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

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023)

Park, Y., Lim, S., Gu, C., & Song, M. (2023)

# Table of Contents

# Introduction

- In modern society, **keeping up with recent trends is crucial** for individuals and organizations to succeed, but it can be a difficult task.
- **Time series forecasting** has emerged as a promising approach, enabling the analysis of temporal data and the **discovery of patterns and trends**.
- Leveraging historical and current trends enables individuals and organizations to **gain a competitive advantage in the market**, particularly within the blockchain domain by making informed future predictions.

# Introduction

- Deep learning methods such as **LSTM and GCNs have shown promise for topic trend predictions** by capturing the temporal and structural dependencies between topics.

- However, **the integration of topic modeling into these models remains unexplored**, preventing a comprehensive grasp of the research field's main themes and sub-topics.

- We propose a new approach that **integrates topic modeling and GCNs to predict upcoming topic trends** within the **blockchain field**.

# Overall Workflow

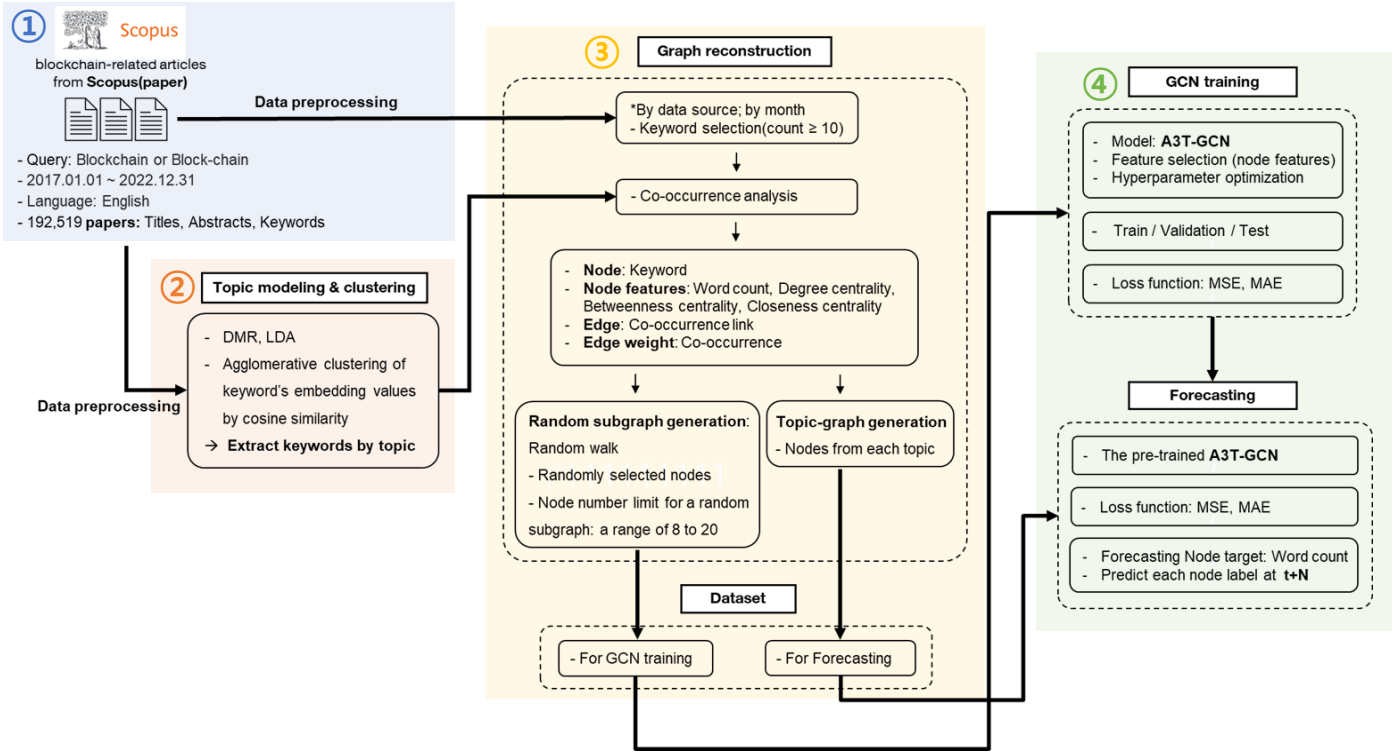



**Figure 1. The Overall Schematic Research Workflow**

# **Methodology:** ① Data Preparation

- Total of **192,519** papers were collected:
  - Target text: **Titles, Abstracts, Keywords** of research papers
  - Period: January 1st, 2017 ~ December 31st, 2022 (**72-month**)
  - Search query: **"Blockchain or Block-chain"** (in Scopus database)

- Based on this, we perform several preprocessing steps:

1) Convert all text to lowercase
2) Divide the text into sentence units using NLTK
3) Tokenize sentences into words
4) Employ NLTK for part-of-speech tagging
5) Retain words tagged as nouns(unigram)
6) Filter out stopwords (commonly used, meaningless, and major topic words)

→ **In this process, focus on relevant and meaningful terms.**

# **Methodology:** ② Topic Modeling and Clustering

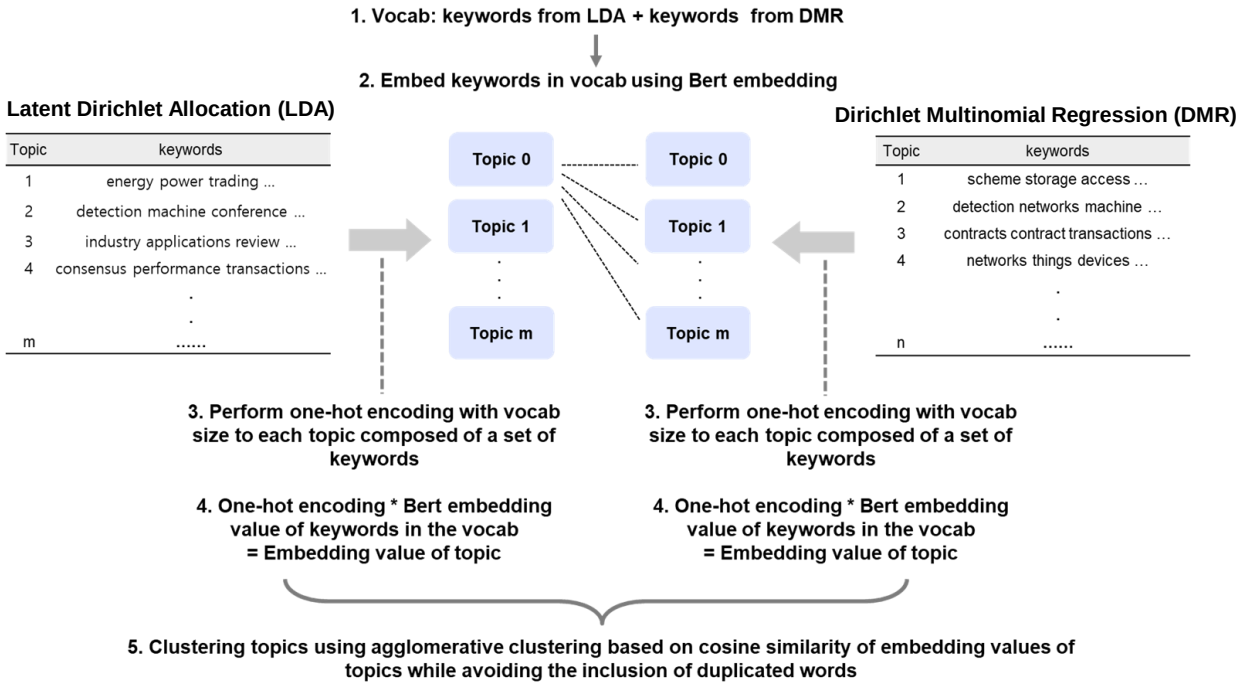



**Figure 2. Topic clustering using agglomerative clustering based on the cosine similarity of their embedding values.**

# Methodology: ③ Graph Reconstruction

## Time Serial Document Graph

- Every node and edge in the graph are annotated with their corresponding **monthly-based** features.
  **(72 time points → 72 time-specific graphs)**

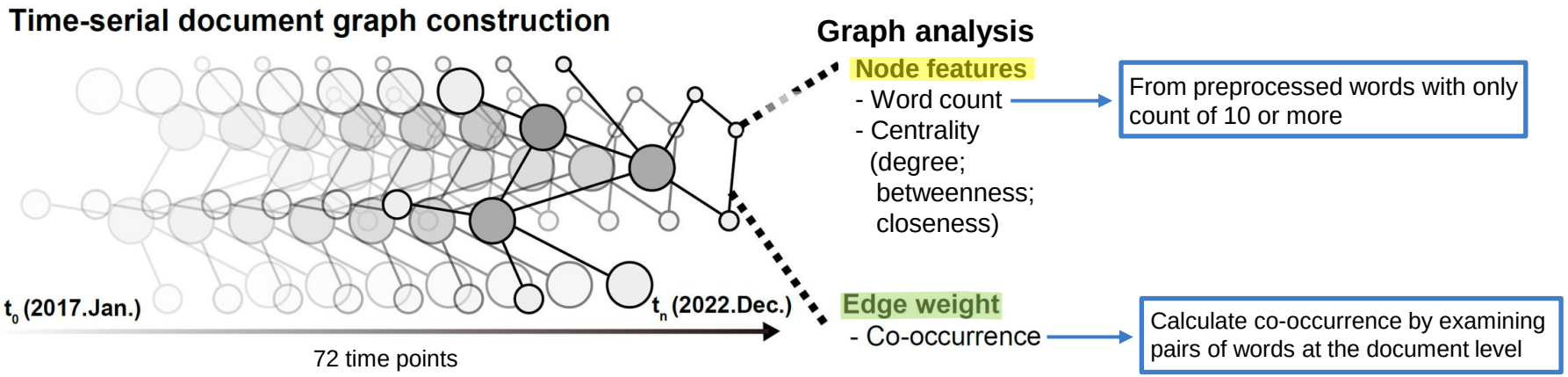



**Figure 3. (a) Document graph**

# Methodology: ③ Graph Reconstruction

- **Time Serial Random Subgraph**
  - Construct random subgraphs using the **random walk** method.
  - Employ random subgraphs for **training GCNs**.

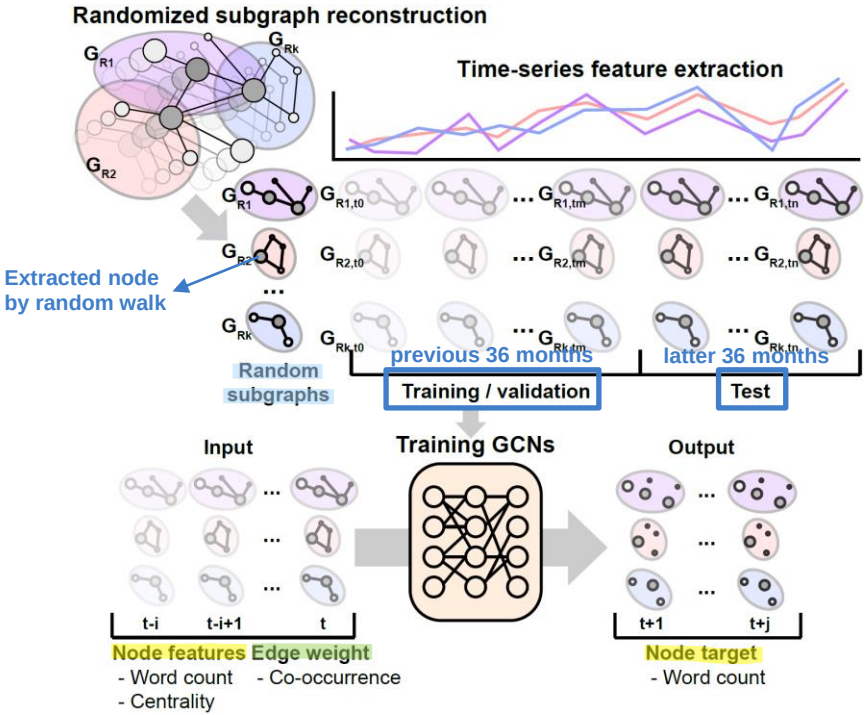



**Figure 3. (b) Random subgraphs**

# Methodology: ③ Graph Reconstruction

- **Time Serial Topic Subgraph**
  - Node: **topic keyword** (from LDA, DMR)
  - Employ topic subgraphs on a pre-trained GCN model for **forecasting.**

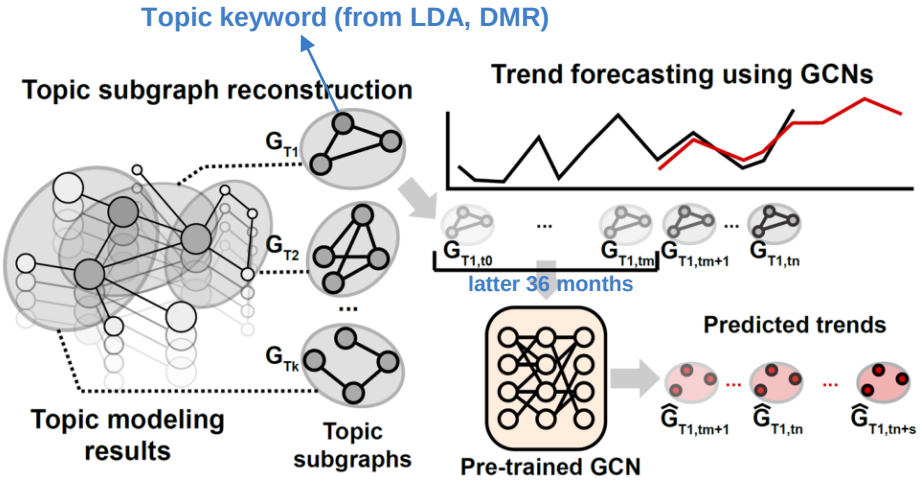



**Figure 3. (c) Topic subgraphs**

# **Methodology:** ④ Topic Trend Forecasting

## A3T-GCN Model

- **A3T-GCN** model effectively captures global variation trends by re-weighting historical information.
- To capture the changes in topic trends, we update the node features, and edge weight on a monthly basis when constructing the subgraphs.

  → Changes in the node's word count are assumed to indicate changes in topic trends.
- **Predict** changes in **word count(Node target)** using information such as centrality and co-occurrence.

# Methodology: ④ Topic Trend Forecasting

## Training A3T-GCN Model

- **Feature selection** on the node features and **hyperparameter optimization**.
- **Train** the **A3T-GCN model** using **random subgraphs** with a fixed lookback window of 12 months to predict **word count (Node target)** for future time periods (1, 3, 6, 9, or 12 months ahead).

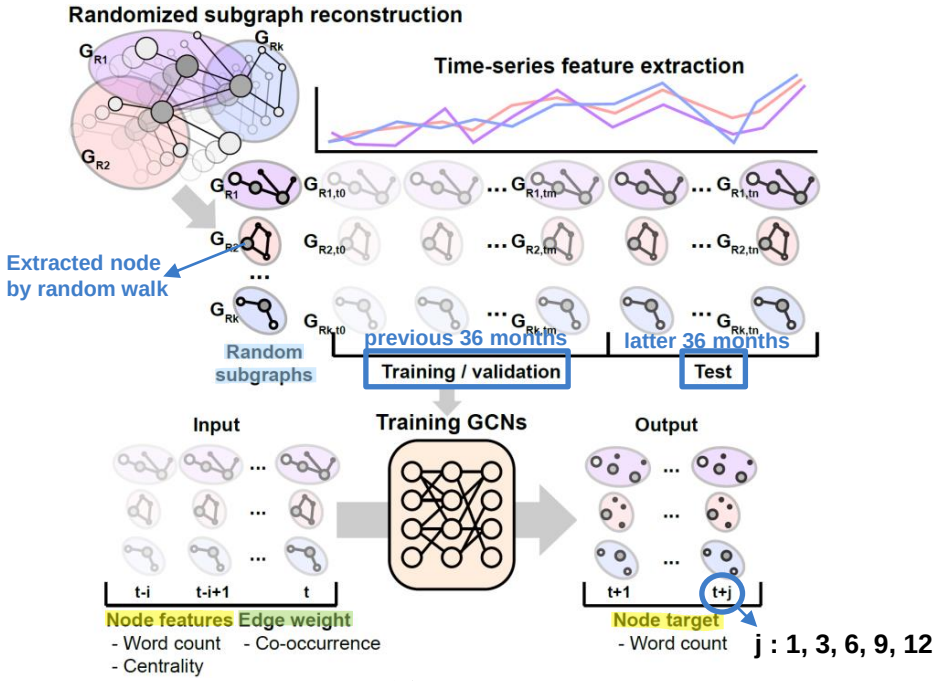



**Figure 3. (b) Random subgraphs**

# Methodology: ④ Topic Trend Forecasting

## Forecasting

- Employ **topic subgraphs** on a pre-trained A3T-GCN model for **forecasting** to predict **word count (Node target)** for future time periods (1, 3, 6, 9, or 12 months ahead).

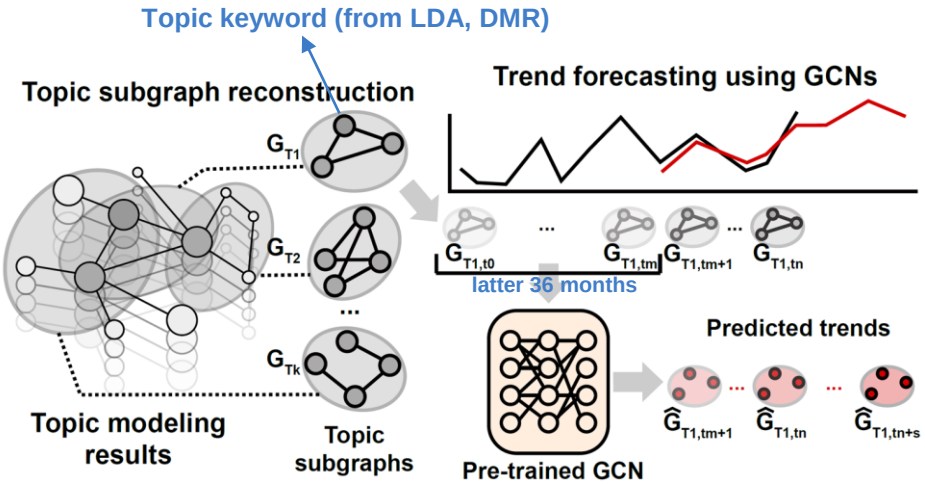



**Figure 3. (c) Topic subgraphs**

# **Results:** Topic Clustering

- This is the result of the clustering process for topic modeling.
- Evaluation Metric: Silhouette Score (Score was **0.8038** which represents good performance)

**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 |

# **Results:** Time-series graphs and feature

- As the timeline of collected data has 72-time points, 72 time-specific subgraphs with word count, co-occurrence, and centralities have been extracted for each topic.

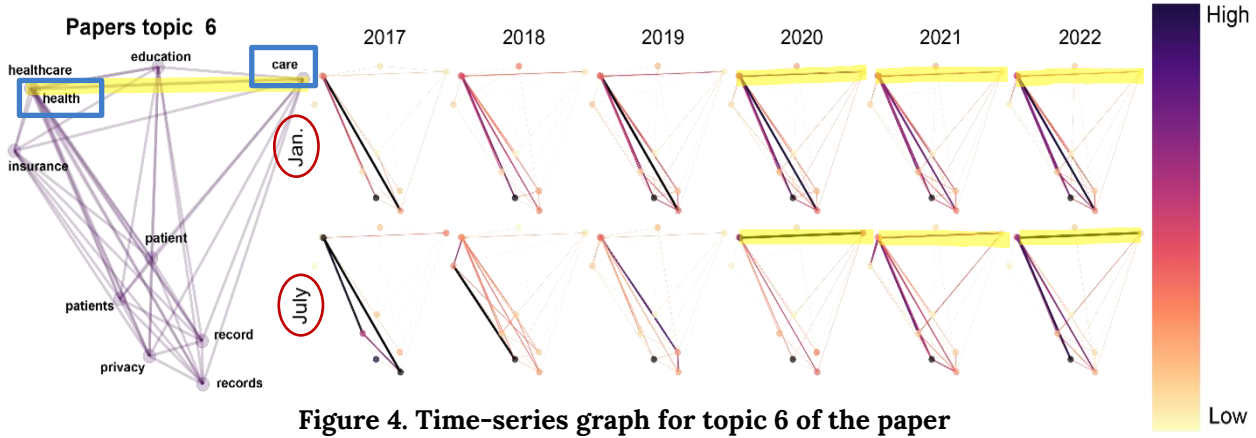



**Figure 4. Time-series graph for topic 6 of the paper**

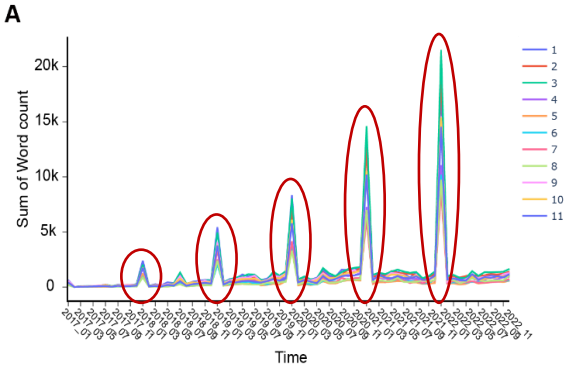
→ As a result, the structure of co-occurrence for topic subgraphs seems dynamic.

# **Results:** Time-series graphs and feature

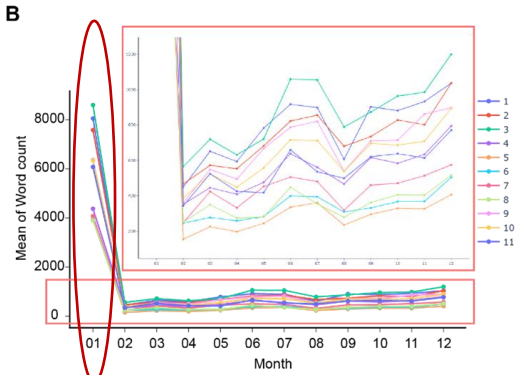
- Seasonality of paper documents, when analyzing word counts of topic keywords.

  → (a), (b) showed **month-specific trends.**

  → Word count in **January** was remarkably increased for all the topics every year.




(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**

- Evaluation metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE).

The lowest MSE, MAE

# **Results:** Topic trend forecasting

**Feature selection**
(betweenness, closeness)

→

**Pre-train A3T-GCN**
(random subgraphs)

→

**Forecasting**
(topic subgraphs)

**Table 2. The results of the feature selection**

| Features | | | MSE | MAE |
|---|---|---|---|---|
| betweenness | closeness | degree | | |
| 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 |

**Table 3. Forecasting Results on Random Subgraphs(test dataset) with optimal A3T-GCN model**

| Random Subgraphs | | |
|---|---|---|
| Forecasting horizon | MSE | MAE |
| 1 | 0.02091 | 0.10093 |
| **3** | **0.01580** | **0.08669** |
| 6 | 0.02370 | 0,10147 |
| 9 | 0.03628 | 0.12595 |
| 12 | 0.04522 | 0.13420 |

**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 |

# **Results:** Topic trend forecasting

- The **blue line** in the graph shows the **actual frequency** of keywords in the topic, while the **orange line** represents the **predicted word count**.
- The **predicted line** closely matches the **ground truth line**, indicating the effectiveness of the topic trend forecasting.

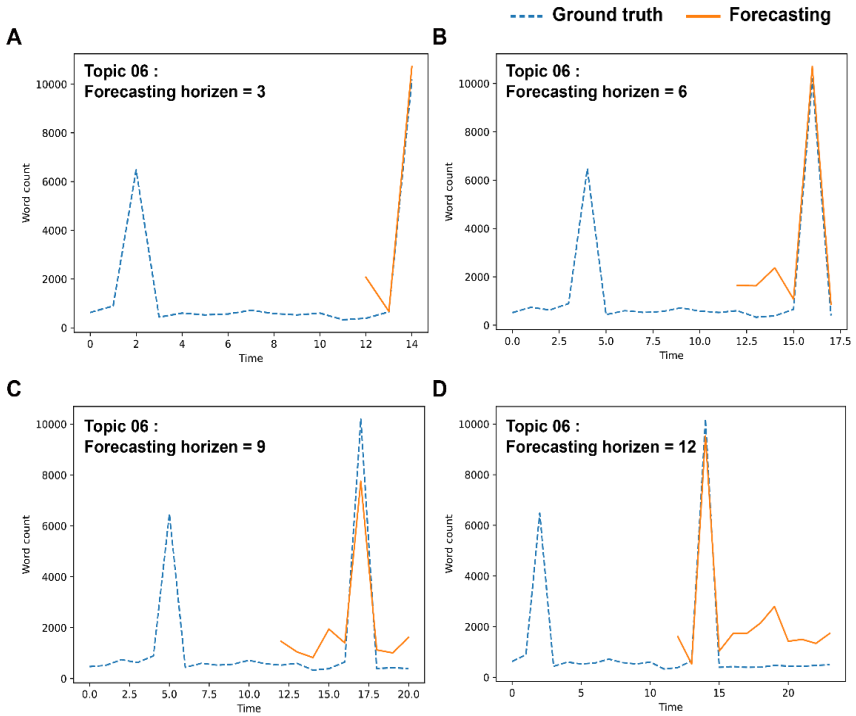



**Figure 6. The trend forecasting results for Topic 6.**

# **Results:** Topic trend forecasting

- The **predicted mean word count** exhibits **similar trends** and values to the **actual mean word count** across all forecasting horizons.

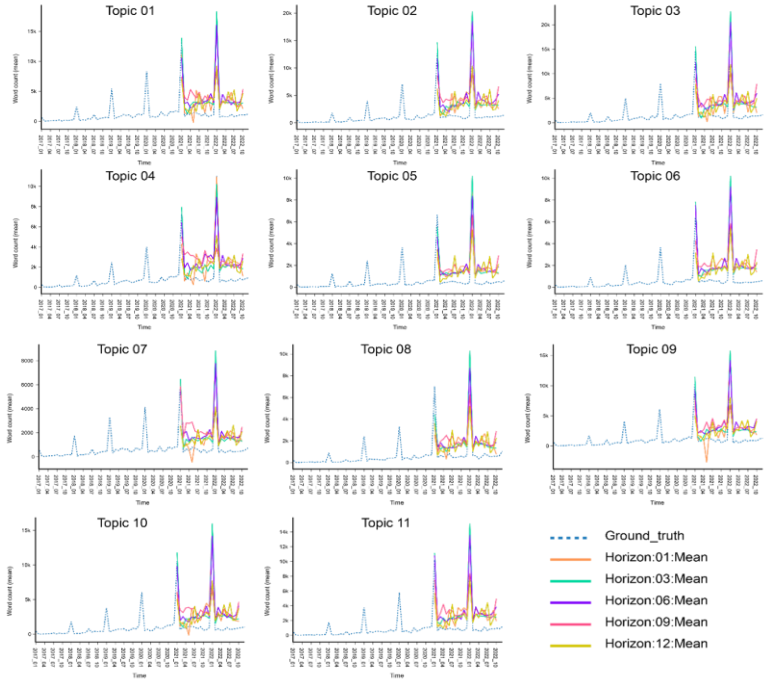



**Figure 7. Mean of word count for the topic trend forecasting models by forecasting horizon.**

# Conclusion

- 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).**

- GCN model shows **great performance in the prediction** of topic trends, even if it was trained using random subgraphs of the overall document.

- Our proposed 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** by providing a powerful and reliable tool for trend forecasting and analysis.

# Conclusion

- Limitation and future work

1. Currently, our model is limited to the application of academic papers and cannot be extended to other data sources.
   → Our next step involves the design and training of GCN models tailored for forecasting topic trends in patent and news data.

2. Apart from A3T-GCN, we did not include experiments comparing our model with other models.
   → 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.

# References


[1] Seyed Mojtaba Hosseini Bamakan, Alireza Babaei Bondarti, Parinaz Babaei Bondarti, and Qiang Qu. 2021. Blockchain technology forecasting by patent analytics and text mining. *Blockchain: Research and Applications* 2, 2 (2021), 100019. DOI:https://doi.org/10.1016/j.bcra.2021.100019

[2] Yijun Zou, Ting Meng, Peng Zhang, Wenzhen Zhang, and Huiyang Li. 2020. Focus on blockchain: A comprehensive survey on academic and application. *IEEE Access* 8, (2020), 187182–187201. DOI:https://doi.org/10.1109/ACCESS.2020.3030491

[3] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016).

[4] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32, 1 (2020), 4–24. DOI:https://doi.org/10.1109/TNNLS.2020.2978386

[5] Weiwei Jiang and Jiayun Luo. 2022. Graph neural network for traffic forecasting: A survey. Expert Syst Appl (2022), 117921. DOI:https://doi.org/10.1016/j.eswa.2022.117921

[6] Xingkun Yin, Da Yan, Abdullateef Almudaifer, Sibo Yan, and Yang Zhou. 2021. Forecasting stock prices using stock correlation graph: A graph convolutional network approach. In 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 1–8. DOI:https://doi.org/10.1109/IJCNN52387.2021.9533510

[7] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926 (2017).

[8] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: A temporal graph convolutional network for traffic prediction. IEEE transactions on intelligent transportation systems 21, 9 (2019), 3848–3858. DOI:https://doi.org/10.1109/TITS.2019.2935152

[9] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In Proceedings of the AAAI conference on artificial intelligence, 922–929. DOI:https://doi.org/10.1609/aaai.v33i01.3301922

[10] Jiandong Bai, Jiawei Zhu, Yujiao Song, Ling Zhao, Zhixiang Hou, Ronghua Du, and Haifeng Li. 2021. A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. ISPRS Int J Geoinf 10, 7 (2021), 485. DOI:https://doi.org/10.3390/ijgi10070485

[11] Mingying Xu, Junping Du, Zhe Xue, Zeli Guan, Feifei Kou, and Lei Shi. 2022. A scientific research topic trend prediction model based on multi-LSTM and graph convolutional network. International Journal of Intelligent Systems 37, 9 (2022), 6331–6353. DOI:https://doi.org/10.1002/int.22846

[12] Ike Vayansky and Sathish A P Kumar. 2020. A review of topic modeling methods. Inf Syst 94, (2020), 101582. DOI:https://doi.org/10.1016/j.is.2020.101582

[13] David Mimno and Andrew McCallum. 2012. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv preprint arXiv:1206.3278 (2012).

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022.

[15] Fionn Murtagh and Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J Classif 31, (2014), 274–295. DOI:https://doi.org/10.1007/s00357-014-9161-z

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

[17] Hyeyoung Kim, Hyelin Park, and Min Song. 2022. Developing a topic-driven method for interdisciplinarity analysis. J Informetr 16, 2 (2022), 101255. DOI:https://doi.org/10.1016/j.joi.2022.101255

[18] Kyle Porter. 2018. Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling. Digit Investig 26, (2018), S87–S97. DOI:https://doi.org/10.1016/j.diin.2018.04.023

[19] Hoyeop Lee, Jueun Kwak, Min Song, and Chang Ouk Kim. 2015. Coherence analysis of research and education using topic modeling. Scientometrics 102, (2015), 1119–1137. DOI:https://doi.org/10.1007/s11192-014-1453-x

[20] Sakun Boon-Itt and Yukolpat Skunkan. 2020. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 6, 4 (2020), e21978. DOI:https://doi.org/10.2196/21978

[21] Yong Fang, Yusong Guo, Cheng Huang, and Liang Liu. 2019. Analyzing and identifying data breaches in underground forums. IEEE Access 7, (2019), 48770–48777. DOI:https://doi.org/10.1109/ACCESS.2019.2910229

[22] Mahedi Hasan, Anichur Rahman, Md Razaul Karim, Md Saikat Islam Khan, and Md Jahidul Islam. 2021. 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, 341–354. DOI:https://doi.org/10.1007/978-981-33-4673-4_27

[23] Qing Xie, Xinyuan Zhang, Ying Ding, and Min Song. 2020. Monolingual and multilingual topic analysis using LDA and BERT embeddings. J Informetr 14, 3 (2020), 101055. DOI:https://doi.org/10.1016/j.joi.2020.101055

[24] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! arXiv preprint arXiv:2004.14914 (2020).

[25] Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. arXiv preprint arXiv:1906.04165 (2019).

[26] Junlong Zhang and Yu Luo. 2017. Degree centrality, betweenness centrality, and closeness centrality in social network. In 2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017), Atlantis press, 300–303. DOI:https://doi.org/10.2991/msam-17.2017.68

[27] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[28] Jae Dong Noh and Heiko Rieger. 2004. Random walks on complex networks. Phys Rev Lett 92, 11 (2004), 118701. DOI:https://doi.org/10.1103/PhysRevLett.92.118701

[29] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. 2007. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Trans Knowl Data Eng 19, 3 (2007), 355–369. DOI:https://doi.org/10.1109/TKDE.2007.46

[30] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20, (1987), 53–65. DOI:https://doi.org/10.1016/0377-0427(87)90125-7

[31] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), IEEE, 747–748. DOI:https://doi.org/10.1109/DSAA49011.2020.00096

[32] Godwin Ogbuabor and F N Ugwoke. 2018. Clustering algorithm for a healthcare dataset using silhouette score value. Int. J. Comput. Sci. Inf. Technol 10, 2 (2018), 27–37. DOI:https://doi.org/10.5121/ijcsit.2018.10203

[33] Aryan Jadon, Avinash Patil, and Shruti Jadon. 2022. A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting. arXiv preprint arXiv:2211.02989 (2022).

# Thank You

# Q & A