Journal of Informetrics 19 (2025) 101639

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Forecasting topic trends of blockchain utilizing topic modeling and deep learning-based time-series prediction on different document types

Yejin Park [a,1], Seonkyu Lim [b,1], Changdai Gu [c,d,1], Arida Ferti Syafiandini [a,e], Min Song [a,*]

[a] Department of Library and Information Science, Yonsei University, Seoul 03722, Republic of Korea
[b] Korea Financial Telecommunications & Clearings Institute, Seoul 06220, Republic of Korea
[c] Dept. of Artificial Intelligence, Yonsei University, Seoul 03722, Republic of Korea
[d] Oncocross Co., Ltd., Saechang-ro, Mapo-gu, Seoul 04168, Republic of Korea
[e] Research Center for Computing, National Research and Innovation Agency, 60111, Indonesia

ARTICLE INFO

Keywords:
time-series forecasting
topic modeling
agglomerative clustering
LSTM
GRU
AGCRN
A3T-GCN
blockchain

ABSTRACT

Topic trends in rapidly evolving domains like blockchain are dynamic and pose prediction challenges. To address this, we propose a novel framework that integrates topic modeling, clustering, and time-series deep learning models. These models include both non-graph-based and graph-based approaches. Blockchain-related documents of three types—academic papers, patents, and news articles—are collected and preprocessed. Random and topic subgraphs are constructed as inputs for model training and forecasting across various time epochs. The four models (LSTM, GRU, AGCRN, and A3T-GCN) are trained on random subgraphs, and the trained models forecast topic trends using topic subgraphs. We also analyze the distinctive characteristics of each document type and investigate the causal relationships between them. The results indicate that non-graph-based models, such as LSTM, perform better on periodic data like academic papers, whereas graph-based models, such as AGCRN and A3T-GCN, excel at capturing non-periodic patterns in patents and news articles. Our framework demonstrates robust performance, offering a versatile tool for blockchain-related trend analysis and forecasting. The code and environments are available at https://github.com/textmining-org/topic-forecasting.

Abbreviations

A3T-GCN	attention temporal graph convolutional network
AGCRN	adaptive graph convolutional recurrent network
ASTGCN	attention-based spatial-temporal graph convolutional networks)
DCRNN	diffusion convolutional recurrent neural network; adaptive graph convolutional recurrent network
DMR	Dirichlet multinomial regression

[*] Corresponding author.
E-mail addresses: yejinpark@yonsei.ac.kr (Y. Park), sklim@kftc.or.kr (S. Lim), cdgu@yonsei.ac.kr (C. Gu), afsyafiandini@yonsei.ac.kr (A.F. Syafiandini), min.song@yonsei.ac.kr (M. Song).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.joi.2025.101639
Received 4 June 2024; Received in revised form 5 January 2025; Accepted 15 January 2025
Available online 22 January 2025

GCNs	graph convolutional networks
GRU	gated recurrent unit
LDA	latent Dirichlet allocation
LSTM	long short-term memory
MAE	mean absolute error
MSE	mean squared error
NLTK	Natural Language Toolkit
RNN	recurrent neural network
STGCN	spatio-temporal graph convolutional networks
T-GCN	temporal graph convolutional network
USPTO	United States patent and trademark office.

1. Introduction

In modern society, keeping up with emerging trends is crucial for individuals and organizations to thrive in a dynamic environment. However, predicting the constant evolution of these trends has posed challenges owing to the complexity and unpredictability of changing interests across various industries. As novel topics continuously emerge, developing robust forecasting methods to capture these shifts accurately has become increasingly complex. One promising approach to address this issue is time-series forecasting, which utilizes historical data to predict future trends.

In the time-series forecasting field, numerous studies related to topic modeling approaches have been reported (Vayansky & Kumar, 2020)—independently or in combination with bibliometric information—to predict trends in specific domains (Ghaffari et al., 2023; Ekin et al., 2023; Das et al., 2023). Additionally, some studies have employed network-based forecasting models. However, few have focused on transforming topic modeling results into graph structures and analyzing them with network-based models for trend forecasting. This gap can be a shortcoming, as effective trend forecasting in complex domains like blockchain requires methods that can concurrently capture the evolving topic structures and their temporal dynamics. In addition, there has been limited research on trend analysis within the blockchain domain—a field characterized by rapid technological advancements and diverse applications (Kumar et al., 2023; El Akrami et al., 2023). Blockchain trend analysis presents unique challenges due to the non-periodic nature of certain trends, making it more complex than other domains (Bamakan et al., 2021; Zou et al., 2020).

Trend prediction requires a framework that can effectively analyze patterns from time-series data and execute various tasks, including classification, anomaly detection, and forecasting. Such frameworks should include reliable models capable of processing time-series data and performing specific predictive tasks. Among various models, graph neural networks (GNNs) (Wu et al., 2020), particularly graph convolutional networks (GCNs) (Kipf & Welling, 2016), have emerged as powerful tools for analyzing time-series data. Previous studies have applied GCN architectures to forecasting tasks, such as traffic flow and stock market trends (Jiang & Luo, 2022; Yin et al., 2021; Li et al., 2017; Zhao et al., 2019; Guo et al., 2019; Bai et al., 2021). However, their application specifically for trend forecasting is still limited. Another deep learning method, long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997), has also demonstrated strong performance in trend forecasting. Some studies (Xu et al., 2022) have integrated LSTM with GCN models to capture temporal and structural dependencies between topics, achieving significant forecasting performance. Despite these advancements, existing approaches still have limitations that need to be addressed for further improvement.

This study proposes a framework that combines topic modeling, clustering, and deep learning models to forecast topic trends. Specifically, we employ four deep learning models, both non-graph-based and graph-based: LSTM, gated recurrent units (GRU), adaptive graph convolutional recurrent networks (AGCRN) (Bai et al., 2020), and attention-based temporal graph convolutional networks (A3T-GCN) (Bai et al., 2021). Blockchain-related documents from three document types—academic papers, patents, and news articles—are collected and preprocessed. Subgraphs, including random and topic subgraphs, are constructed as inputs for model training and forecasting. Additionally, we analyze the distinctive characteristics of each document type and explore the causal relationships across document types. The findings demonstrate that non-graph-based models, such as LSTM, perform better on periodic data like academic papers. In contrast, graph-based models, such as AGCRN and A3T-GCN, are more effective in capturing non-periodic patterns in patents and news articles.

2. Related Works

Blockchain was initially introduced as the underlying technology for the cryptocurrency and bitcoin. As the name suggests, a blockchain is a chain of blocks that stores committed transactions and continuously grows as new blocks are added to the system. It operates in a decentralized environment, eliminating the need for intermediaries to validate and verify each transaction (Litke et al., 2019). Blockchain is widely implemented across various systems and domains because of its decentralized and secure nature. Given its broad implementation, trends in blockchain research are continuously evolving, making it a key area of focus for researchers.

Trend analysis explores the direction of topics in specific domains. It has primarily been applied to information extracted from two document types: academic (mining scientific documents) and society-based (processing news articles or social media text). A previous study even compiled the two sources, working with documents collected from USPTO, Google Scholar, news, books, and web pages, and predicted the technology trends (Segev et al., 2015). Ena et al. (2016) introduced a technology trend monitoring methodology that integrates data from nine diverse sources: scientific articles, patent documents, social media, foresight projects, conference proceedings, web pages, published dissertations, and SlideShare presentations—to analyze technological evolution comprehensively.

Furthermore, Li et al. (2019) utilized patent and Twitter (now X) data to monitor the trends of emerging technologies, aiding in understanding technology development and forecasting by comparing the results of patent analysis with Twitter data mining. Ingrole et al. (2021) analyzed microneedle technology across scientific literature, patents, clinical trials, and internet/social media, highlighting robust and growing activity. Although previous studies in trend analysis have contributed valuable insights utilizing different data sources with various characteristics, they mainly focused on highly technical fields. They restricted their focus to conventional technologies or specific application areas. Few studies have extensively explored trends in the blockchain domain.

Trend analysis is usually associated with temporal analysis, which utilizes topic modeling to understand how topics evolve. Vayansky and Kumar (2020) studied topic modeling approaches. They found that conventional methods, such as LDA, typically provide a static snapshot of topic distributions but do not effectively capture the dynamism inherent in longitudinal datasets. The model ignores temporal variations, which are crucial for a deeper understanding of topic evolution. Hence, temporal topic modeling techniques such as dynamic topic models and topics over time (Wang & McCallum, 2006) predict trends more accurately. Still, they require complex parameter tuning and have scalability issues. To overcome these challenges, combining some conventional topic modeling methods and GCN architectures (Kipf & Welling, 2016) may perform better trend analysis with less complex computation.

Some studies have utilized topic modeling for trend forecasting in various domains. For example, Gupta et al. (2022) employed LDA-based topic modeling to predict research trends, focusing on identifying current topics without integrating time-series forecasting models. Similarly, Yu and Xiang (2023) employed LDA to explore topics and trends in the field of artificial intelligence. However, their approach was limited to static analysis without incorporating dynamic temporal models (Gupta et al., 2022), (Yu & Xiang, 2023). In addition, Zou et al. (2024) combined LDA with ARIMA to predict energy policy trends in China; however, ARIMA utilizes only temporal information and cannot handle spatial dependencies (Zou et al., 2024). Recent approaches have shifted beyond conventional LDA-based topic modeling, focusing on trend analysis using large language models (LLMs) such as BERT. For example, Soru and Marshall (2024) proposed a method that uses LLMs for more precise topic extraction and trend analysis, which overcomes the limitations of static analysis and provides flexibility to capture subtle nuances in the text (Soru, T., & Marshall, J., 2024). Wang et al. (2023) utilized BERTopic to track topic changes across various domains dynamically, demonstrating better performance than conventional LDA models (Wang et al., 2023). These approaches highlight the common limitation of focusing only on temporal information or static analysis without fully capturing the complex interdependencies of the data. Our approach addresses these limitations by employing GNN-based models that capture temporal and spatial information, enhancing the prediction performance for complex tasks. This integrated framework allows for a more comprehensive analysis of trends, as demonstrated in our experiments comparing LSTM and gated recurrent units (GRU) (Cho et al., 2014) with GNN-based models.

GCNs are a variant of GNNs, which adapt convolutions to process graph signals (Kipf & Welling, 2016). GCNs can aggregate graph signals within the neighborhood of each node. There are three categories of GCNs: classic CNN-based, propagation-based, and other related general frameworks. Several GNN architectures, including diffusion convolutional recurrent neural networks (DCRNNs) (Li et al., 2017), temporal graph convolutional networks (T-GCNs), and spatio-temporal graph convolutional networks (STGCNs) (Yu et al., 2017), capture the spatial and temporal dependencies and reach state-of-the-art performance in executing downstream tasks. Several GCN architectures, such as AGCRN (Bai et al., 2020) and A3T-GCN (Zhu et al., 2020), employ an attention layer. By re-weighting the influence of historical information, these two models capture the global variation trend more accurately than other GCN architectures.

3. Material and Methods

3.1. Data Preparation

We collect data published between January 1st, 2017, and December 31st, 2023, from the Scopus (academic papers), USPTO

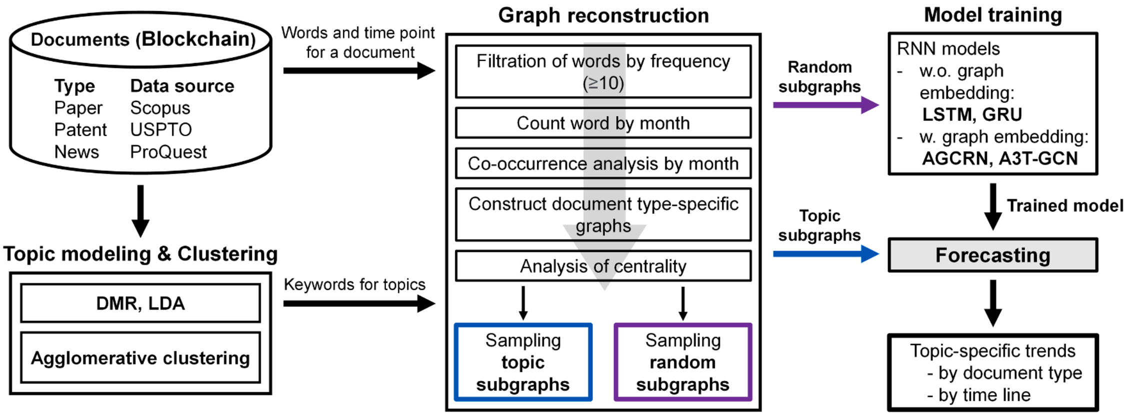


Fig. 1. Overall schematic research workflow. This study comprises five main steps: data preparation, topic modeling and clustering, graph reconstruction, model training, and topic trend forecasting.

(patents), and ProQuest (news articles) database using the search query "Blockchain or Blockchain" (Fig. 1). Specifically, we extract the title, abstract, and keywords from 56,300 academic papers, the title and abstract from 13,524 patents, and the title and full text from 36,991 news articles. This comprehensive dataset forms the basis of our analysis, as depicted in Fig. 1.

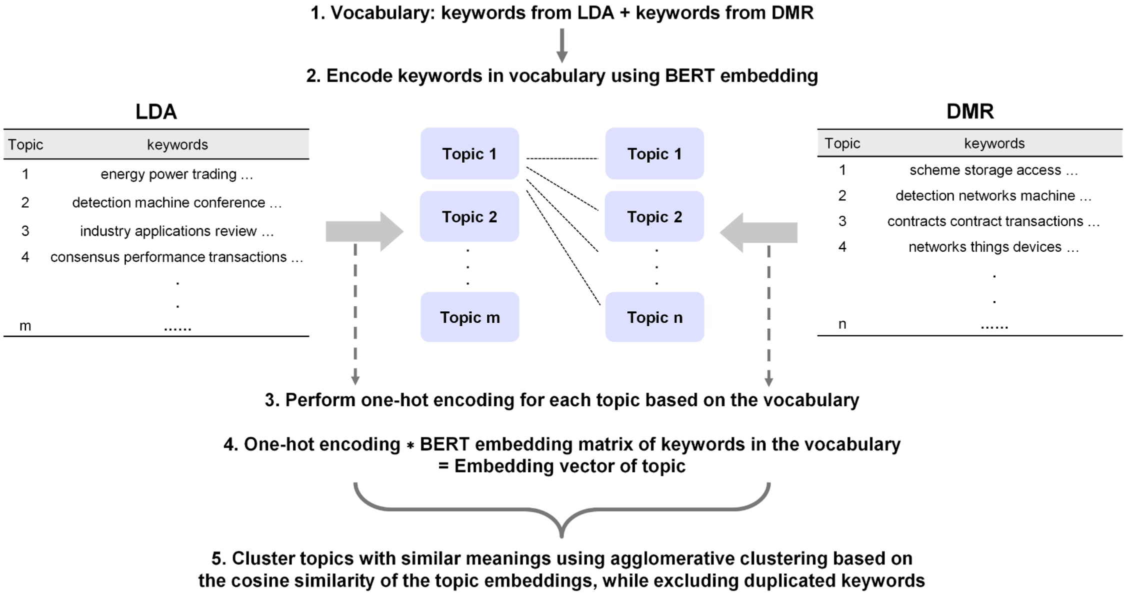
Preprocessing steps are undertaken to prepare the dataset for further analysis. The text of all documents is converted to lowercase and segmented into sentences using the Natural Language Toolkit (NLTK) library. Domain-specific multi-word blockchain terms are standardized into corresponding tokens to ensure accurate representation as single entities (e.g., ’Cryptocurrency Mining’ is converted to ’Cryptocurrency-Mining’). Next, the sentences are tokenized into words, and part-of-speech tagging is performed with NLTK. Only words tagged as nouns are retained, as they are typically more informative for analysis. Finally, stop words—commonly used function words and high-frequency terms identified based on Zipf’s law—are filtered out. Removing these words allows the analysis to focus on more relevant and meaningful content within the dataset.

3.2. Topic Modeling and Clustering

For topic modeling, we employ two methods—latent Dirichlet allocation (LDA) (Blei et al., 2003) and Dirichlet-multinomial regression (DMR) (Mimno & McCallum, 2012)—to generate topics for each document type. Selecting the optimal number of topics is crucial in topic modeling because it significantly affects the model’s performance during training. Generally, perplexity and coherence measures have been utilized to determine the optimal number of topics. Lower perplexity indicates more accurate predictions, while higher coherence indicates better semantic consistency in the topic results. Therefore, one (Boon-Itt & Skunkan, 2020) or both (Fang et al., 2019; Hasan et al., 2021) measures can determine the optimal number of topics. Here, we utilize both coherence and perplexity as indicators to determine the optimal number of topics. To identify the ideal number of topics, we search for the intersection point where the coherence value increases while the perplexity value decreases rapidly.

While BERTopic offers an automated approach for topic modeling utilizing HDBSCAN, its performance on our datasets is suboptimal. Specifically, BERTopic generates highly unbalanced topic counts—13 for academic papers, 3 for patents, and 289 for news articles—and yields consistently negative silhouette scores (-0.06 for academic papers, -0.04 for patents, and -0.11 for news articles), indicating poor cluster performance. These results suggest that BERTopic’s automated clustering approach is unsuited for our datasets, likely because of their varying sizes and complexities. Detailed BERTopic modeling results, including topic numbers and keywords for academic papers, patents, and news articles, are provided in Supplementary Table S1. In contrast, LDA and DMR give better control over topic generation, making them more appropriate for this study.

After topic modeling, We employ a clustering approach to merge the topics generated by LDA and DMR. Our approach is inspired by previous research (Lee et al., 2015), which utilizes cosine similarity to integrate results from LDA and DMR. Lee’s research has demonstrated that this merged approach yields more effective results than using either method alone (Kim et al., 2022; Porter, 2018). For clustering, we compare two methods for computing the embedding value of each topic: element-wise multiplication and sentence embedding. The element-wise multiplication method is detailed in Fig. 2, while sentence embedding generates a topic embedding by treating the collection of keywords within a topic as a single sentence. When evaluating topic clusters formed by both approaches, the element-wise multiplication method produces more cohesive clusters with semantically similar keywords. Based on these results, we select element-wise multiplication as the preferred method for generating topic embeddings.


1. Vocabulary: keywords from LDA + keywords from DMR
2. Encode keywords in vocabulary using BERT embedding
LDA
DMR
Topic 1
Topic 2
Topic m
Topic 1
Topic 2
Topic n
3. Perform one-hot encoding for each topic based on the vocabulary
4. One-hot encoding ∗ BERT embedding matrix of keywords in the vocabulary
= Embedding vector of topic
5. Cluster topics with similar meanings using agglomerative clustering based on
the cosine similarity of the topic embeddings, while excluding duplicated keywords

Topic	keywords
1	energy power trading ...
2	detection machine conference ...
3	industry applications review ...
4	consensus performance transactions ...
	.
	.
m	......

Topic	keywords
1	scheme storage access ...
2	detection networks machine ...
3	contracts contract transactions ...
4	networks things devices ...
	.
	.
n	......

Fig. 2. Topic clustering utilizing agglomerative clustering based on the cosine similarity of topic embeddings.

Fig. 2 illustrates the steps involved in topic clustering. Agglomerative clustering (Murtagh & Legendre, 2014; Müllner, 2011) is employed for merging, utilizing the average linkage method with cosine similarity as the affinity measure. The distance threshold is set to 0.05, as this value produces the highest silhouette score.

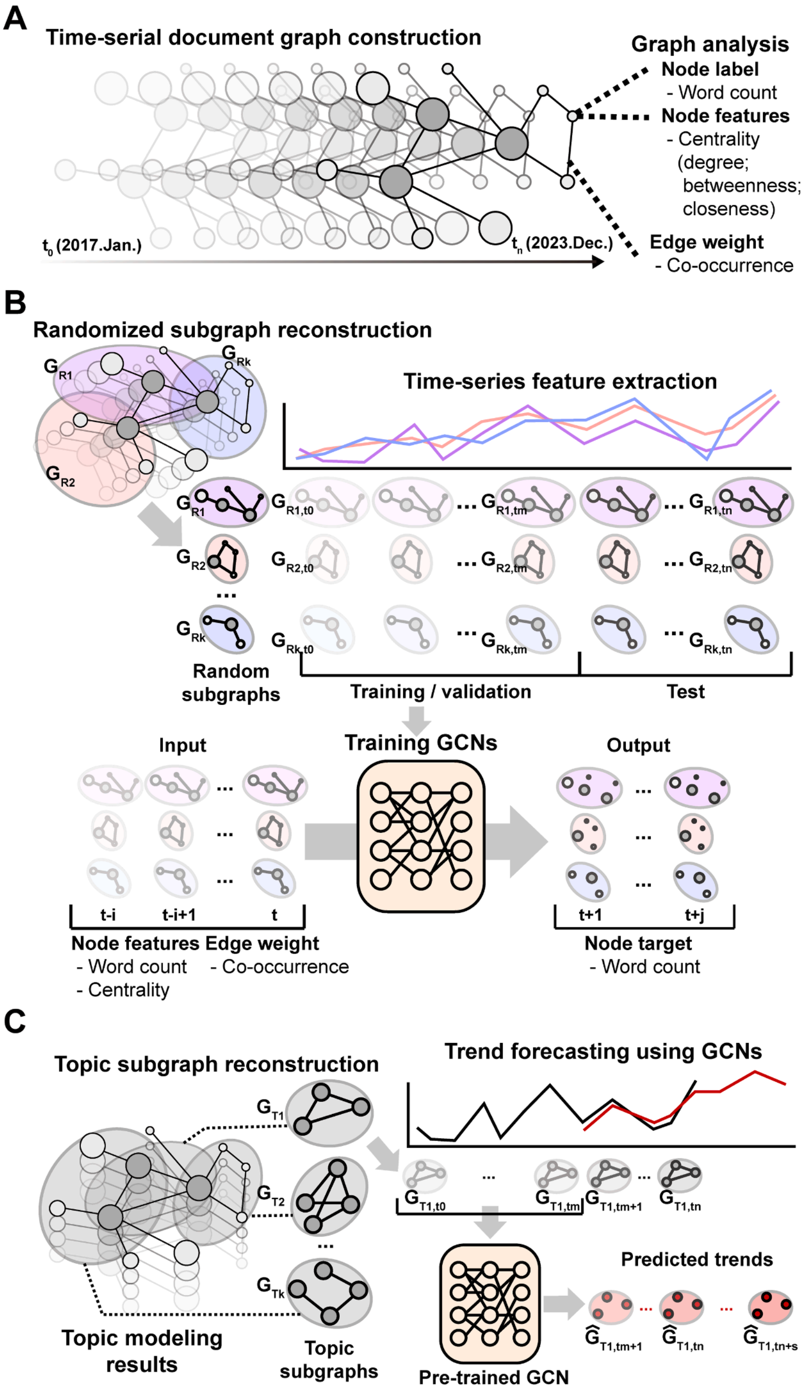


Fig. 3. Schematic flow of graph reconstruction and topic forecasting. A) A time-serial document graph is constructed for each document type—blockchain-related academic papers, patents, and news articles—published between 2017 and 2023, resulting in three distinct document graphs. Word count, co-occurrence, and centralities are calculated for each month. B) Random subgraphs are extracted from the document graph and split into training/validation and test time spans. The GCNs are trained with the time-serial random subgraphs. Word count, centrality of nodes, and co-occurrence data are utilized to predict the word count of nodes for the future timeline. C) Node features and edge weights of topic keywords from the time span (t_{m+1} to t_n) are utilized for forecasting.

1. Combine keywords from topics generated by DMR and LDA to create a unified vocabulary.
2. Use BERT embeddings (Xie et al., 2020; Sia et al., 2020; Miller, 2019) to encode the vocabulary as dense vectors.
3. Perform one-hot encoding for each topic based on the vocabulary, representing topics as sparse binary vectors.
4. Compute topic embeddings by multiplying the one-hot vectors with the BERT embedding matrix, aggregating keyword embeddings into a single dense vector for each topic.
5. Cluster topics with similar meanings using agglomerative clustering based on the cosine similarity of the topic embeddings while excluding duplicated keywords.

3.3. Graph Reconstruction

3.3.1. Data for Document Graph

To construct the document graph, we begin by analyzing the frequency of preprocessed words for each document type (e.g., academic papers, patents, and news articles) across all documents within that type. Subsequently, we focus on retaining only words with a frequency exceeding 10 over the entire time span. This approach allows us to focus on more meaningful and informative words while minimizing computational costs, especially for news articles. Using the refined word set, the co-occurrence of word pairs is measured for all documents. For each document, all possible word pair combinations are identified, with a co-occurrence weight of 1 assigned to each pair to mitigate bias caused by variations in document length. Finally, the word count and co-occurrence data are aggregated on a monthly basis, categorized by document type.

3.3.2. Document Graph

Using the monthly word count and co-occurrence data, a time-serial document graph is reconstructed for each document type (Fig. 3A). In the graph, every node is annotated with its respective monthly and whole-time word counts. Edge weights represent the monthly and whole-time co-occurrence values for each word pair. Additionally, various node centrality measures, including degree centrality, betweenness centrality, and closeness centrality (Zhang & Luo, 2017), are employed for node features. For edge weight to calculate the node centrality, inverted co-occurrence counts are utilized to implement edge distances between nodes.

3.3.3. Random and Topic Subgraphs

Usually, predicting topic trends and their corresponding keywords requires utilizing the entire document graph (i.e., all nodes and edges). However, computational resource limitations—particularly memory usage—restrict the amount of data that can be processed simultaneously. To minimize computational cost, we construct randomly sampled subgraphs (Fig. 3B), which include enough nodes from the document graph rather than loading the entire document graph.

The random subgraphs are reconstructed by iteratively sampling nodes from the document graph with a random walk method, a common node sampling technique in graph analysis and machine learning (Noh & Rieger, 2004; Fouss et al., 2007). The number of nodes in each subgraph is randomly chosen between 8 and 20 nodes, a range determined by topic clustering results, where each topic comprises between 8 and 20 distinct keywords. For random sampling, a seed node is first chosen randomly from the document graph and used to construct a seed node pool. This pool is expanded until its number of nodes reaches the limit, randomly selected between 8 and 20. During expansion, another random node is appended to the node pool iteratively, which connects one of the nodes of the node pool. When selecting a new random node, the selection probability is weighted by the connectivity of the node pool. After defining the node pool for a random subgraph, time-specific features for the nodes of the node pool and edges between the nodes are utilized to reconstruct a time-serial random subgraph.

For training ordinary time-series prediction models, the dataset is typically split into training, validation, and test sets according to distinct time epochs. However, our blockchain-related dataset spans a relatively short period. To address this limitation, we divide the timeline into two parts—early (t_0 to t_m) and late (t_{m+1} to t_n)—using the early period for training and validation, and the late period for test, thereby ensuring temporal independence between these sets (Fig. 3B, upper right). To further minimize data overlap across the training, validation, and test sets, we sample 2,000 time-specific random subgraphs for training, 500 for validation, and 500 for test. Moreover, any random subgraph that contains topic keywords is excluded from the training phase.

For the time-serial topic subgraphs, nodes representing keywords within each topic and their connecting edges are extracted from the document graph. These nodes and edges, along with time-specific node features and edge weights, are used to reconstruct time-serial topic subgraphs for each topic (Fig. 3C). The test time span (t_{m+1} to t_n) is utilized to build these subgraphs.

Because the number of nodes in the subgraphs varies, placeholder nodes are appended for some subgraphs (random and topic) to standardize input and output shapes, ensuring a consistent number of nodes across samples. The node feature for these placeholders is set to 0, and no edge associated with them is utilized.

3.4. Topic Trend Forecasting

This study applies two types of time-series prediction models: RNN models without graph embedding, namely LSTM (Hochreiter & Schmidhuber, 1997) and GRU, and with graph embedding, such as AGCRN (Bai et al., 2020) and A3T-GCN (Bai et al., 2021). For the non-graph-based models, only subgraphs' node features (e.g., word count and centralities) are utilized. In contrast, for the graph-based models, edge weights (e.g., co-occurrence value) are also used along with the node features. We update the node features and edge weights monthly to capture changes in topic trends, as variations in word count can indicate shifts in these trends. By predicting changes in word count with information such as node centrality and co-occurrence, the four models offer an effective approach for accurately forecasting topic trends over time.

3.4.1. Training the Models

Feature selection on the node features and hyperparameter optimization are applied to optimize the models. For feature selection, three centrality types (betweenness, closeness, and degree) are assessed in the permutation of usage. For the forecasting horizon, which indicates the output timespan, 1, 3, 6, 9, and 12 are evaluated, and the fixed lookback timespan is 12 months. The random subgraphs are divided into distinct time steps to facilitate the training and evaluation of the models. The initial 48 months of data are designated for training and validation, while the subsequent 36 months are utilized for the test. Among 2,500 randomly sampled clusters from the initial 48 months, 2,000 are used for training, and the remaining 500 are used for validation by each document type. Similarly, 500 random subgraphs for the later 36 months are used for the test. The node features and edge weights of random subgraphs (with edge weights used only for AGCRN and A3T-GCN) are utilized during training to predict the future word count of nodes.

3.4.2. Forecasting of Topic Trend

We perform forecasting using the trained models, which are trained on random subgraphs, by providing topic subgraphs as input. Node features and edge weights of topic subgraphs (with edge weights used only for AGCRN and A3T-GCN) are employed to predict the future word count of nodes. The forecasting process involves predicting the outcome for future time periods, specifically 1, 3, 6, 9, and 12 months ahead. This forecasting is conducted with a fixed training lookback window of 12 months, meaning the model makes predictions based on the previous 12 months of topic subgraph data (Fig. 3C). We utilize the later 36 months features of the topics for forecasting, which is the timeline of the test dataset, to ensure heterogeneity in the time span for the topic forecasting from training and validation.

3.5. Computational Environments and Accessibility

All experiments were conducted in the following software and hardware environments: UBUNTU 18.04 LTS / CentOS, PYTHON 3.7.11, NETWORKX 2.6.3, PYTORCH 1.11.0, CUDA 11.4.48, NVIDIA Driver 417.22, i9 CPU, and NVIDIA Corporation GA102GL [RTX A6000]. For more detail, codes and environments are available at https://github.com/textmining-org/topic-forecasting.

4. Results

4.1. Topic Modeling and Clustering

Topics for each document type were generated using the LDA and DMR models (see Section 3.2): the optimal number of topics was 10 for academic papers, 19 for patents, and 32 for news articles. The clustering method then integrated the results from both topic models (see Supplementary Tables S2, S3, and S4). The clustering performance was evaluated using the silhouette score (Rousseeuw, 1987; Shahapure & Nicholas, 2020; Ogbuabor & Ugwoke, 2018), which is commonly utilized for clustering evaluation. If the silhouette score exceeded 0.5, the clustering has been considered to show good performance. The score was 0.7 for academic papers, 0.6 for patents, and 0.6 for news articles.

For further analysis, we calculated the semantic similarity between topics from academic papers, patents, and news articles. We used cosine similarity on SentenceTransformer (stsb-roberta-large) embeddings of the keywords for each topic derived from the topic clustering results. Only combinations with a similarity score greater than 0.5 were considered to ensure meaningful semantic relationships. Among these, the combination of Papers Topic 10, Patents Topic 13, and News Topic 7 showed the highest similarity scores, as shown in Table 1 (Papers-Patents: 0.823, Papers-News: 0.914, Patents-News: 0.798). Consequently, this triplet of topics was selected for further analysis, as detailed in Fig. 4, to investigate the relationships and trends shared across these distinct document types.

4.2. Causal Relationship across Document Types

The causal relationships between academic papers, patents, and news articles were explored to identify how trends in one document type influence the others. The analysis was based on monthly word counts for Papers Topic 10, Patents Topic 13, and News Topic 7, which exhibited high semantic similarity and shared thematic overlap (Table 1). This topic set was selected as strong candidates for examining causal interactions based on their relevance to a joint research focus.

Fig. 4 illustrates temporal patterns and potential causal relationships across different document types. Peaks in news often precede increases in scientific activity, suggesting that news coverage may stimulate academic research. However, this influence's magnitude is

Table 1
Semantic similarity between topics from papers, patents, and news.

Papers Topic Number	Patents Topic Number	News Topic Number	Similarity (Papers-Patents)	Similarity (Papers-News)	Similarity (Patents-News)
10	13	1	0.823	0.781	0.739
10	**13**	**7**	**0.823**	**0.914**	**0.798**
10	13	14	0.823	0.853	0.769

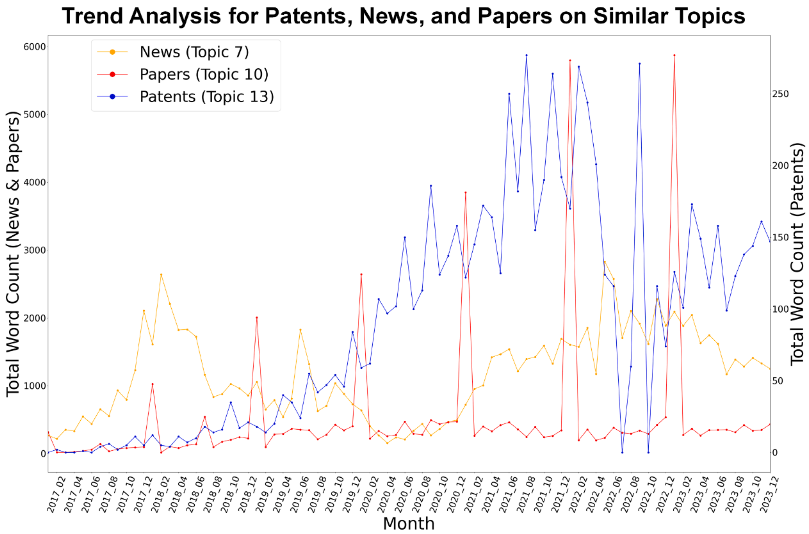


Fig. 4. Monthly word count trends for News Topic 7, Papers Topic 10, and Patents Topic 13 from January 2017 to December 2023. These topics have similar meanings based on the semantic similarity analysis (Table 1). The x-axis represents the timeline divided into two-month intervals, with the left y-axis showing word counts for news articles and academic papers and the right y-axis displaying word counts for patents. These topics, selected for their high semantic similarity, represent a shared research theme.

smaller than the immediate response seen in news trends. Patents exhibit fewer but sharper peaks, reflecting selective innovation processes influenced by news and scientific findings. This highlights how news and research dynamics shape scientific and technological advancements.

In addition to descriptive trend analysis, we conducted Granger causality tests to quantitatively evaluate the influence of one document type on another over various time lags (Fig. 5). These tests examined how trends in academic papers, patents, and news articles impact each other over time. A p-value ≤ 0.05 was used to determine statistically significant causal relationships, providing insights into the temporal interactions between document types. News articles significantly influenced both academic papers and patents, with delays of 4–7 months for academic papers and 2 months for patents (Figs. 5A and 5B). However, neither academic papers nor patents significantly affected news trends within 12 months (Figs. 5C and 5E). Academic papers and patents, on the other hand, exhibited a bidirectional relationship, with academic papers exerting a stronger influence on patents than vice versa (Figs. 5D and 5F). Specifically, patent trends were significantly affected by paper trends with a lag of 4–12 months. These findings highlight the cascading effects of topic trends from news articles to academic papers and patents, as well as the mutual influence between academic papers and patents.

4.3. Time-series Graphs and Features

Supplementary Fig. S1 illustrates changes in node and edge features within time-serial subgraphs from 2017 to 2023, focusing on January and July each year, independent of model training. The topic subgraph examples for each document type reveal temporal variations in node and edge features. Some edges, such as "program"-"computer" in Patents Topic 9 (Supplementary Fig. S1B), consistently show strong connections, while others, like "implementation"-"application" in Papers Topic 3 (Supplementary Fig. S1A), exhibit stronger links in January than in July, indicating seasonality in some cases.

Seasonality may trigger inaccurate bias for training; thus, further investigations were conducted on the word count for each topic by time (Fig. 6). All the topics of academic papers had an increased word count in January compared to other months in both overall time period and month-averaged graph (Figs. 6A and 6B). For Topic 4 of academic papers, the average sum of word count for January was approximately 10-fold higher than in other months (Fig. 6B). This might be due to the yearly publication policies of some journals. For patents (Figs. 6C and 6D), slight peaks were observed at the end of each quarter for a year (Mar., Jun., Sep., and Dec.). In contrast, news articles (Figs. 6E and 6F) exhibited moderate fluctuations by month, as illustrated in the monthly-averaged graphs. For yearly trends, academic papers and patents generally increased over time, whereas news articles did not. Taken together, the time-serial topic subgraphs and word count trends for topics confirmed that academic papers have strong seasonality, with a markedly increased number of documents published in January. Furthermore, the word count of topics of other document types demonstrated moderate seasonality when averaging by month. This result indicates that at least yearly data are required for forecasting topic trends to avoid prediction failure due to seasonality. Therefore, the lookback timespan was fixed to 12 months for all the model training and forecasting.

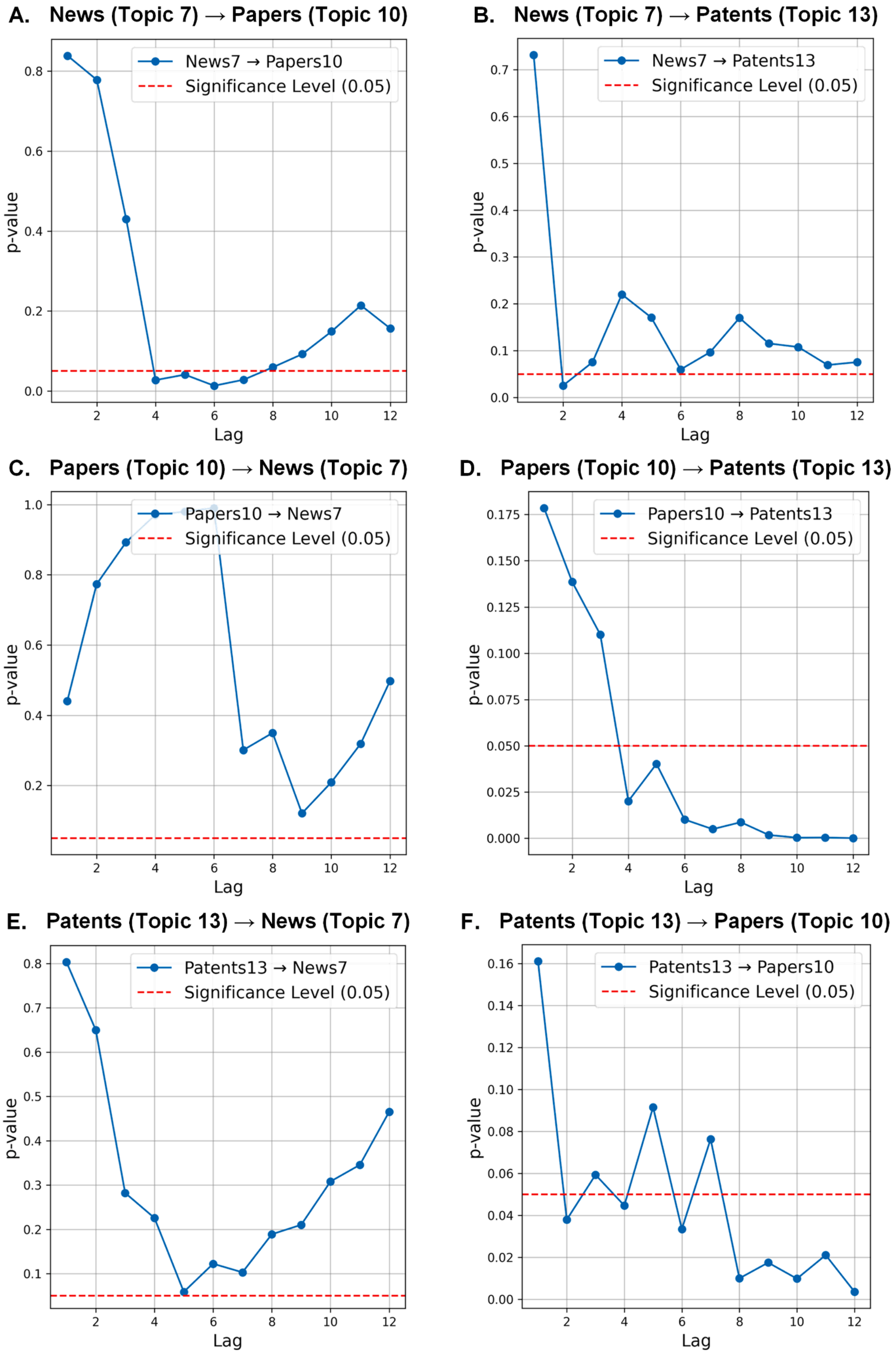


Fig. 5. P-value trends for Granger causality tests between News Topic 7, Papers Topic 10, and Patents Topic 13, with lags of 1–12 months. A) Influence of news on papers, B) Influence of news on patents, C) Influence of papers on news, D) Influence of papers on patents, E) Influence of patents on news, and F) Influence of patents on papers. The red dashed line indicates the significance threshold (p = 0.05), highlighting significant causal relationships at specific time lags.

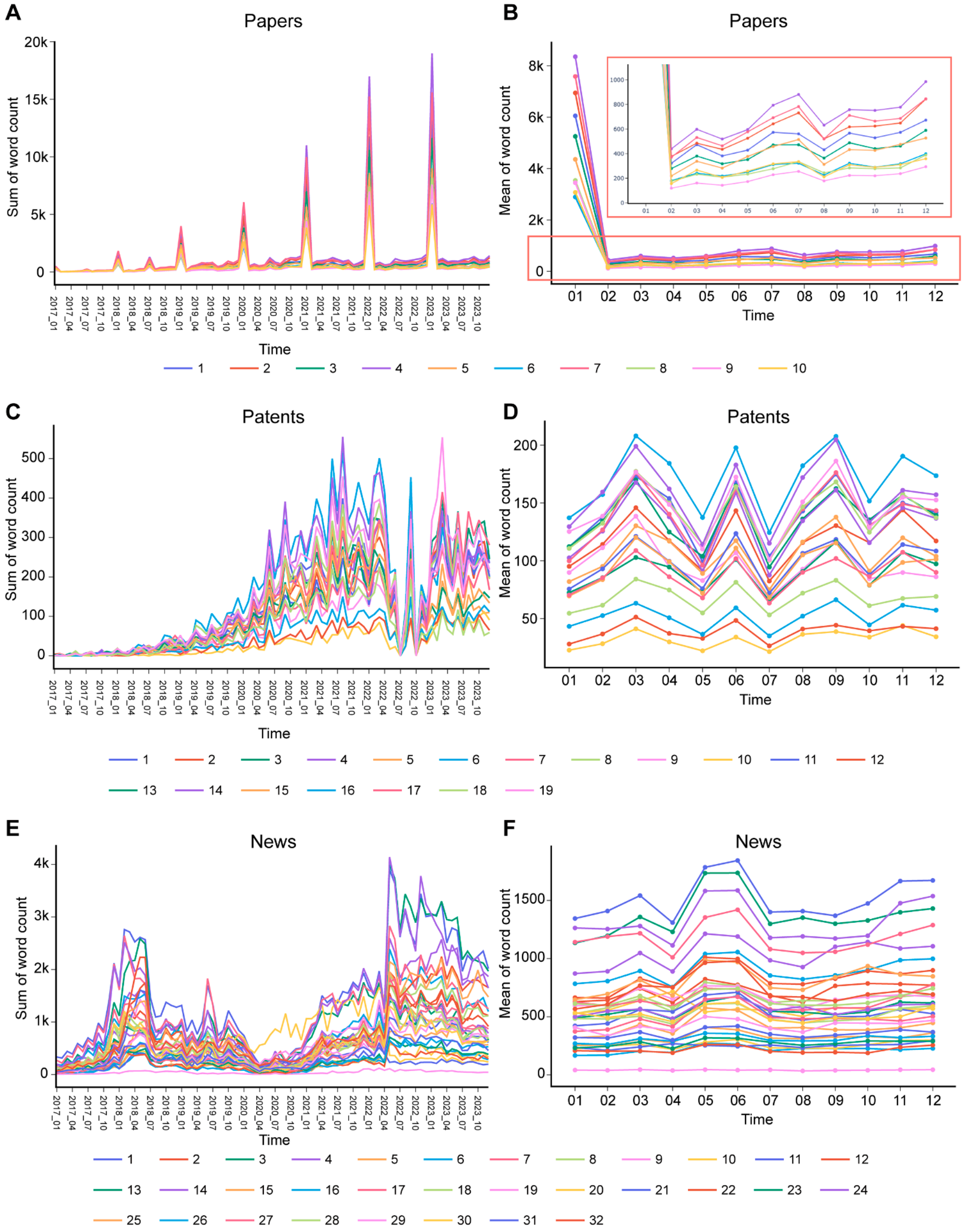


Fig. 6. Seasonality of document types. Word count for each topic across the entire period (A, C, and E) is described, along with month-averaged word count (B, D, and F). The timeline demonstrates month-specific trends for academic papers (A and B), with a significant spike in word counts every January for all topics. This consistent rise can be linked to the publication of special issues or conference proceedings that generally occur at the start of the year. For patents (C and D), the graphs depict seasonality with notable fluctuations in word counts during certain months. These peaks may be influenced by the closure of fiscal quarters, the timing of specific industry product launches, or deadlines for patent filings requiring increased documentation and reporting. News articles (E and F) exhibited no particular trend by month, but the trend by year varied significantly, especially for early 2018 or 2022 to 2023. This variation is likely due to significant events, changes in government policy, or political activities during these periods.

4.4. Topic Trend Forecasting

4.4.1. Training Models

We trained four types of models—LSTM, GRU, AGCRN, and A3T-GCN—using two evaluation metrics: mean squared error (MSE) and mean absolute error (MAE) (Jadon et al., 2022). Feature selection was performed by evaluating various combinations of node centrality features (Table 2; Supplementary Table S5). For academic papers and patents, the use of all centrality types yielded the best performance on average and in most cases. However, for predicting trends in news articles, a single type of centrality produced the best performance in both average MSE and MAE. Therefore, we adopted all types of centralities for academic papers and patents but only degree centrality for news articles. Along with feature selection, hyperparameters for each model were also optimized (Supplementary Table S6) with selected centrality features for each corresponding document type.

4.4.2. Forecasting of Topic

After training, we utilized the trained LSTM, GRU, AGCRN, and A3T-GCN models to forecast topic trends across various forecasting horizons. The evaluation results are presented in Table 3. The predicted values were compared with the actual ground truth values, and the differences were assessed using MSE and MAE as evaluation metrics. Our analysis results indicate that the models' performance varied significantly depending on the document type and the complexity of the topic dynamics. LSTM and GRU were sufficiently robust in forecasting trends for topics within academic papers, which exhibit periodicity and more predictable patterns. These models capably handled the cyclic nature inherent in academic papers, likely due to the consistent publication schedules and recurrent themes.

In contrast, the AGCRN and A3T-GCN models, which incorporate edge information, outperformed forecasting topics from patents and news articles. These documents exhibit higher complexity, with rapidly evolving topics and unpredictable trend shifts driven by external factors and innovations, unlike the regular periodicity of academic papers. The additional context provided by the edge information in the GCN models allows for a more nuanced understanding of the relational dynamics between words, enhancing their predictive capacity in environments with complex and irregular patterns. This analysis underscores the importance of selecting an appropriate model based on the data's characteristics and the forecasting task's specific requirements.

We conducted additional analysis based on the results of our forecasting experiments. Fig. 7 shows the ground truth values and the predicted word counts for topic keywords from academic papers (Topic 10), patents (Topic 13), and news articles (Topic 7)—topics identified as highly similar in Table 1—predicted by four different models at time step t+12. This time point was selected as it effectively captures longer-term trends and fluctuations in topic popularity. Predictions for the other time steps, including t+1, t+3, t+6, and t+9, are provided in Supplementary Fig. S2. The word count predictions for each model and the ground truth values are visualized utilizing histograms for comparison. The results indicate that model performance varies with document characteristics: non-graph-based models, such as LSTM, perform better on periodic data (e.g., academic papers), while graph-based models, such as AGCRN and A3T-GCN, excel at capturing non-periodic patterns in patents and news articles.

5. Conclusions

This study proposed a topic trend forecasting framework that integrates topic modeling, clustering methods, and time-series deep learning models. This approach predicted topic trends by forecasting the word count of topic keywords over various time periods, capturing both temporal dynamics and structural relationships across academic papers, patents, and news articles. A significant insight from this study is the necessity of selecting models that align with the intrinsic characteristics of each document type. LSTM effectively captured the periodic patterns present in academic papers. In contrast, graph-based models such as AGCRN and A3T-GCN were better suited for handling the irregular and non-periodic patterns observed in patents and news articles. Furthermore, our analysis identified potential causal relationships between trends in academic papers, patents, and news articles, highlighting how emerging topics in the news can shape subsequent academic research and eventually drive technological innovations documented in patents.

Despite the promising results of this study, the reliance on conventional topic modeling techniques such as LDA and DMR has certain limitations. While these methods provided greater control over topic generation for our datasets, they primarily captured word co-occurrences without fully accounting for semantic relationships. This can lead to redundancy and reduced cohesion in topic representation. For instance, semantically similar terms like "fund" and "funding" may be grouped into separate topics due to word frequency or context differences. Although BERTopic's automated clustering was not well-suited for our datasets' varying sizes and complexities, its use of sentence embeddings offers the potential for capturing richer semantic nuances. Future research could explore hybrid approaches that integrate conventional methods with advanced techniques like BERTopic to generate more cohesive and contextually meaningful topic clusters, especially for diverse and complex datasets.

Another avenue for advanced research involves developing models enabling transfer learning across different document types. In this study, we trained models independently for each document type. Transferable models could offer a more holistic and comprehensive view of emerging trends. Furthermore, reliance on high-frequency words may limit the framework's ability to detect emerging

Table 2
Average loss of trained models for feature selection.

Features*	Papers		Patents		News	
	MSE	MAE	MSE	MAE	MSE	MAE
b + c	0.02251	0.07661	0.08743	0.02251	0.08027	0.18391
b + d	0.02021	0.07243	0.08118	0.02021	0.07724	0.18362
c + d	0.02196	0.07395	0.08235	0.02196	0.07729	0.18267
b + c + d	**0.01707**	**0.06833**	**0.08096**	**0.01707**	0.07812	0.18442
B	0.02100	0.07411	0.08377	0.02100	0.07797	0.18370
D	0.01965	0.07156	0.08230	0.01965	**0.07565**	0.18100
C	0.02518	0.07966	0.08850	0.02518	0.07578	**0.18059**

*Features utilized here were between centrality (b), closeness centrality (c), and degree centrality (d).

Table 3
Forecasting performance with optimized setting.

Document type	Forecasting horizon	LSTM		GRU		AGCRN		A3T-GCN	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Papers	1	**0.00502**	**0.02750**	0.10398	0.23736	0.01183	0.06401	0.06277	0.10245
("b+c+d ")	3	**0.00626**	**0.03455**	0.08008	0.22605	0.14303	0.31606	0.09126	0.23469
	6	**0.01026**	**0.04072**	0.06275	0.17251	0.02040	0.08086	0.05588	0.16316
	9	**0.00937**	**0.04654**	0.09534	0.22426	0.03024	0.08863	0.08967	0.21604
	12	**0.01914**	**0.06821**	0.09393	0.19479	0.04600	0.11949	0.11431	0.24569
Patents	1	0.40082	0.55728	0.32090	0.48332	**0.06489**	**0.17746**	0.23997	0.38669
("b+c+d ")	3	0.24452	0.39253	0.20349	0.35404	**0.06267**	**0.16884**	0.06347	0.18153
	6	0.20812	0.36720	0.27596	0.41763	**0.06514**	**0.17815**	0.13985	0.27388
	9	0.23964	0.38718	0.16091	0.29940	**0.06639**	**0.18659**	0.19631	0.31939
	12	0.25379	0.40154	0.29069	0.42520	**0.07043**	**0.19825**	0.20494	0.34243
news	1	0.37391	0.50412	0.16170	0.27009	0.11025	0.26337	**0.08533**	**0.22026**
("d")	3	0.27244	0.40685	0.39654	0.52860	0.15997	0.28773	**0.08084**	**0.20372**
	6	0.21712	0.38729	0.26615	0.40259	0.22563	0.36178	**0.07536**	**0.18744**
	9	0.32235	0.44136	0.22032	0.36263	0.19968	0.32852	**0.07591**	**0.18748**
	12	0.31237	0.46915	0.23989	0.37252	0.20284	0.32531	**0.07961**	**0.19070**

*Features utilized here were between centrality (b), closeness centrality (c), and degree centrality (d).

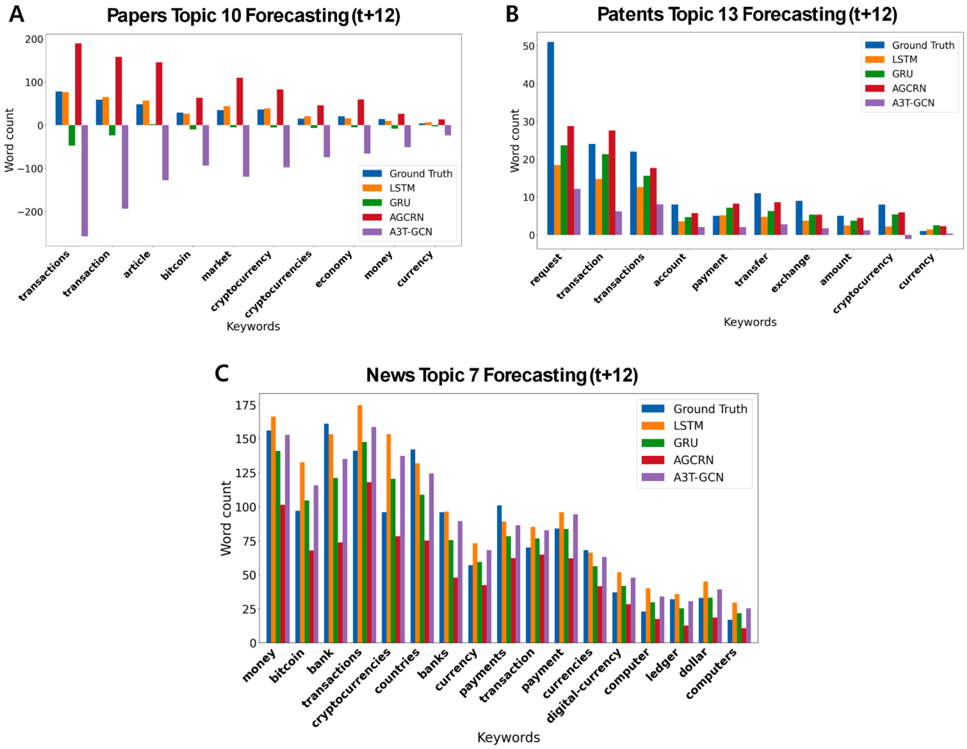


Fig. 7. Word count prediction results for topic keywords in papers (Topic 10), patents (Topic 13), and news (Topic 7) at time step t+12 are presented. The ground truth values (actual values; blue) and predicted word counts for each keyword (x-axis) within the corresponding topics are presented. The prediction models include LSTM (orange), GRU (green), AGCRN (red), and A3T-GCN (purple). (A) For academic papers, the LSTM model closely matches the ground truth values, demonstrating its ability to accurately capture the periodic patterns characteristic of scholarly publications, which are typically influenced by predictable cycles such as conference schedules or journal release dates. (B) For patents, the AGCRN model outperformed the others, effectively capturing variations in word counts. This suggests that AGCRN is better suited for predicting trends in patent data, which are often shaped by irregular external factors such as fiscal cycles and fluctuating market demands. (C) For news articles, the A3T-GCN model provided the most accurate predictions, closely aligning with the ground truth values. This highlights the effectiveness of graph-based models like A3T-GCN in modeling the non-periodic patterns and complex relationships inherent in news data, which are often influenced by sudden events or policy changes.

trends driven by mid- or low-frequency terms. Incorporating methods such as TF-IDF to refine word selection could enhance the understanding of topic evolution, enabling the identification of less prominent but potentially impactful trends.

Finally, while our current research focused on blockchain-related topics, extending the framework's application to other domains is crucial. Such an extension would allow us to evaluate the generalizability of our methodology and its effectiveness in forecasting trends in various fields. By aligning models with the unique attributes of each document type, we aim to establish a robust, domain-agnostic tool capable of anticipating and responding to emerging trends across diverse research areas.

CRediT authorship contribution statement

Yejin Park: Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Seonkyu Lim:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Changdai Gu:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Arida Ferti Syafiandini:** Writing – review & editing, Writing – original draft, Software, Methodology. **Min Song:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

None.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5B1104865).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.joi.2025.101639.

References

Bai, J., Zhu, J., Song, Y., Zhao, L., Hou, Z., Du, R., & Li, H. (2021). A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. *ISPRS International Journal of Geo-Information, 10*(7), 485. https://doi.org/10.3390/ijgi10070485

Bai, L., Yao, L., Li, C., Wang, X., & Wang, C. (2020). Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems, 33*, 17804–17815. https://doi.org/10.5555/3495724.3497218

Bamakan, S. M. H., Bondarti, A. B., Bondarti, P. B., & Qu, Q. (2021). Blockchain technology forecasting by patent analytics and text mining. *Blockchain: Research and Applications, 2*(2), Article 100019. https://doi.org/10.1016/j.bcra.2021.100019

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. Jan.

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance, 6*(4), e21978. https://doi.org/10.2196/21978

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv. arXiv, 1409* (1259). https://doi.org/10.3115/v1/W14-4012

Das, K., Patel, J. D., Sharma, A., & Shukla, Y. (2023). Creativity in marketing: Examining the intellectual structure using scientometric analysis and topic modeling. *Journal of Business Research, 154*, Article 113384. https://doi.org/10.1016/j.jbusres.2022.113384

Ekin, C. C., Polat, E., & Hopcan, S. (2023). Drawing the big picture of games in education: A topic modeling-based review of past 55 years. *Computers & Education, 194*, Article 104700. https://doi.org/10.1016/j.compedu.2022.104700

El Akrami, N., Hanine, M., Flores, E. S., Aray, D. G., & Ashraf, I. (2023). Unleashing the potential of blockchain and machine learning: Insights and emerging trends from bibliometric analysis. *IEEE access : practical innovations, open solutions*. https://doi.org/10.1109/ACCESS.2023.3298371

Ena, O., Mikova, N., Saritas, O., & Sokolova, A. (2016). A methodology for technology trend monitoring: the case of semantic technologies. *Scientometrics, 108*(3), 1013–1041. https://doi.org/10.1007/s11192-016-2024-0

Fang, Y., Guo, Y., Huang, C., & Liu, L. (2019). Analyzing and identifying data breaches in underground forums. *IEEE access : practical innovations, open solutions, 7*, 48770–48777. https://doi.org/10.1109/ACCESS.2019.2910229

Fouss, F., Pirotte, A., Renders, J. M., & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 355–369. https://doi.org/10.1109/TKDE.2007.46

Ghaffari, M., Aliahmadi, A., Khalkhali, A., Zakery, A., Daim, T. U., & Yalcin, H. (2023). Topic-based technology mapping using patent data analysis: A case study of vehicle tires. *Technological Forecasting and Social Change, 193*, Article 122576. https://doi.org/10.1016/j.techfore.2023.122576

Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention-based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI conference on artificial intelligence, 33*(01), 922–929. https://doi.org/10.1609/aaai.v33i01.3301922

Gupta, R. K., Agarwalla, R., Naik, B. H., Evuri, J. R., Thapa, A., & Singh, T. D. (2022). Prediction of research trends using LDA based topic modeling. *Global Transitions Proceedings, 3*(1), 298–304. https://doi.org/10.1016/j.gltp.2022.03.015

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021). Normalized approach to find optimal number of topics in latent Dirichlet allocation (LDA). In M. S. Kaiser, A. Bandyopadhyay, M. Mahmud, & K. Ray (Eds.), Proceedings of international conference on trends in computational and cognitive engineering: Proceedings of TCCE 2020 (pp. 341-354). Springer Singapore.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ingrole, R. S. J., Azizoglu, E., Dul, M., Birchall, J. C., Gill, H. S., & Prausnitz, M. R. (2021). Trends of microneedle technology in the scientific literature, patents, clinical trials and internet activity. *Biomaterials, 267*, Article 120491. https://doi.org/10.1016/j.biomaterials.2020.120491

Jadon, A., Patil, A., & Jadon, S. (2022). A comprehensive survey of regression based loss functions for time-series forecasting. *arXiv. arXiv, 2211*, 02989. https://doi.org/10.48550/arXiv.2211.02989

Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications, 207*, Article 117921. https://doi.org/10.1016/j.eswa.2022.117921

Kim, H., Park, H., & Song, M. (2022). Developing a topic-driven method for interdisciplinarity analysis. *Journal of Informetrics, 16*(2), Article 101255. https://doi.org/10.1016/j.joi.2022.101255

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv. arXiv, 1609*, 02907. https://doi.org/10.48550/arXiv.1609.02907

Kumar, S., Lim, W. M., Sivarajah, U., & Kaur, J. (2023). Artificial intelligence and blockchain integration in business: trends from a bibliometric-content analysis. *Information Systems Frontiers, 25*(2), 871–896. https://doi.org/10.1007/s10796-022-10279-0

Lee, H., Kwak, J., Song, M., & Kim, C. O. (2015). Coherence analysis of research and education using topic modeling. *Scientometrics, 102*, 1119–1137. https://doi.org/10.1007/s11192-014-1453-x

Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change, 146*, 687–705. https://doi.org/10.1016/j.techfore.2018.06.004

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv. arXiv, 1707*, 01926. https://doi.org/10.48550/arXiv.1707.01926

Litke, A., Anagnostopoulos, D., & Varvarigou, T. (2019). Blockchains for supply chain management: architectural elements and challenges towards a global scale deployment. *Logistics, 3*(1), 5. https://doi.org/10.3390/logistics3010005

Miller, D. (2019). Leveraging BERT for extractive text summarization on lectures. *arXiv. arXiv, 1906*, 04165. https://doi.org/10.48550/arXiv.1906.04165

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *arXiv. arXiv, 1206*, 3278. https://doi.org/10.48550/arXiv.1206.3278

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv. arXiv, 1109*, 2378.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification, 31*, 274–295. https://doi.org/10.1007/s00357-014-9161-z

Noh, J. D., & Rieger, H. (2004). Random walks on complex networks. *Physical Review Letters, 92*(11), Article 118701. https://doi.org/10.1103/PhysRevLett.92.118701

Ogbuabor, G., & Ugwoke, F. N. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. AIRCC's. *International Journal of Computer Science and Information Technology*, 27–37.

Porter, K. (2018). Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling. *Digital Investigation, 26*, S87–S97. https://doi.org/10.1016/j.diin.2018.04.023

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1109/TSC.2014.2338855

Segev, A., Jung, S., & Choi, S. (2015). Analysis of Technology Trends Basedon Diverse Data Sources. *IEEE Transactions on Services Computing, 8*(6), 903–915. https://doi.org/10.1109/TSC.2014.2338855

Shahapure, K. R., & Nicholas, C. (2020). *In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747–748). IEEE. https://doi.org/10.1109/DSAA49011.2020.00096

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! *arXiv. arXiv, 2004*, 14914. https://doi.org/10.48550/arXiv.2004.14914

Soru, T., & Marshall, J. (2024). *In 2024 IEEE 18th International Conference on Semantic Computing (ICSC)* (pp. 285–288). IEEE.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems, 94*, Article 101582. https://doi.org/10.1016/j.is.2020.101582

Wang, X., & McCallum, A. (2006). Topics over time. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 424–433). New York: Association for Computing Machinery. https://doi.org/10.1145/1150402.1150450.

Wang, Z., Chen, J., Chen, J., & Chen, H. (2023). Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, 1–26.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Xie, Q., Zhang, X., Ding, Y., & Song, M. (2020). Monolingual and multilingual topic analysis using LDA and BERT embeddings. *Journal of Informetrics, 14*(3), Article 101055. https://doi.org/10.1016/j.joi.2020.101055

Xu, M., Du, J., Xue, Z., Guan, Z., Kou, F., & Shi, L. (2022). A scientific research topic trend prediction model based on multi-LSTM and graph convolutional network. *International Journal of Intelligent Systems, 37*(9), 6331–6353. https://doi.org/10.1002/int.22846

Yin, X., Yan, D., Almudaifer, A., Yan, S., & Zhou, Y. (2021, July). Forecasting stock prices using stock correlation graph: A graph convolutional network approach. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533510.

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18) (pp. 3634-3640). IJCAI. https://doi.org/10.24963/ijcai.2018/505.

Yu, D., & Xiang, B. (2023). Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. *Expert systems with applications, 225*, Article 120114. https://doi.org/10.1016/j.eswa.2023.120114

Zhang, J., & Luo, Y. (2017, March). Degree centrality, betweenness centrality, and closeness centrality in social network. In M. Gholami, R. Jiwari, K. Weller (Eds.), 2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017) (pp. 300-303). Atlantis press. https://doi.org/10.2991/msam-17.2017.68.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., … Li, H. (2019). T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems, 21*(9), 3848–3858. https://doi.org/10.1109/TITS.2019.2935152

Zhu, J., Song, Y., Zhao, L., Hou, Z., Du, R., & Li, H. (2020). A3T-GCN: Attention temporal graph convolutional network for traffic forecasting. *International Journal of Geo-Information, 10*(7), 485. https://doi.org/10.3390/ijgi10070485

Zou, T., Guo, P., Li, F., & Wu, Q. (2024). Research topic identification and trend prediction of China's energy policy: A combined LDA-ARIMA approach. *Renewable Energy, 220*, Article 119619. https://doi.org/10.1016/j.renene.2023.119619

Zou, Y., Meng, T., Zhang, P., Zhang, W., & Li, H. (2020). Focus on blockchain: A comprehensive survey on academic and application. *IEEE access : practical innovations, open solutions, 8*, 187182–187201. https://doi.org/10.1109/ACCESS.2020.3030491