of topics

Figure 7 illustrates the outcomes of the LDA-based topic modeling through an intertopic distance map alongside a Top-30 most salient terms visualization, providing an overview of the corpus's underlying thematic structure. Displayed in the left panel is the intertopic distance map produced using multidimensional scaling (MDS), in which each circle corresponds to a topic, and the circle size reflects its relative prominence within the dataset. Spatial proximity reflects semantic similarity between topics, while the PC1 and PC2 axes capture the main variance in topic–word distributions. A marginal topic distribution inset summarizes the overall contribution of each topic to the corpus.

The right panel illustrates the 30 most important terms, with blue bars indicating overall term frequency and red overlays showing topic-specific relevance. Frequently occurring terms such as blockchain, contract, smart, security, consensus, and privacy indicate strong representation of blockchain, cybersecurity,

IoT, and digital transformation themes. These terms align primarily with topic 5 (smart contracts and machine learning) and topic 8 (blockchain security and consensus mechanisms), highlighting blockchain's expanding role across industries.

Security-related terms, including privacy, IoT, and scheme, emphasize ongoing concerns in IoT, cloud services, and secure communication, corresponding to topics 4 and 7. Meanwhile, terms such as digital, business, management, and research reflect digital transformation and AI-driven innovation associated with topics 1 and 2. Healthcare-related terms (health, patient, and medical) and energy-related terms (energy, power, and trading) align with topics 3 and 9, respectively, indicating increasing adoption of blockchain and smart technologies in healthcare systems and sustainable energy markets.

The findings align with prior research in topic modeling within text mining, where LDA has been widely used to uncover hidden thematic structures in textual datasets. Similar to previous studies that have analyzed emerging technologies through topic modeling, this study highlights the prevalence of blockchain and IoT as dominant research themes. However, while some studies have emphasized blockchain's role in finance and cybersecurity, our results indicate a broader scope, including supply chains, healthcare, and energy. A major strength of this research lies in its ability to systematically uncover thematic structures within academic abstracts, providing a structured view of technological trends. The identification of research gaps, particularly in blockchain applications in healthcare and energy, further adds value to the field.

4. CONCLUSION

This study examined the thematic structure and development trends of blockchain engineering research from a global perspective by applying topic evolution analysis and LDA topic modeling to 3,665 WoS articles published between 2019 and 2024. The analysis identified nine major topics and revealed a clear shift in research focus toward emerging technologies such as blockchain and IoT and their applications in supply chains, healthcare, and energy systems, highlighting blockchain's growing role in technological innovation. The results provide a structured overview of topic relationships and research dynamics, demonstrating that the LDA model effectively captures the core thematic areas despite moderate coherence. As the first large-scale application of LDA topic modeling in blockchain engineering, this work offers valuable theoretical and methodological insights for future research; however, its scope is limited by reliance on English-language abstracts from a single database. Future studies should extend this approach to full-text, multilingual datasets to achieve a more comprehensive understanding of blockchain engineering research trends.

5. FUTURE DIRECTIONS

This study offers a foundational examination of hidden topics and emerging trends in blockchain engineering through LDA topic modeling. However, several promising directions for future research can enhance and extend this work. The authors have used this model on blockchain engineering in the aim of creating scenarios for development of an educational game on smart contract programming "smart you" taking into account key topics of future trends. Future studies could perform a longitudinal analysis with finer temporal granularity to detect micro-trends within specific blockchain applications, covering applications such as DeFi, non-fungible tokens (NFTs), and privacy-enhancing solutions. Investigating the interplay between emerging topics and industry adoption would also offer meaningful insights into the real-world implications of research trends.

ACKNOWLEDGEMENTS

The authors would like to express our sincere gratitude to the Bolashak scholarship program and tutors from King's College London for their guidance and support in helping complete this research paper.

FUNDING INFORMATION

The article was supported by the project from the Ministry of Science and Higher Education of the Republic of Kazakhstan, No. BR24992975 "Development of a digital twin of a food processing enterprise using AI and IIoT technologies (2024-2026)."

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dinara Zhaisanova	✓	✓	✓	✓	✓			✓	✓	✓	✓			✓
Madina Mansurova	✓					✓	✓	✓		✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY

WoS core collection dataset utilized in this study are available in Github at https://github.com/DinaraZhay/LDA-Topic-Modeling/blob/main/block_all.xls. Also, the source code with all graphs written in Python is available at the link: https://github.com/DinaraZhay/LDA-Topic-Modeling/blob/main/_Block_ipynb_.ipynb.

REFERENCES

[1] J. Angelis and E. R. da Silva, "Blockchain adoption: a value driver perspective," *Business Horizons*, vol. 62, no. 3, pp. 307–314, May 2019, doi: 10.1016/j.bushor.2018.12.001.

[2] P. Garg, B. Gupta, K. N. Kapil, U. Sivarajah, and S. Gupta, "Examining the relationship between blockchain capabilities and organizational performance in the Indian banking sector," *Annals of Operations Research*, vol. 348, no. 3, pp. 1513–1546, May 2025, doi: 10.1007/s10479-023-05254-0.

[3] A. Hughes, A. Park, J. Kietzmann, and C. A.-Brown, "Beyond bitcoin: what blockchain and distributed ledger technologies mean for firms," *Business Horizons*, vol. 62, no. 3, pp. 273–281, May 2019, doi: 10.1016/j.bushor.2019.01.002.

[4] H. J. Scholl and M. P. R. Bolívar, "Regulation as both enabler of technology use and global competitive tool: the gibraltar case," *Government Information Quarterly*, vol. 36, no. 3, pp. 601–613, Jul. 2019, doi: 10.1016/j.giq.2019.05.003.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4–5, pp. 993–1022, 2003, doi: 10.7551/mitpress/1120.003.0082.

[6] P. Mennig, "Who cares about agriculture? analyzing German parliamentary debates on agriculture and food with structural topic modeling," *Food Policy*, vol. 130, Jan. 2025, doi: 10.1016/j.foodpol.2024.102788.

[7] M. F. Özmantar, K. Gökdağ, T. Hangül, and G. Agaç, "Research themes and trends in the field of teacher educators: a topic modelling study," *Teaching and Teacher Education*, vol. 148, Oct. 2024, doi: 10.1016/j.tate.2024.104696.

[8] S. Y. Park, X. Wang, Y. Oh, S. M. Hong, and S. H. Woo, "Application of structural topic modeling in a literature review of air transport," *Journal of Air Transport Management*, vol. 122, 2025, doi: 10.1016/j.jairtraman.2024.102708.

[9] D. Yu and B. Xiang, "Discovering topics and trends in the field of artificial intelligence: using LDA topic modeling," *Expert Systems with Applications*, vol. 225, Sep. 2023, doi: 10.1016/j.eswa.2023.120114.

[10] A. Shukla, P. Jirli, A. Mishra, and A. K. Singh, "An overview of blockchain research and future agenda: insights from structural topic modeling," *Journal of Innovation and Knowledge*, vol. 9, no. 4, 2024, doi: 10.1016/j.jik.2024.100605.

[11] L. Zheng, Z. He, and S. He, "A topic model-based knowledge graph to detect product defects from social media data," *Expert Systems with Applications*, vol. 268, Apr. 2025, doi: 10.1016/j.eswa.2024.126313.

[12] C. Sharma, S. Sharma, and Sakshi, "Latent Dirichlet allocation (LDA) based information modelling on blockchain technology: a review of trends and research patterns used in integration," *Multimedia Tools and Applications*, vol. 81, no. 25, pp. 36805–36831, Oct. 2022, doi: 10.1007/s11042-022-13500-z.

[13] J. Lee, H. Zo, and T. Steinberger, "Exploring trends in blockchain publications with topic modeling: implications for forecasting the emergence of industry applications," *ETRI Journal*, vol. 45, no. 6, pp. 982–995, Dec. 2023, doi: 10.4218/etrij.2022-0257.

[14] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," *2018 Annual National Seminar*, 2008, pp. 1–9.

[15] J. H. Lee and M. J. Ostwald, "Latent Dirichlet allocation (LDA) topic models for space syntax studies on spatial experience," *City, Territory and Architecture*, vol. 11, no. 1, Jan. 2024, doi: 10.1186/s40410-023-00223-3.

[16] C. Sharma, I. Batra, S. Sharma, A. Malik, A. S. M. S. Hosen, and I.-H. Ra, "Predicting trends and research patterns of smart cities: a semi-automatic review using latent Dirichlet allocation (LDA)," *IEEE Access*, vol. 10, pp. 121080–121095, 2022, doi: 10.1109/ACCESS.2022.3214310.

[17] B. Yin and C.-H. Yuan, "Detecting latent topics and trends in blended learning using LDA topic modeling," *Education and Information Technologies*, vol. 27, no. 9, pp. 12689–12712, Nov. 2022, doi: 10.1007/s10639-022-11118-0.

[18] J. Han, G. Liu, and Y. Yang, "Latent Dirichlet allocation-based topic mining analysis of educational scientific research projects based on 2360 NSF education projects," *TEM Journal*, pp. 865–875, May 2023, doi: 10.18421/TEM122-32.

[19] H. Zhang, T. Daim, and Y. P. Zhang, "Integrating patent analysis into technology roadmapping: a latent Dirichlet allocation based technology assessment and roadmapping in the field of blockchain," *Technological Forecasting and Social Change*, vol. 167, Jun. 2021, doi: 10.1016/j.techfore.2021.120729.

[20] J. Zimmermann, L. E. Champagne, J. M. Dickens, and B. T. Hazen, "Approaches to improve preprocessing for latent Dirichlet allocation topic modeling," *Decision Support Systems*, vol. 185, Oct. 2024, doi: 10.1016/j.dss.2024.114310.

[21] Z. Luo, L. Liu, S. Ananiadou, and Q. Xie, "Graph contrastive topic model," *Expert Systems with Applications*, vol. 255, Dec. 2024, doi: 10.1016/j.eswa.2024.124631.

[22] B. B. Ozkara, M. Karabacak, K. Margetis, W. Smith, M. Wintermark, and V. S. Yedavalli, "Trends in stroke-related journals: examination of publication patterns using topic modeling," *Journal of Stroke and Cerebrovascular Diseases*, vol. 33, no. 6, Jun. 2024, doi: 10.1016/j.jstrokecerebrovasdis.2024.107665.
[23] N. Akar and T. Yörük, "A topic modeling-based analysis for the outcomes of psychological contract breaches and violations in organizations: current research trends and future agenda," *Heliyon*, vol. 10, no. 14, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34908.
[24] S. Alkamli and R. Alabduljabbar, "Understanding privacy concerns in ChatGPT: a data-driven approach with LDA topic modeling," *Heliyon*, vol. 10, no. 20, Oct. 2024, doi: 10.1016/j.heliyon.2024.e39087.
[25] S. M. Saqib, S. Ahmad, A. H. Syed, T. Naeem, M. Fahad, F. M. Alotaibi, "Analysis of latent Dirichlet allocation and non-negative matrix factorization using latent semantic indexing," *International Journal of Advanced and Applied Sciences*, vol. 6, no. 10, pp. 94–102, Oct. 2019, doi: 10.21833/ijaas.2019.10.015.
[26] F. Horasan, "Latent semantic indexing-based hybrid collaborative filtering for recommender systems," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 10639–10653, Aug. 2022, doi: 10.1007/s13369-022-06704-w.
[27] S. Roy, S. Morrell, L. Zhao, and R. Homayouni, "Large-scale identification of social and behavioral determinants of health from clinical notes: comparison of latent semantic indexing and generative pretrained transformer (GPT) models," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, Oct. 2024, doi: 10.1186/s12911-024-02705-x.
[28] H. K. Alay, "Evaluating research trends on the emerging blockchain technology in the fields of business and management: a systematic review," *Journal of Emerging Economies and Policy*, vol. 7, no. 2, pp. 409-417, 2022.
[29] A. G. Gad, D. T. Mosa, L. Abualigah, and A. A. Abohany, "Emerging trends in blockchain technology and applications: a review and outlook," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 9, pp. 6719–6742, Oct. 2022, doi: 10.1016/j.jksuci.2022.03.007.
[30] A.-W. Harzing and S. Alakangas, "Google scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison," *Scientometrics*, vol. 106, no. 2, pp. 787–804, Feb. 2016, doi: 10.1007/s11192-015-1798-9.
[31] M.-A. V.-Baceta, M. Thelwall, and K. Kousha, "Web of Science and Scopus language coverage," *Scientometrics*, vol. 121, no. 3, pp. 1803–1813, Dec. 2019, doi: 10.1007/s11192-019-03264-z.
[32] J. Zhu and W. Liu, "A tale of two databases: the use of Web of Science and Scopus in academic papers," *Scientometrics*, vol. 123, no. 1, pp. 321–335, Apr. 2020, doi: 10.1007/s11192-020-03387-8.
[33] D. Zhaisanova and M. Mansurova, "Blockchain concept for the educational purposes: bibliometric analysis and conceptual structure," *Procedia Computer Science*, vol. 231, pp. 753–758, 2024, doi: 10.1016/j.procs.2023.12.142.

BIOGRAPHIES OF AUTHORS

Dinara Zhaisanova earned, from the Al Farabi Kazakh National University, a B.S. in Computer Science (2009) and a Ph.D. in Innovation Management (2022). She is now the acting associate professor of the Department of the Artificial intelligence and Big Data. Her research interests include innovation technologies, management of technologies, blockchain, cryptography, machine learning, data mining, and natural language processing. She can be contacted at email: zhaisanova15@gmail.com.

Madina Mansurova is candidate of physical and mathematical sciences. She is currently head of the Department of Artificial intelligence and Big Data at Al-Farabi Kazakh National University. Her research interests include big data analytics, neural networks, and the development of information technology for various industrial applications. She has contributed extensively to scientific literature in these fields, with multiple publications in recognized journals. Her research has been supported by national and international funding bodies, including the Ministry of Education and Science of the Republic of Kazakhstan. In addition, she serves on the editorial board of several academic journals and regularly reviews for journals such as the IEEE Transactions on Neural Networks and Learning Systems and Information Systems. She can be contacted at email: madina.mansurova@kaznu.kz.