IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol. 003, LIIR077
DOI: https://doi.org/10.52731/liir.v003.077

Unsupervised Detection of Domain Switching in Thai Multidisciplinary Online News

Chotanansub Sophaken [*], Kantapong Vongpanich [*],
Akkharawoot Takhom [†], Prachya Boonkwan [†],
Thepchai Supnithi [†]

Abstract

Electronic news has become a popular method of keeping up with digital information, where news tracking is more accessible and reaches a broader variety of audiences. However, ambiguous contexts can be an obstacle for news consumption, causing online disputes, cyberbullying, and political radicalization. This paper demonstrates a network text analysis with a generative statistical model called Latent Dirichlet allocation to extract terminologies and generate a co-occurrence network across multidisciplinary knowledge. The network points out that each terminology corresponds to different domains which are to recognize interpretations of news readers.

Keywords: Latent Dirichlet Allocation, Network Text Analysis, Natural Language Processing, Multidisciplinary Knowledge.

1 Introduction

Electronic news has become one of the most popular media during the past three decades. With the emergence of the Internet, online news sources offer minute-by-minute information and multimedia streaming, e.g., sound recording, video clips, and animations, which cannot be presented in traditional paper-based newspapers. In Thailand, there are 48.59 million Internet users (January 2021) [https://datareportal.com/reports/digital-2021-thailand], most of which actively consuming online news on various platforms; e.g., news agent websites, social media, and news threads.

However, with an abridged edition time, multiple knowledge fields may have to be coherently integrated in a single news article. That puts the burden onto the audience, as they must recognize the context switching regularly. The gist of information is oftentimes distorted in multidisciplinary documents due to context misunderstanding causing linguistic ambiguity potentially leading to online disputes, cyberbullying, and political radicalization.

To overcome this issue, there are two problems to address: (1) how to determine the context automatically in a large pool of texts and (2) how to understand the contextual connection among different of the documents. In the first problem, unsupervised clustering can be applied to each text chunk to explore the knowledge domains. In the latter problem, semantic interpretation must

[*] King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[†] National Electronics and Computer Technology Center, Pathum Thani, Thailand

be applied to multidisciplinary documents, so that the connection of words can be visualized for human interpretation.

In this paper, the authors present a statistical model that recognizes context switching in multidisciplinary documents in an unsupervised fashion. This model is composed of two components: clustering module and word-networking module. The first component is based on the Latent Dirichlet Allocation (LDA) [2], whereby text chunks are categorized into domains based on the distribution of words w.r.t. the domains using unsupervised learning. The latter component is based on Network Text Analysis (NTA) [1], whereby connection of keywords is analyzed and visualized as a graph of words (GoW).

In recent literature, LDA models were applied to enhance the NTA approach, for instance, Takhom et al. [3] and Vaz et al. [4] raised multidisciplinary contexts. On the other hand, this paper demonstrates the LDA model to analyze multidisciplinarity in Thai electronic news. The authors aim to identify multiple domains of a topic modeling technique and represent a relationship within multidisciplinary contexts through a co-occurrence network.

The rest of the paper is organized as follows: Section (2) describes methodology and network text analysis workflow. Section (3) shows study case on Thai electronic news. Section (4) explains quantitative contributions of co-occurrence network and metrics used to evaluate the model. Section (5) gives elaboration on the result given from section (4). Section (6) discusses about related work and contrast from other papers. Section (7) is a conclusion to overall research, including future works.

2 A Cross-domain Statistical Model

A cross-domain statistical model concentrates on reducing ambiguity and identifying the topic of the corpus pairs in multidisciplinary contexts by using LDA to cluster corpus into a determined number of domains and visualize the result using NTA approaches. The NTA workflow has six phases, described as follows:

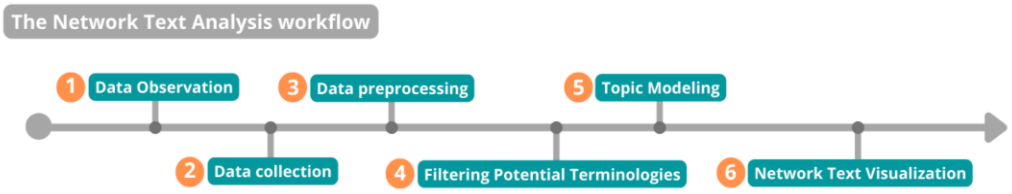


Figure 1: A cross-domain statistical model has six phases of the NTA workflow

Data observation (Phase 1) is a survey and type selection with data sources used in the analysis and determining the scope and features of the data to suit the objective. Proper features of the data will positively influence the performance of the model. In this project, the data type is plain text from electronic news articles. As illustrated in Figure 2, electronic news has four typical components: (a) uniform resource locator, commonly known as URL, (b) Headline (c) publisher information consisting of publisher's name, date of issue, and specific domain, and (d) news content.

Figure 2: Four typical components of electronic news.

Data collection (**Phase 2**) is the process of scrapping and collecting data from determined sources, and then storing them in a format that corresponds with further exploratory analysis, model evaluation, and feature extraction. In this project, the authors extracted contents of the Thai electronic news article from reliable sources using a web data extraction technique commonly known as web scraping, then stored them in a tabular format containing news content, domain, keywords, publish date, and source.

Data preprocessing (**Phase 3**) is the process of cleaning and preparing collected data in better condition and a more efficient format for further analysis. Divided into two sub-processes: (1) data cleaning is to screen the data by removing insignificant or unreliable data to optimize the quantity of the data, making corrections or fulfill missing data, to be more efficient on collected data and further improve the reliability of the analysis, and (2) data formatting is to transform the data into the most appropriate and efficient format for machine analysis. For this project, the data transforms into formats that are most appropriate for each different technique.

Filtering potential terminologies (**Phase 4**) is the process of determining the keywords, a terminology that has a high influence on the domain of the article, according to the specified feature or metrics. This project will filter the keywords by mainly focusing on frequency range - inverse frequency, resulting in a list of keywords that possess the potential for creating a bag-of-words, and furthermore, an important feature in topic modeling

Topic Modeling (**Phase 5**) is to classify terms according to related domains using a topic model. A cross-domain statistical model in this paper has chosen LDA in lexical data segmentation through the reference list of terminologies and features from the bag-of-words.

Network Text Visualization (**Phase 6**), Visualizing the data in the form of Network Text Analysis, by connecting the relationship in each terminology with different terms and attributes. In this project, the network of co-occurrence is represented.

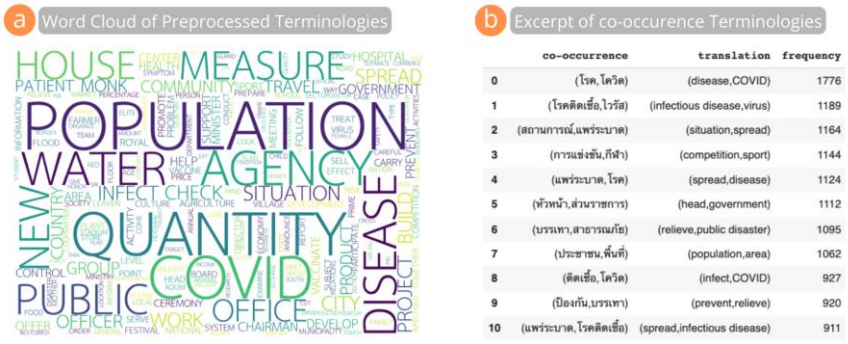
3 Case Study

Our case study is to apply method mentioned in section 2 to Thai news articles.

Phase 1, the authors decided to use data from electronic news articles in 2021 (Jan - Dec) from reliable source, Open Government Data of Thailand (http://data.go.th) which possessed quality, quantity, non-biased, and variety of domains to be collected.

Phase 2, the database only provided URL of the news articles, so web scraping techniques are used to extract the news content. After conducting exploratory data analysis, 8,981 articles are collected, estimates of article length is from 100 to 500 words long.

Phase 3, the collected data are processed in two sub-processes as follows: (1) Segment content from each article into a list using tokenizer. Then, insignificant terminology will be removed by NLP techniques as follows: stop-word removal, part-of-speech tagging and term filtering. (2) Applying bigram detection to the filtered corpus list from sub-processed (1) resulting in a list of co-occurrence terminology with cumulative frequency.



	co-occurrence	translation	frequency
0	(โรค,โควิด)	(disease,COVID)	1776
1	(โรคติดเชื้อ,ไวรัส)	(infectious disease,virus)	1189
2	(สถานการณ์,แพร่ระบาด)	(situation,spread)	1164
3	(การแข่งขัน,กีฬา)	(competition,sport)	1144
4	(แพร่ระบาด,โรค)	(spread,disease)	1124
5	(หัวหน้า,ส่วนราชการ)	(head,government)	1112
6	(บรรเทา,สาธารณภัย)	(relieve,public disaster)	1095
7	(ประชาชน,พื้นที่)	(population,area)	1062
8	(ติดเชื้อ,โควิด)	(infect,COVID)	927
9	(ป้องกัน,บรรเทา)	(prevent,relieve)	920
10	(แพร่ระบาด,โรคติดเชื้อ)	(spread,infectious disease)	911

Figure 3: (a) word cloud of preprocessed terminologies (b) excerpt of co-occurrence terminologies

Phase 4, contribute bag of words shows in Figure 4 (a) by applying TF-IDF vectorizer to the corpus lists from Phase 3, sub-process (1) to extract potential terminology from each article, resulting in a bag of words shows in Figure 4 (b).

Figure 4: (a) process of bag-of-words contribution (b) excerpt of bag-of-words table

Phase 5, contribute topic modeling using LDA on corpus lists in Phase 3, sub-process (1), using bag of words from Phase 4 to create N-by-k feature matrix shows in Figure 5, resulting in corpus occurring probability in each clustered domain.

Figure 5: The processes of contribution to the feature table of the LDA model

Phase 6, visualize as network text where nodes and its size represent terminologies and degree centrality, the edges and its width represent co-occurrence between nodes and frequency of co-occurrence. Result shows in Figure 6.

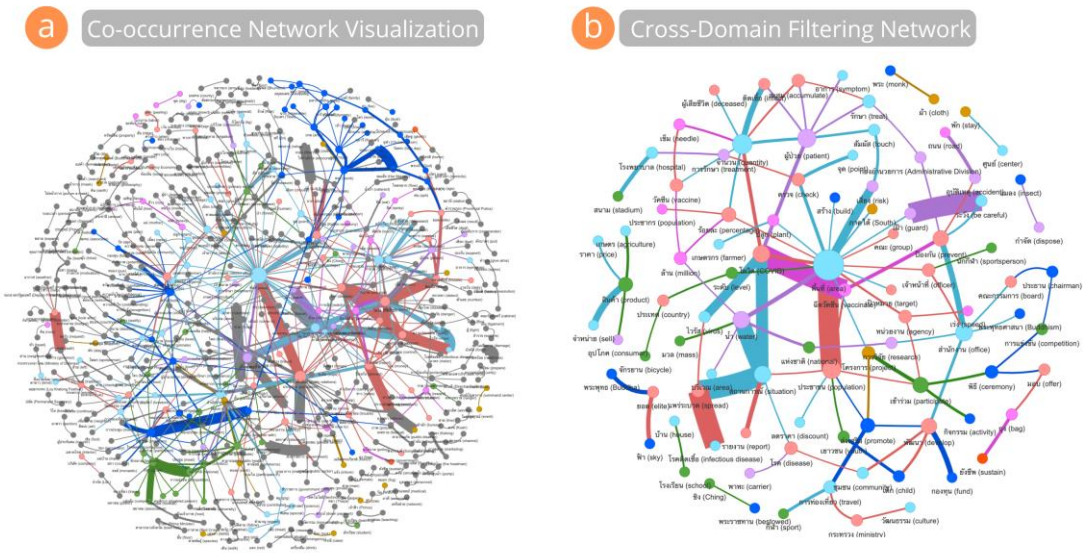


Figure 6: (a) co-occurrence network visualization (b) filtering cross-domain in co-occurrence network

4 Quantitative Contribution of Co-occurrence Network

The probabilistic calculation for quantitative analysis contains the following relevant theories. Dirichlet distribution is the calculation of the proportion and distribution density of the data. can also describe the relationship that occurs with the probability density function. as quantitative evaluation of topic modeling. The log probability of term-topic distribution [5] as determined from Equation (1):

$$\log \boldsymbol{Prob}(\boldsymbol{w}) = \boldsymbol{log}\frac{\boldsymbol{P}(\boldsymbol{T}|\boldsymbol{w})}{\boldsymbol{P}(\boldsymbol{T})} \tag{1}$$

Within this project, the focus was on presenting a network of co-occurrence terminology. It is a network that shows the relationship of terminologies that are used or appear together in a sentence in a significant way. The factors that determine the characteristics of the network are common occurrence frequencies of pairs, network centrality, and domains. data visualization in network graphs. For quantitative evaluation, the network centrality measure [6] can be used to rank nodes and edges, a cross-domain in co-occurrence relation that represent via edge. Edges are ranked according to the edge-betweenness centrality which can be calculated by Equation (2):

$$\mathrm{c_B(e)} = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)} \tag{2}$$

5 Result and Discussion

5.1 Results of Topic Modeling using LDA

Table 1. Result of topic modeling using LDA. P(T): Probability of topic distributions

Topic	Color	Top-5 most relevant terminologies	$P(T)$
1	Red	วัคซีน (Vaccine), ผู้เสียชีวิต(Deceased), ติดเชื้อ (Infect), โควิด(COVID), อ้าง (Claim)	0.171
2	Cyan	หมู (Pig),เกษตรกร (Farmer), ราคา (Price), เนื้อ (Meat), อาหาร (Food)	0.151
3	Blue	พระ (Monk), บรม (Arch), เสด็จ (Come), ถวาย (Give), ราช (Royal)	0.136
4	Yellow	ติดเชื้อ (Infect), ดิน (Dirt), สถานประกอบการ (Establishment), เตือนภัย (Alarm), ใช้จ่าย (Spend)	0.124
5	Green	กีฬา (Sport), การแข่งขัน (Competition), เกม (Game), นักกีฬา(Athelete), ฟุตบอล (Football),	0.119
6	Purple	น้ำ (Water), วิทยาศาสตร์ (Science), นวัตกรรม (Innovation), คลัง(Storage), คาดการณ์ (Predict)	0.116
7	Pink	ฉีดวัคซีน (Vaccination), ทางหลวง (Highway), ผู้โดยสาร(Passenger), ประชากร (Population), ขับรถ (Driving)	0.114
8	Orange	การเลือกตั้ง (Election), สมาชิกสภา (Council members), การคัดเลือก(Selection), วิ่ง (Run), ผู้แทนราษฎร (Citizen representative)	0.069

As shown in table 1, each terminology is clustered into domains and represented by a distinct color seen in Figure 6. The results from intertopic distance map generated based on LDAvis [7] in Figure 7, shows that multidisciplinarity commonly appear across the corpus.

However, corpus clustered in Topic 3 and Topic 5 are completely isolated from other topics. This this reasonable considering Topic 5 has a great portion of terminologies relate to sports which distinctively use in news articles that particularly associate with sports. Similarly, Topic 3 has a great portion of terminologies originate from Bali (पालि), Sanskrit (पाळि) and Khmer (ខ្មែរ) which commonly use in news article that significantly associate with Thai Buddhist or Thai royals, as shown in Table 1.

Nonetheless, by retrieving root terminology of Topic 3 from Thai WordNet [8]. The authors found that a great portion of retrieved terminology has appeared in other topics. Therefore, corpus clustering can be affected by a particular language-style [9].

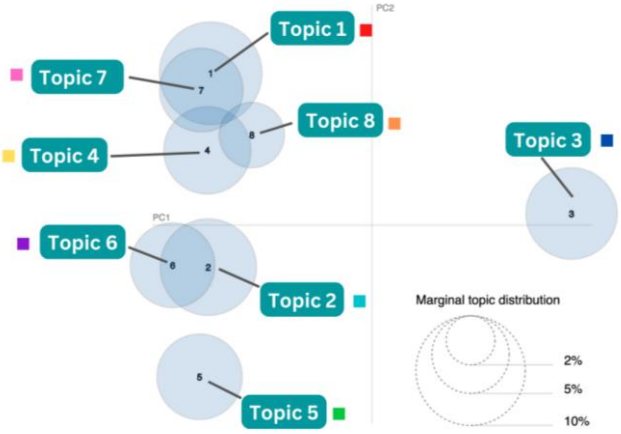


Figure 7: Intertopic distance map generated based on LDAvis [7]

5.2 Results of Cross-Domain with Co-Occurrence Network

Table 2. Result of cross-domain with co-occurrence network. ***Prob***: log probability of term distribution in the topic (normalized), EBC: Edge betweenness centrality.

Co-occurrences Terms		Domain		*Prob*		Frequency	EBC
w_1	w_2	w_1	w_2	w_1	w_2		
(a) Top-5 co-occurrences frequency							
สถานการณ์ (Situation)	แพร่ระบาด (Pandemic)	Topic2	Topic1	0.0516	0.0517	**1163**	0
ประชาชน (Population)	พื้นที่ (Area)	Topic1	Topic2	0.0518	0.0586	1062	0
แพร่ระบาด (Pandemic)	โรคติดเชื้อ (Infectious disease)	Topc1	Topc2	0.0516	0.0518	911	**398**
ฉีดวัคซีน (Vaccinate)	โควิด (COVID)	Topic7	Topic1	**0.0764**	0.0678	744	0
สถานการณ์ (Situation)	น้ำ (Water)	Topic2	Topic6	0.0516	**0.0699**	583	0
(b) Top-5 highest co-occurrences EBC							
พื้นที่ (Area)	สร้าง (Costruct)	Topic2	Topic3	**0.0586**	0.0493	91	**8484**
ส่งเสริม (Promote)	ประชาชน (Population)	Topic3	Topic1	0.0499	**0.0518**	97	6275
พื้นที่ (Area)	จุด (Point)	Topic2	Topic7	0.0585	0.0496	92	4693
ประชาชน (Population)	สถานการณ์ (Situation)	Topic1	Topic2	0.0518	0.0516	**117**	4578
เจ้าหน้าที่ (Officer)	พื้นที่ (Area)	Topic1	Topic2	0.0491	0.0585	94	4442

In table 2 (a), the co-occurrences that have the highest frequency are mostly associated with COVID-19, for example (สถานการณ์ /"Situation", แพร่ระบาด /"Pandemic"), (แพร่ระบาด /"Pandemic", โรคติดเชื้อ /"Infectious disease") and (ฉีดวัคซีน /"Vaccinate", โควิด /"COVID") gain its frequency due to the COVID-19 pandemic situation in 2021 but there's no considerable relevance between individual term probability and frequency. However, as shown in table 2 (b), co-occurrences with higher EBC tend to have lower frequency compared to (a) which has correlatively lower EBC. This relation between EBC and frequency can be represented by exponential relation and illustrated as reveals in Figure 8.

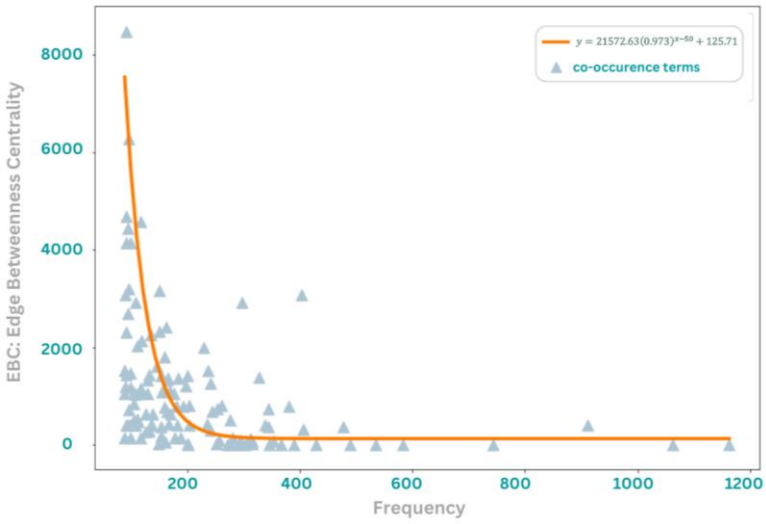


Figure 8: Exponential relation between EBC and co-occurrence frequency.

6 Related Works

Many cross-domain network text analysis research papers are published, such as Takhom et al. [3] proposed the NTA approach with semantic embedding techniques. To reduce misunderstandings in case of cross-disciplinary concept discovery in case studies from question-answering information. Sun et al. [10] proposed a cross-disciplinary approach to understand relationship in the field of Coronavirus disease 2019 (COVID-19). Both NTA research share a contribution of key features on domain indicated by conceptual codebook whereas this paper attempts to implement topic modeling with entire articles to cluster terminologies into their most relevance domain.

Kim & Gil [11] proposed a classification approach with the LDA technique for the extraction of significant words from the abstracts of each article. Sheikha [12] set up a data mining to gather Coronavirus disease information by applying the LDA technique with latent semantic analysis (LSA) to represent a significant correlation between social media data and World Health Organization (WHO) data. Both papers use LDA to cluster data into a particular domain, while this work provides an overview of cross-domain in co-occurrence network. Instead of indicating cross-domain directly, the authors weigh up the probability of terminology distribution in the topic with centrality through network analysis.

7 Conclusion and Future Work

This paper presents a lexical identification technique with LDA to generate a co-occurrence network. To support the text data analysis from Thai electronic news and reduce the cross-domain ambiguity that appear in news articles with various disciplines. To make a model evaluation, the authors have chosen quantitative assessment approach. The results of applying this approach to study case from reliable Thai electronic news in 2021 shows that cross-domain ambiguity occurs in Thai electronic news can be identify and describe its contextual correlation by using LDA and NTA approach along with human interpretation is capable of considerable reducing the cross-domain ambiguity occurred in Thai electronic news.

Furthermore, the authors plan to improve the capability of reducing cross-domain ambiguity and enhance the ability to identify data before using the topic modeling by semantic relationship extraction and information extraction [9]. This is the automatic extraction of knowledge from documents and electronic sources. The expected results of the data extraction process will be in the form of a triple structure [10], and an important part of creating a knowledge graph [11].

Acknowledgment

This paper is partially supported by the National Science and Technology Development Agency (NSTDA) and Siam Commercial Bank (SCB) under Junior Science Talent Project (JSTP-SCB) bachelor's scholarship program. Secondly, the authors are immensely grateful to Language and Semantic Technology Laboratory (LST), National Electronics and Computer Technology Center (NECTEC) who provided insight, necessary resources, and expertise that greatly assisted the research. Finally, the authors would like to show our gratitude to Computer Engineer Department, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand for providing the laboratory and necessary equipment.

References

[1] R. Popping, “Knowledge graphs and network text analysis,” *Soc. Sci. Inf.*, vol. 42, no. 1, pp. 91–106, 2003, doi: 10.1177/0539018403042001798.

[2] D. M. Blei, A. Y. Ng, and M. T. Jordan, “Latent dirichlet allocation,” *Adv. Neural Inf. Process. Syst.*, vol. 3, no. Jan, pp. 993–1022, 2002.

[3] A. Takhom, P. Boonkwan, M. Ikeda, S. Usanavasin, and T. Supnithi, “Reducing miscommunication in cross-disciplinary concept discovery using network text analysis and semantic embedding,” in *The 6th Joint International Semantic Technology Conference, CEUR Workshop Proceedings 1741*, 2017, vol. 2000, pp. 20–31, [Online]. Available: http://ceur-ws.org/Vol-2000/paos2017_paper3.pdf.

[4] A. S. Vaz et al., “The progress of interdisciplinarity in invasion science,” Ambio, vol. 46, no. 4, pp. 428–442, 2017, doi: 10.1007/s13280-017-0897-7.

[5] J. Chuang, C. D. Manning, και J. Heer, ‘Termite: Visualization techniques for assessing textual topic models’, στο Proceedings of the international working conference on advanced visual interfaces, 2012, σσ. 74–77.

[6] Brandes, Ulrik. "On variants of shortest-path betweenness centrality and their generic computation". Social networks vol. 30, no.2, pp. 136-145, 2008.

[7] C. Sievert και K. Shirley, ‘LDAvis: A method for visualizing and interpreting topics’, στο Proceedings of the workshop on interactive language learning, visualization, and interfaces, 2014, σσ. 63–70.

[8] D. Leenoi, et al., “A Construction of Thai WordNet through Translation Equivalence”, in *The 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2022), November 5-7, 2022*

[9] P. Eckert και J. R. Rickford, Style and sociolinguistic variation. Cambridge University Press, 2001.

[10] J. Sun *et al.*, “COVID-19: epidemiology, evolution, and cross-disciplinary perspec-tives,” *Trends Mol. Med.*, vol. 26, no. 5, pp. 483–495, 2020.

[11] S. W. Kim and J. M. Gil, “Research paper classification systems based on TF-IDF and LDA schemes,” *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, pp. 1–21, 2019, doi: 10.1186/s13673-019-0192-7.

[12] H. Sheikha, “Text mining Twitter social media for Covid-19 Comparing latent semantic analysis and latent Dirichlet allocation.” 2020, [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-32567.

[13] T. H. Nguyen, B. Plank, and R. Grishman, “Semantic representations for domain adaptation: A case study on the tree kernel-based method for relation extraction,” in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of*

Natural Language Processing, Proceedings of the Conference, 2015, vol. 1, pp. 635–644, doi: 10.3115/v1/p15-1062.

[14] A. Kumar and S. Dinakaran, "Textbook to triples: Creating knowledge graph in the form of triples from AI TextBook," *arXiv Prepr. arXiv2111.10692*, 2021.

[15] A. Takhom, D. Leenoi, C. Sophaken, P. Boonkwan, and T. Supnithi, "An Approach of Network Analysis Enhancing Knowledge Extraction in Thai Newspapers Contexts," *J. Intell. Informatics Smart Technol.*, vol. 6, no. October 2021, pp. 19–24, 2021, [Online]. Available: https://jiist.aiat.or.th/assets/uploads/1635853027829tBupD1635602106085fdegH39.pdf.