Forecasting Future Topic Trends in the Blockchain Domain: Using Graph Convolutional Network

Yejin Park[1,†], Seonkyu Lim[1,2,†], Changdai Gu[1,3,†] and Min Song[1,*]

[1]Yonsei University, 50, Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea
[2]Korea Financial Telecommunications & Clearings Institute, 432, Nonhyeon-ro, Gangnam-gu, Seoul, Republic of Korea
[3]Oncocross Co., Ltd., 11, Saechang-ro, Mapo-gu, Seoul, Republic of Korea

Abstract

Keeping up with evolving trends is crucial in modern society, but it can be challenging. Time series forecasting analysis has emerged as a promising approach to analyze data over time and identify trend patterns, particularly in fields like blockchain where rapid advancements occur. Graph convolutional networks (GCNs) have shown promise for analyzing structured data, but their effectiveness in domains other than traffic and stock forecasting remains unclear. Efforts to incorporate GCNs for forecasting topic trends have limitations, such as not integrating topic information. To address these limitations, we propose a new approach that combines topic modeling techniques and GCNs for forecasting future topic trends in the blockchain domain. We select an attention temporal graph convolutional network (A3T-GCN) model for its ability to capture global variation trends. Using paper data from the Scopus database, we preprocess the data, identify potential topics using Dirichlet Multinomial Regression and Latent Dirichlet Allocation, and apply agglomerative clustering. We construct two graphs, the random subgraph, and the topic graph, incorporating node features (word count and centralities) and edge weights (co-occurrence). The A3T-GCN model is trained on the random subgraph for forecasting, and the topic graph is used to predict future topic trends in the blockchain with pre-trained models. Our objective is to track key topics and leading keywords shaping the field. The proposed approach has implications for researchers, businesses, and policymakers in understanding topic trends. The paper concludes by presenting the methodology, experimental findings, and future research directions.

Keywords

Forecasting, Topic modeling, Agglomerative clustering, A3T-GCN, Blockchain

1. Introduction

In modern society, where trends are constantly evolving, keeping up with the latest trends is crucial for individuals and organizations to succeed, yet it can be a challenging task. To overcome this challenge, time series forecasting analysis has emerged as a promising approach, enabling individuals to analyze data over time and identify patterns and trends. This approach holds particular relevance in fields like blockchain [1, 2], where rapid technological advancements are commonplace. By leveraging past and present trends, individuals and organizations can make informed predictions about the future, which can give them a competitive edge in the market.

Recently, graph convolutional networks (GCNs) [3] which are variants of graph neural networks (GNNs) [4] have emerged as a promising approach for analyzing structured data, including time-series data. However, most of the existing studies have focused on applying GCNs to forecasting in the field of traffic and stock [5, 6, 7, 8, 9, 10] leaving open the question of whether GCNs can be effectively applied to other domains. Efforts to incorporate deep learning methods such as Long Short-Term Memory (LSTM) and GCNs to capture the temporal and structural dependencies between topics [11] have shown promise in sophisticated forecasting topic trends, but there are still limitations to consider. Notably, these models have yet to incorporate topic modeling [12] method that can be used to identify the main themes and sub-topics in a corpus of documents and can help provide a more nuanced understanding of the research field. Therefore, there is a need to develop more advanced methods that can integrate deep learning techniques with topic modeling analysis to better capture the complex dynamics of topic trends and the intellectual structure of research fields. Such methods could potentially improve the accuracy of forecasting topic trends and provide valuable insights for researchers and decision-makers in various domains.

To address these limitations, we propose a new approach for forecasting future topic trends in the blockchain domain, which combines topic modeling techniques and GCNs. Given the rapid evolution of the blockchain field and the importance of accurate topic forecasting, our proposed method has implications for researchers, businesses, and policymakers. To identify

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online
*Corresponding author.
†These authors contributed equally.
yejinipark@yonsei.ac.kr (Y. Park); sklim@kftc.or.kr (S. Lim); cdgu@yonsei.ac.kr (C. Gu); min.song@yonsei.ac.kr (M. Song)

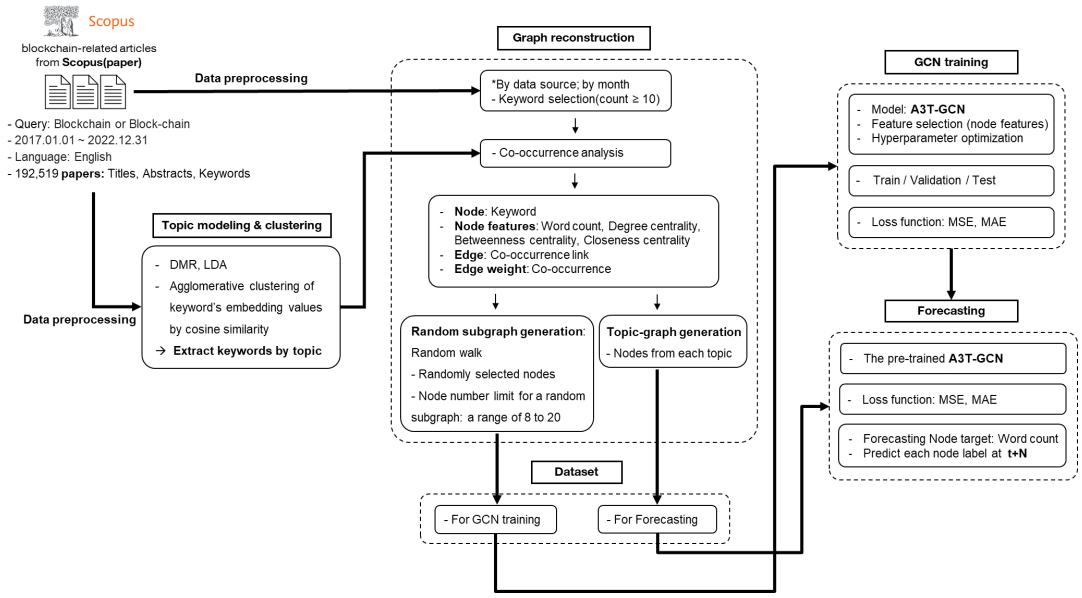


Figure 1: The overall schematic research workflow.

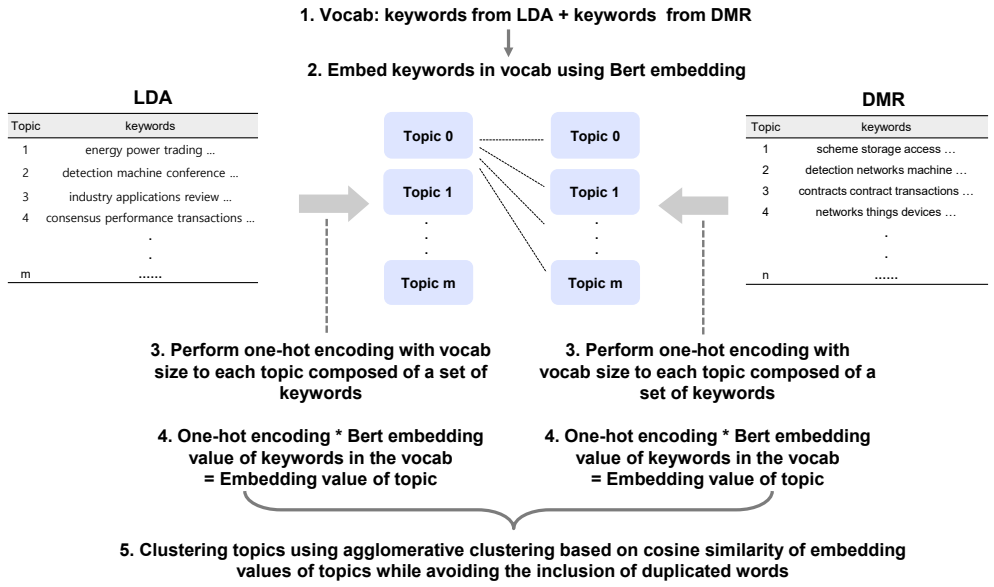


Figure 2: Topic clustering using agglomerative clustering based on the cosine similarity of their embedding values.

the most suitable GCN model for our task of topic trend forecasting, we analyze several GCN models, including the diffusion convolutional recurrent neural network (DCRNN), temporal graph convolutional network (T-GCN), attention based spatial-temporal graph convolutional network (ASTGCN), and attention temporal graph convolutional network (A3T-GCN). DCRNN [7], which employs a bidirectional random walk, captures spatial dependencies. T-GCN [8] combines GCNs and Gated Recurrent Unit (GRU) to capture both spatial and tem-

Table 1
The results of topic clustering.

Topic	Keywords
1	contracts; contract; ethereum; software; applications; voting; blockchains; smart; execution; framework; platform; service; architecture
2	scheme; privacy; access; authentication; encryption; control; storage; vehicles; secure; signature; storage; identity; protection
3	iot; internet; things; devices; networks; edge; communication; privacy; architecture; healthcare; health; applications
4	energy; power; trading; market; grid; electricity; transaction; consumption; demand; resources; resource
5	learning; detection; machine; networks; conference; algorithm; proceedings; topics; papers; prediction; image
6	health; healthcare; records; education; patients; patient; privacy; record; care; insurance
7	bitcoin; cryptocurrency; transactions; transaction; cryptocurrencies; payment; market; currency; money; price
8	supply; traceability; food; industry; logistics; chains; products; quality; manufacturing; production
9	consensus; nodes; block; protocol; algorithm; transaction; performance; transactions; mining; blockchains
10	service; trust; identity; platform; privacy; storage; services; solution; records; integrity
11	research; industry; review; applications; literature; adoption; application; economy; innovation; intelligence

poral dependencies. ASTGCN [9] employs an attention mechanism to capture dynamic correlations in spatial and temporal dimensions. A3T-GCN [10] uses the attention mechanism to improve T-GCN. Among these models, we select A3T-GCN, which appropriately captures the global variation trend by re-weighting the influence of historical information.

To conduct our research, we collect paper data from the Scopus database over a five-year period and extract titles, abstracts, and keywords. After preprocessing the collected data, we employ Dirichlet Multinomial Regression (DMR) [13] and Latent Dirichlet Allocation (LDA) [14] techniques to identify potential topics. We then apply agglomerative clustering [15, 16] to the resulting topic keywords from both models. Next, we proceed to construct two distinct graphs: the first one is known as the random subgraph, which comprises keywords with a count of 10 or more, encompassing both topic keywords and other keywords. The second graph, referred to as the topic graph, solely consists of the topic keywords. The graph reconstruction process involves incorporating node features, including word count and centralities, as well as edge weights derived from co-occurrence analysis. These node features and edge weights are updated on a monthly basis, taking into account changes in keyword word count as indications of shifts in topic trends. Using the random subgraph, we train the A3T-GCN model to forecast topic trends at different time intervals, specifically 1, 3, 6, 9, and 12 months into the future. We use the topic graph to predict future topic trends in the blockchain domain at a point $t+n$ (where n represents the time interval, such as 1 month, 3 months, 6 months, 9 months, or 12 months). These predictions are based on the pre-trained A3T-GCN model.

We focus on the blockchain field, chosen for its potential to bring about transformation in sectors such as finance, supply chain, and healthcare. Our objective is to track leading keywords and identify the main topic that drives industry development. Specifically, we aim to address the research question: Which primary blockchain topics will have a substantial influence on the future of the blockchain industry? This question is crucial for understanding the key factors that will shape the industry's trajectory.

The remainder of this paper is structured as follows: Section 2 presents the methodology that we have proposed, Section 3 details our experimental findings, and finally, we conclude in Section 4.

2. Methodology

Figure 1 shows the overall workflow of the current study.

2.1. Data Preparation

We collect data from research papers published between January 1st, 2017, and December 31st, 2022, from the Scopus database using the search query "Blockchain or Block-chain" (Figure 1). This search yields a total of 192,519 research papers. From these papers, we extract relevant information such as titles, abstracts, and keywords. To prepare the extracted data for further analysis, We perform several preprocessing steps. Firstly, we convert all the text to lowercase. Then, we divide the text into sentence units using the Natural Language Toolkit (NLTK) library. Subsequently, we tokenize the sentences into words and employ NLTK for part-of-speech tagging. We specifically retain words that are tagged as nouns since they are typically more informative for our analysis. In the final preprocessing step, we filter out stopwords, which include commonly used words, meaningless words, and major topic words. Stopwords tend to occur frequently and do not contribute much to the overall understanding of the text. By removing them, we aim to focus on more relevant and meaningful terms within the dataset.

2.2. Topic Modeling and Clustering

In this study, we utilize two widely used topic modeling methods, Latent Dirichlet Allocation (LDA) and Dirichlet Multinomial Regression (DMR). While some previous

works have only employed either DMR or LDA [17, 18], others [19] have shown that combining both methods can yield more effective results. Therefore, we utilize both LDA and DMR to generate topics and obtain the topic distribution throughout the literature. By employing topic modeling, we are able to identify the topics present in the entire document set, determine the relative proportion of each topic in every document, and analyze the distribution of words associated with each topic.

Selecting the optimal number of topics is a crucial step in topic modeling as it significantly affects the model's performance during training. Generally, the perplexity and coherence measures are used 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 [20] or both [21, 22] measures can be used to determine the optimal number of topics. In this study, we utilize both coherence and perplexity as indicators for determining 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.

To achieve a more diverse and precise set of topics, we employ a clustering approach to merge the topics generated by LDA and DMR. Our approach is inspired by previous research [19] which utilized cosine similarity to merge the results from LDA and DMR. In this study, we conduct experiments to compare two methods: element-wise multiplication and sentence embedding, in order to obtain the embedding value for each topic. The element-wise multiplication method involves embedding keywords and multiplying them with the one-hot encoding of each topic, while the sentence embedding method treats keywords within a topic as a sentence and obtains the embedding value for the entire topic. Our findings reveal that the element-wise multiplication method achieves a higher silhouette score of 0.8038 during clustering. Additionally, when examining the resulting keywords from topic clustering using both approaches, the element-wise multiplication method outperforms the alternative method by generating more cohesive clusters with semantically similar keywords. Considering these results and the improved clustering performance, we select the element-wise multiplication method as the preferred approach for constructing topic embeddings.

Initially, topics are generated using both LDA and DMR, with each topic consisting of a set of keywords with similar meanings. Then, we merge similar topics using the following steps, which are presented in Figure 2:

1. Create a vocab by combining the keywords within the topic generated by DMR and LDA.
2. Embed the keywords in the vocab using Bert [23, 24, 25].
3. Perform one-hot encoding with vocab size to each topic composed of a set of keywords.
4. To obtain the embeddings for each topic, we multiply the one-hot encoded representation of the topic with the corresponding Bert embeddings of the keywords in the topic.
5. Cluster topics with similar meanings using agglomerative clustering based on cosine similarity of embedding values of topics while avoiding the inclusion of duplicated words.

Agglomerative clustering is employed for topic merging, utilizing the average linkage method with cosine as the affinity measure. The distance threshold is set to 0.05, as it yields the highest silhouette score.

2.3. Graph Reconstruction

2.3.1. Data for document graph

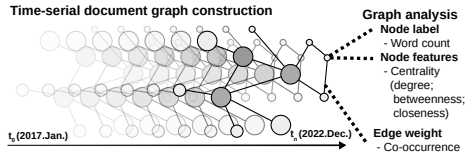
For the document graph, we perform a word count analysis on the preprocessed words in the corpus. We only consider words with a count exceeding 10 throughout the whole time span, aiming to focus on more meaningful and informative words. To calculate monthly co-occurrence, we examine pairs of words at the document level for every month. For each document, we count all combinations of pairs of words with equal weight, disregarding repeated occurrences of identical words to avoid any potential skew in the co-occurrence calculation that may result from variations in document length. All the word counts and co-occurrences are calculated on a monthly basis.

2.3.2. Document graph

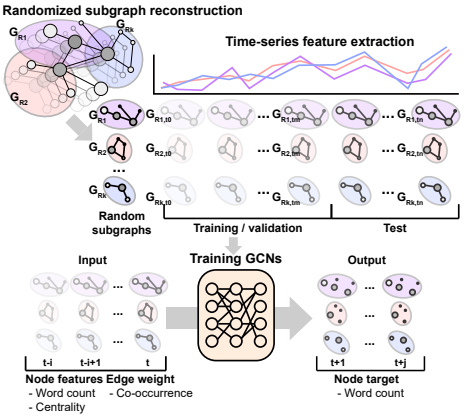
To reconstruct a time-serial document graph (Figure 3 (a)), we employ word count and co-occurrence data on a monthly basis. In the graph, every word node is annotated with its respective monthly and whole-time word count, while the co-occurrence edges are annotated with the monthly and whole-time co-occurrence values between the words they represent at each month. To evaluate the centrality of nodes, we employ various methods such as degree centrality, betweenness centrality, and closeness centrality [26]. To impose edge distance between nodes, we use inverted co-occurrence counts in calculating centralities.

2.3.3. Random and topic subgraphs

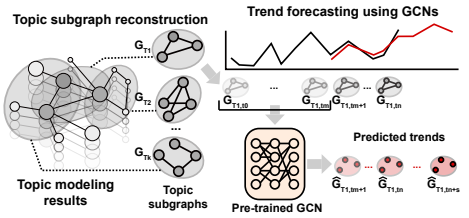
In order to predict the topic trend and its corresponding keywords, it is necessary to utilize the entire word nodes and their corresponding edges from the whole document during training. However, due to memory limitations, it becomes necessary to restrict data utilization. For this,



(a) Document graph



(b) Random subgraphs



(c) Topic subgraphs

Figure 3: Graph reconstruction and topic forecasting. (a) Time-serial document graph has been constructed using abstracts of blockchain-related papers published between 2017 to 2022. Word count, co-occurrence, and centralities have been calculated by each month. (b) Random subgraphs have been extracted, using the document graph, and split into training/-validation and test time span. Using the time-serial random subgraphs, GCNs have been trained. To train the GCN models, word count and centrality of nodes and co-occurrence data were used, to predict the word count of nodes for the future timeline. (c) Node features and edge weights of the keywords of topics for recent months have been utilized to pre-trained GCNs to predict each corresponding trend.

we employ randomly clustered or selected subgraphs that contain a sufficient number of nodes to cover the entire document graph (Figure 3 (b)).

To construct the random subgraph, we initially extract word nodes using the random walk method, which is a common node sampling technique in graph analysis and machine learning tasks [27, 28] (Figure 3 (b)). The number of nodes in each subgraph is randomly chosen between 8 and 20, as the number of nodes for the topic clustering results ranges from 10 to 15. For random selection of nodes, a seed node is randomly chosen from the document graph and used to construct a primary random node pool. Based on the seed node, another node is appended to the random node pool, ensuring the connection of the newly selected to the random node pool. The randomness of the selection process for a new random node is weighted with the connectivity of the random node pool to the new random node candidates. Then, the selected word-related node annotations and edges features are extracted from the document graph, to reconstruct a time-serial random subgraph. Through the random node selection and time-serial random subgraph reconstruction process, we construct 2,000 random subgraphs for each training, validation, and test dataset. We use early (t_0 to t_m) time epoch data of the time-serial random subgraphs for training and validation dataset, and late time epoch data (t_{m+1} to t_n) for the test dataset, to ensure time-independence between training/validation and test timeline (Figure 3 (b) upper right).

For the time-serial topic subgraphs, we extract each corresponding keyword-related node and edge feature from the document graph. The extracted features are used to reconstruct time-serial topic subgraphs for each topic (Figure 3 (c)). Each time-serial topic subgraphs of the test time span (t_{m+1} to t_n) are used for forecasting based on the pre-trained A3T-GCN.

The features facilitated to A3T-GCN include node features (word count and centralities) and edge weight (co-occurrence) for both random subgraphs and topic graphs on a monthly basis. Since the number of nodes for the subgraphs varies, we impose placeholder nodes to each node pool and impute them with zero values for nodes and null values for related edges.

2.4. Topic Trend Forecasting

In this study, we use the A3T-GCN model that can effectively capture global variation trends by re-weighting the influence of historical information. Our approach involves constructing an A3T-GCN consisting of nodes that represent keywords in the graph. Each node has features that reflect the word count of its corresponding keyword and the keyword's centrality within the graph. The edge weight between nodes is determined by the co-occurrence of keyword pairs. To capture the changes in topic trends, we update the node features, and edge weight on a monthly basis, as changes in keyword word count are assumed to indicate changes in topic trends. By predicting changes in word count using information such as keyword centrality and co-occurrence, our proposed A3T-GCN model offers an effective approach for accurately forecasting topic trends over time. Therefore, our

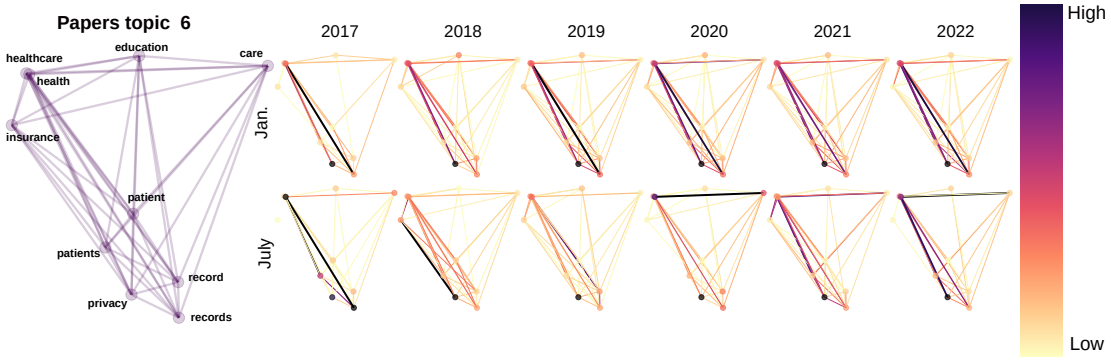
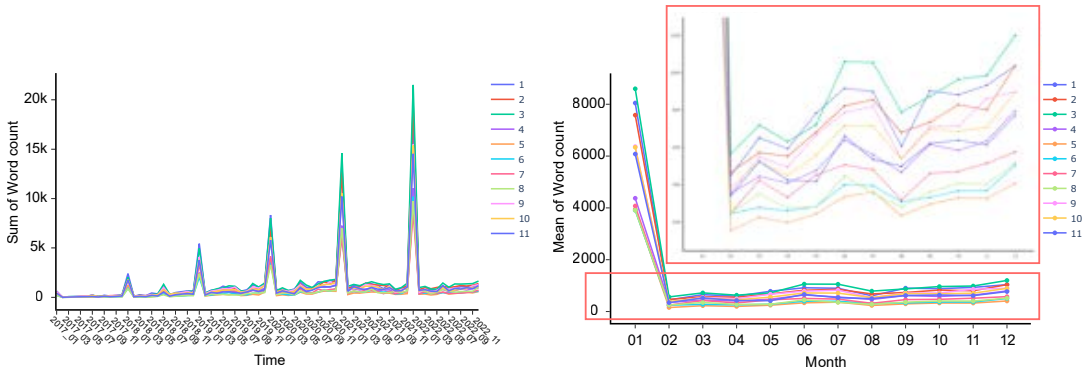


Figure 4: Time-series graph for topic 6 of the paper. Topic keywords, which have been determined using LDA and DMR, were used to extract topic-specific subgraphs for the document graph. As the timeline of the collected data ranged from Jan. of 2017 to Dec. of 2022, 72 monthly time-specific subgraphs were obtained for a topic. (Here, the word count and co-occurrence of the 6th topic in Jan. and Dec. for each year were shown as an example.)



(a) The whole timeline word count for each topic of papers

(b) The average word count by month for paper topics

Figure 5: Seasonality of paper documents. There is seasonality of paper documents, when analyzing word counts of topic keywords. For the whole timeline word count for each topic of papers (a), it showed month-specific trends. Word count in January was remarkably increased for all the topics every year. Also, at the average word count by month for paper topics (b), January of all the paper topics showed the highest word count than the other months.

model is a valuable tool for a wide range of applications.

2.4.1. Training A3T-GCN model

To optimize the A3T-GCN model, we perform feature selection on the node features and conduct hyperparameter optimization. This allows us to identify the most relevant features and tune the model parameters for improved performance. Subsequently, we train individual models to predict word count for future time periods, specifically 1, 3, 6, 9, or 12 months ahead. The training process utilizes random subgraphs, with a fixed training lookback window of 12 months. To facilitate the training and evaluation of the models, the random subgraphs are divided into distinct time steps. The initial 36 months of data are designated for training and validation, while the subsequent 36 months are used for testing. For the training and validation phase, we utilize 2,000 random subgraphs extracted from the first 36 months of the overall dataset. Similarly, for testing, we employ 2,000 random subgraphs from the later 36 months.

2.4.2. Forecasting of the topic

Using the pre-trained models on the random subgraph, we conduct forecasting on the topic graph. The forecasting process involve predicting the outcome for future time periods, specifically 1, 3, 6, 9, and 12 months ahead. This forecasting is carried out using a fixed training lookback window of 12 months, meaning the model used the

past 12 months of topic graph data to make predictions for the future (Figure 3c). To ensure heterogeneity in the time span for the topic forecasting from training or validation, we use later 36 months features of the topics, which is the identical timeline to the test dataset.

3. Experiment Results

3.1. Environments

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].

3.2. Topic Modeling and Clustering

We used the method described in Section 2.2 to determine the optimal number of topics for both LDA and DMR models. The optimal number of topics was found to be 10. After generating topics using both models, we performed clustering as outlined in Figure 2, and the result of the clustering is presented in Table 1. To evaluate the clustering performance, we used the Silhouette Score [29, 30, 31], which is a commonly used method for clustering evaluation. The Silhouette Score obtained for this study was 0.8038, which exceeds the threshold of 0.5, indicating good clustering performance.

3.3. Time-series Graphs and Features

We constructed the document graph using word count and co-occurrence of the data and extracted topic-specific subgraphs from the document graph (Figure 4), and three types of centralities were calculated at each time point. As the timeline of collected data has 72-time points, 72 time-specific subgraphs with word count, co-occurrence, and pre-calculated centralities have been extracted for each topic. As shown in Figure 4, co-occurrence between "health" and "record" was dominant from 2017 to 2019, but gradually decreased. On the other hand, co-occurrences between "health" and "care", and "health" and "privacy" were relatively more dominant in 2020 to 2022. As a result, the structure of co-occurrence for topic graphs seems not static when comparing all the pairs of nodes, but with partial structural movement by time.

We further investigated the month-specific trend to analyze the seasonality of the document (Figure 5). Figure 5 (a) and Figure 5 (b) demonstrate that the word count of all topics exhibited a noticeable increase in January each year compared to other months. However, even when excluding the January papers, we observed elevated word count tendencies in other months such as July and December (Figure 5 (b)). To account for this seasonality, we utilized 12 months of data as the input timespan for A3T-GCN training.

Table 2
The results of the feature selection.

Features			MSE	MAE
BC	CC	DC		
O	X	X	0.01941	0.09639
X	O	X	0.02591	0.11322
X	X	O	0.02092	0.10229
O	**O**	**X**	**0.01764**	**0.09616**
X	O	O	0.01873	0.09704
O	X	O	0.02067	0.10174
O	O	O	0.01826	0.09719

*BC: betweenness centrality; CC: closeness centrality; DC: degree centrality

Table 3
Forecasting results on random subgraphs with optimal A3T-GCN model.

Random Subgraphs		
Forecasting horizon	MSE	MAE
1	0.02091	0.10093
3	**0.0158**	**0.08669**
6	0.02370	0,10147
9	0.03628	0.12595
12	0.04522	0.13420

3.4. Topic Trend Forecasting

We assessed the performance of the A3T-GCN model using two evaluation metrics, Mean Squared Error (MSE) and Mean Absolute Error (MAE) [32].

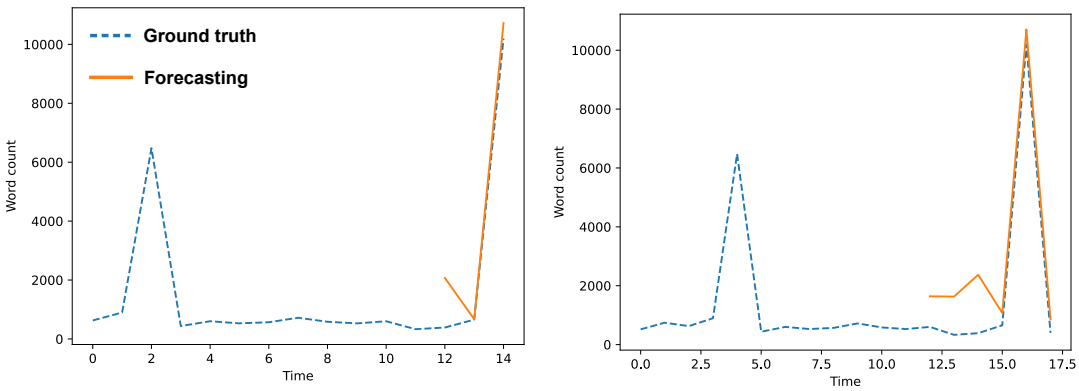
$$MSE = \frac{\sum_{i=1}^{n}(\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2}{n}, \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\frac{\mathbf{Y}_i - \hat{\mathbf{Y}}_i}{\mathbf{Y}_i}|, \tag{2}$$

where $\mathbf{Y}_i$ is the i-th element of $\mathbf{Y}$, n is the number of elements.

3.4.1. Training A3T-GCN model

To optimize the A3T-GCN model, we performed feature selection by trying various node feature combinations. Along with feature selection, we conducted hyperparameter optimization by varying the learning rate with values of 1e-2, 1e-3, and 1e-4. Table 2 displays the results of the feature selection procedure, which were assessed through MSE and MAE. Based on these results, we selected the



(a) Forecasting horizon = 3

(b) Forecasting horizon = 6

(c) Forecasting horizon = 9

(d) Forecasting horizon = 12

Figure 6: The trend forecasting results for Topic 6 with forecasting horizon of (a) 3, (b) 6, (c) 9 and (d) 12.

node feature combination that yielded the lowest MSE value, which included betweenness centrality and closeness centrality as node features.

After conducting feature selection and hyperparameter optimization, we trained the A3T-GCN model to forecast future trends for horizons of 1, 3, 6, 9, and 12 months using a training dataset and validation dataset consisting of 2000 random subgraphs. As mentioned previously, we fixed the training lookback window at 12. The performance of the model was evaluated on a test dataset consisting of 2000 random subgraphs, and the results of the evaluation are presented in Table 3 with the evaluation metrics MSE and MAE.

3.4.2. Forecasting of the topic

We utilized the pre-trained A3T- GCN model to perform topic trend forecasting on topic graphs for each forecasting horizon. Table 4 presents the forecasting results using the evaluation metrics MSE and MAE. As the forecasting horizon increased, the MSE and MAE values also increased, but we observed an exceptional case for the forecasting horizon of 3, which had the lowest MSE and MAE value.

Table 4
Forecasting results on topic graphs with pre-trained A3T-GCN model.

Topic Graphs		
Forecasting horizon	**MSE**	**MAE**
1	0.01342	0.08850
3	**0.00820**	**0.07501**
6	0.01618	0.09025
9	0.02926	0.10925
12	0.03055	0.10695

We present the results of our topic trend forecasting using visualizations that depict the actual and predicted

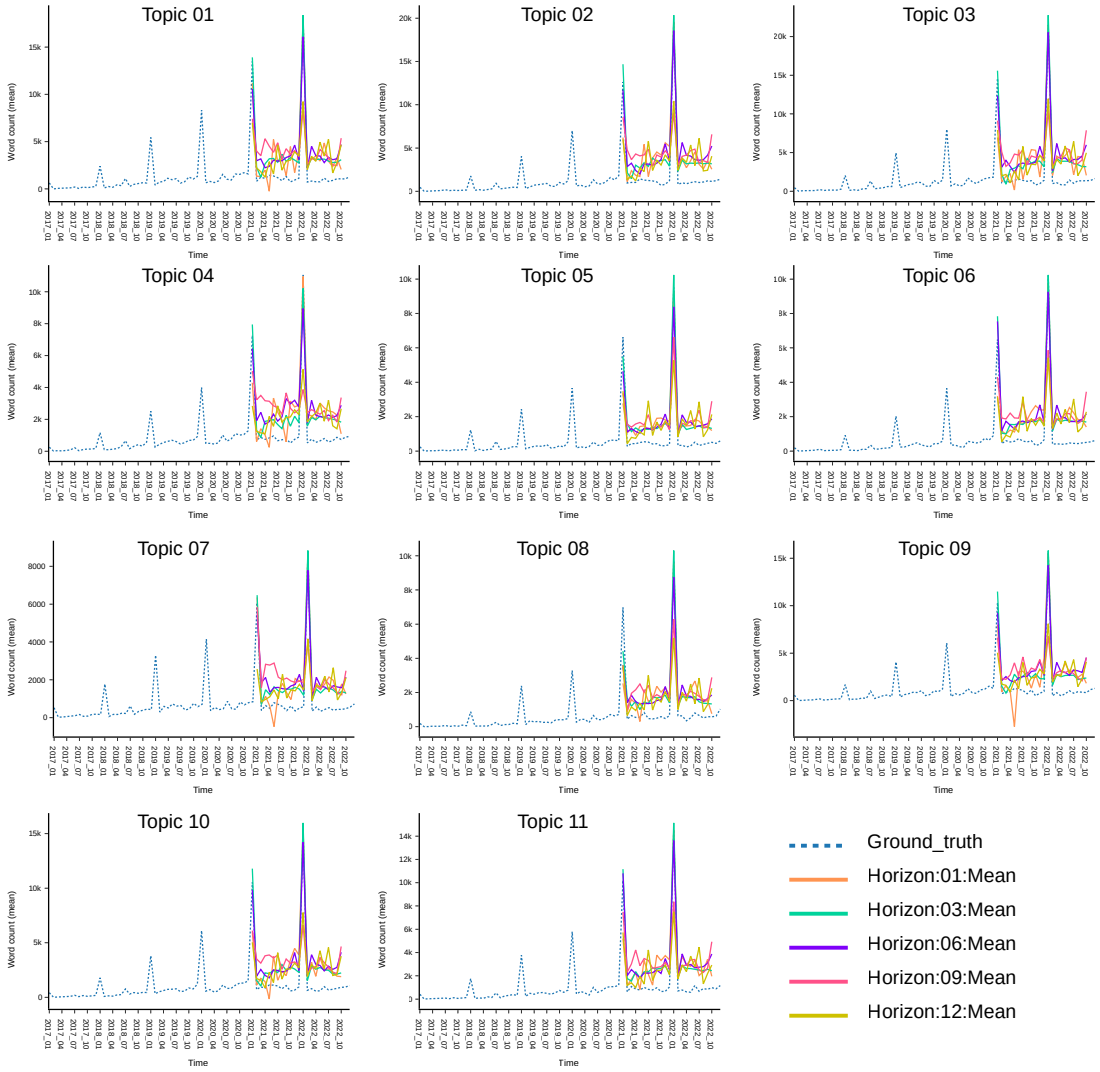


Figure 7: Mean of word count for the topic trend forecasting models by forecasting horizon.

word count of keywords for each topic. The blue line in the graph shows the actual frequency of keywords in the topic, while the orange line represents the predicted word count. Figure 6 shows the results for Topic 6, while Appendix A provides the results for all topics. As illustrated in Figure 6, the predicted line closely matches the ground truth line, indicating the effectiveness of the topic trend forecasting. To provide a more comprehensive understanding of our topic trend forecasting models, we also generated visualizations of the mean word count for each topic across the forecasting horizons, which are presented in Figure 7. The predicted mean word count exhibits similar trends and values to the actual mean word count across all forecasting horizons.

4. Conclusions

In this paper, we propose a novel approach for forecasting future topic trends in the blockchain domain using a combination of topic modeling techniques and graph convolutional networks (GCNs). For the application of our approach to the paper data, GCN model shows great performance on the prediction of topic trend, even if it was trained using random subgraphs of the overall document. The proposed approach addresses the limitations of previous studies by capturing the complex dynamics

of topic trends and the intellectual structure of research fields.

This study shows that paper data have seasonality that can be leveraged for experiments. Our methodology significantly enhances the prediction of topic trends, as demonstrated by experimental results. This approach has implications for researchers, businesses, professionals, and policymakers, as it can provide valuable insights for making informed predictions about the future in the rapidly evolving blockchain field. Although our approach has not been extensively explored in previous studies, our experiments demonstrate its potential for forecasting future topic trends.

Furthermore, we made an attempt to apply our approach to patent data using a pre-trained A3T-GCN model, however, the results did not meet our expectations. As a result, we are currently unable to apply our model to data sources other than academic papers. Our next step involves the design and training of GCN models tailored for forecasting topic trends in patent and news data. We aim to explore various architectural designs and hyperparameters to improve the accuracy and robustness of the models. Moreover, we plan to compare our proposed approach with other state-of-the-art time-series methodologies, including both deep-learning and traditional methods, to demonstrate its effectiveness and superiority in future research. Our ultimate goal is to contribute to the advancement of research in the blockchain domain and related fields by providing a powerful and reliable tool for trend forecasting and analysis. Our proposed approach has the potential to be applied to a wide range of real-world applications, such as financial forecasting, risk management, and market trend analysis.

5. Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2B5B02002359).

References

[1] S. M. H. Bamakan, A. B. Bondarti, P. B. Bondarti, Q. Qu, Blockchain technology forecasting by patent analytics and text mining, Blockchain: Research and Applications 2 (2021) 100019.

[2] Y. Zou, T. Meng, P. Zhang, W. Zhang, H. Li, Focus on blockchain: A comprehensive survey on academic and application, IEEE Access 8 (2020) 187182–187201.

[3] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907 (2016).

[4] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE transactions on neural networks and learning systems 32 (2020) 4–24.

[5] W. Jiang, J. Luo, Graph neural network for traffic forecasting: A survey, Expert Systems with Applications (2022) 117921.

[6] X. Yin, D. Yan, A. Almudaifer, S. Yan, Y. Zhou, Forecasting stock prices using stock correlation graph: A graph convolutional network approach, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.

[7] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, arXiv preprint arXiv:1707.01926 (2017).

[8] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, H. Li, T-gcn: A temporal graph convolutional network for traffic prediction, IEEE transactions on intelligent transportation systems 21 (2019) 3848–3858.

[9] S. Guo, Y. Lin, N. Feng, C. Song, H. Wan, Attention based spatial-temporal graph convolutional networks for traffic flow forecasting, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 922–929.

[10] J. Bai, J. Zhu, Y. Song, L. Zhao, Z. Hou, R. Du, H. Li, A3t-gcn: Attention temporal graph convolutional network for traffic forecasting, ISPRS International Journal of Geo-Information 10 (2021) 485.

[11] M. Xu, J. Du, Z. Xue, Z. Guan, F. Kou, L. Shi, A scientific research topic trend prediction model based on multi-lstm and graph convolutional network, International Journal of Intelligent Systems 37 (2022) 6331–6353.

[12] I. Vayansky, S. A. Kumar, A review of topic modeling methods, Information Systems 94 (2020) 101582.

[13] D. M. Mimno, A. McCallum, Topic models conditioned on arbitrary features with dirichlet-multinomial regression., in: UAI, volume 24, Citeseer, 2008, pp. 411–418.

[14] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[15] F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?, Journal of classification 31 (2014) 274–295.

[16] D. Müllner, Modern hierarchical, agglomerative clustering algorithms, arXiv preprint arXiv:1109.2378 (2011).

[17] H. Kim, H. Park, M. Song, Developing a topic-driven method for interdisciplinarity analysis, Journal of Informetrics 16 (2022) 101255.

[18] K. Porter, Analyzing the darknetmarkets subreddit for evolutions of tools and trends using lda topic

modeling, Digital Investigation 26 (2018) S87–S97.
[19] H. Lee, J. Kwak, M. Song, C. O. Kim, Coherence analysis of research and education using topic modeling, Scientometrics 102 (2015) 1119–1137.
[20] S. Boon-Itt, Y. Skunkan, et al., Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study, JMIR Public Health and Surveillance 6 (2020) e21978.
[21] Y. Fang, Y. Guo, C. Huang, L. Liu, Analyzing and identifying data breaches in underground forums, IEEE Access 7 (2019) 48770–48777.
[22] M. Hasan, A. Rahman, M. R. Karim, M. S. I. Khan, M. J. Islam, Normalized approach to find optimal number of topics in latent dirichlet allocation (lda), in: Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020, Springer, 2021, pp. 341–354.
[23] Q. Xie, X. Zhang, Y. Ding, M. Song, Monolingual and multilingual topic analysis using lda and bert embeddings, Journal of Informetrics 14 (2020) 101055.
[24] S. Sia, A. Dalmia, S. J. Mielke, Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!, arXiv preprint arXiv:2004.14914 (2020).
[25] D. Miller, Leveraging bert for extractive text summarization on lectures, arXiv preprint arXiv:1906.04165 (2019).
[26] J. Zhang, Y. Luo, Degree centrality, betweenness centrality, and closeness centrality in social network, in: 2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017), Atlantis press, 2017, pp. 300–303.
[27] J. D. Noh, H. Rieger, Random walks on complex networks, Physical review letters 92 (2004) 118701.
[28] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, IEEE Transactions on knowledge and data engineering 19 (2007) 355–369.
[29] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.
[30] K. R. Shahapure, C. Nicholas, Cluster quality analysis using silhouette score, in: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), IEEE, 2020, pp. 747–748.
[31] G. Ogbuabor, F. Ugwoke, Clustering algorithm for a healthcare dataset using silhouette score value, Int. J. Comput. Sci. Inf. Technol 10 (2018) 27–37.
[32] A. Jadon, A. Patil, S. Jadon, A comprehensive survey of regression based loss functions for time series forecasting, arXiv preprint arXiv:2211.02989 (2022).

A. Visualization of the topic trend forecasting Results by forecasting horizon for Each Topic