Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.1, February 2025, pp. 1~7
e-ISSN: 2548-6861

Knowledge Discovery Through Topic Modeling on GoPartner User Reviews Using BERTopic, LDA, and NMF

Metti Detricia Pratiwi [1], Ken Ditha Tania [2]*
* Sistem Informasi, Universitas Sriwijaya
mettidetriciaa@gmail.com [1], kenya.tania@gmail.com [2]

Article Info

Article history:

Received 2024-10-29
Revised 2024-11-07
Accepted 2024-11-14

Keyword:

Knowledge Discovery,
Topic Modeling,
BERTopic,
LDA,
NMF.

ABSTRACT

Transportation and food delivery services are one of the driving sectors of the digital economy in Indonesia. The e-Conomy SEA 2023 report shows that the transportation and food delivery services sector experienced a decrease in GMV in 2023 by 8% from the previous year. The decline in GMV indicates a decrease in transaction value in the transportation and food delivery service sector. GoPartner is an application developed by GoTo to assist driver partners in carrying out various services in the gojek application which is one of the applications engaged in the transportation sector and food delivery services. Drivers as people who provide services directly to consumers are certainly one of the factors that influence customer behavior in using services. To find out the problems faced by drivers, this research conducts knowledge discovery through topic modeling on GoPartner application reviews using BERTopic, LDA, and NMF, each of these methods has a different approach. Based on the research results and the quality of the topics generated, BERTopic and LDA have better quality in analyzing GoPartner user reviews.

I. INTRODUCTION

Various science, information, and communication innovations have impeded Indonesia's digital economy's growth rate. According to the e-Conomy SEA 2023 report, five sectors drive the digital economy: e-commerce, online travel, transportation and food delivery, online media, and digital financial services. The expansion of these sectors is closely related to the widespread use of smartphones and internet access among the public, which has led to an increased demand for online application-based services and the creation of job opportunities [1].

One of the entities engaged in the transportation and food delivery service sector in Indonesia is PT GoTo Gojek Tokopedia Tbk (GoTo). Goto provides application-based services, including e-commerce, transportation, and food delivery. GoPartner is one of the applications managed by GoTo to facilitate Gojek Drivers in receiving orders, monitoring earnings, and accessing information. GoPartner is integrated with the services available on Gojek. Through Gojek, users can transact with various services, such as GoRide, GoCar, GoFood, GoMart, and others. The alignment of these two applications in connecting Gojek drivers and customers affects user satisfaction with the application and the services provided. The quality of service provided by Gojek drivers is one of the factors that influences customer satisfaction with Gojek.

In addition, the satisfaction of Gojek drivers is also equally important. Complaints from Gojek drivers about GoPartner, which GoTo provides as a facility for drivers to perform their duties, represent one aspect that requires examination. This is supported by data from the e-Conomy SEA 2023 report, which shows that among various sectors, the transportation and food delivery sectors experienced a decline in Gross Merchandise Value (GMV) of 8% in 2023 compared to the previous year. This fact indicates a change in consumer purchasing behavior. These changes certainly involve all factors related to the service ecosystem, and one of the driving factors is the Gojek driver. The obstacles or conveniences that gojek drivers feel when using the GoPartner application will certainly affect the user experience in transactions using the Gojek application. Gojek drivers' reviews of applications that facilitate them to carry out their work as people who directly provide services to customers can represent real operational

conditions and factors that affect the decline in service performance provided to customers. Gojek drivers' opinions regarding their satisfaction with GoPartner can be seen through reviews on the Google Play Store. These reviews provide information that describes actual user experiences with the application [2].

Keywords frequently appear in a text from each review and become words that represent a topic [3]. Through topic modeling, unsupervised machine learning can automatically detect patterns, phrases, and cluster sets of documents represented by sets of words and expressions that appear together [4]. This machine learning technique is widely used in Natural Language Processing (NLP) applications to analyze unstructured textual data and automatically discover abstract topics [5]. Therefore, applying topic modeling is necessary to gain insight through knowledge discovery of emerging patterns.

Knowledge discovery has become an important research to support decision-making [6], [7], [8]. Topic modeling is one of the most frequently used computer-assisted text data mining applications for knowledge discovery [8]. Knowledge discovery approaches contribute to social computing by utilizing information processing technologies in data mining processes to identify hidden patterns in individual behavior [9]. Discovering new information and insight gained through hidden patterns in GoPartner user reviews can help decision-making improve strategies.

Various methods can do topic modeling with different approaches such as Latent Dirichlet Allocation (LDA) which is a generative statistical model, Non-Negative Matrix Factorization (NMF) which performs topic extraction with a linear algebra approach, BERTopic, and Top2Vec with a embedding approach [10]. In LDA, a probabilistic model, latent topics are discovered in the document collection by analyzing word distributions [11]. NMF is a statistical model that factorizes or decomposes a non-negative input matrix into two non-negative sub-matrices [12]. Meanwhile, BERTopic utilizes BERT embeddings and clustering algorithms to generate semantically meaningful topic representations [13]. In some previous studies, LDA is the most often used method for topic modeling, such as in research conducted by Mutmainah et al. (2023) applied BiLSTM and LDA for sentiment analysis and topic modeling in telemedicine applications on the Google Play Store which resulted in topics with coherence score of 0.6437 in positive sentiment, 0.6132 in negative sentiment, and 0.6296 in neutral sentiment [14]. The next application of LDA was carried out by Fahlevvi and Azhari (2022) adding n-grams after the preprocessing stage, resulting in 5 topics and 20 passes with a coherence value of 0.53 [15]. Another study conducted by Ogunleye et al. (2023) compared topic modeling approaches in the banking context, that study aims to introduce the use of Kernel Principal Component Analysis (KernelPCA) and K-means on the BERTopic architecture with the results of the application of KernelPCA and K-means managed to become a model with the highest coherence score 0.864 [16].

Based on the problems and previous research, this study applied topic modeling using LDA, NMF, and BERTopic to GoPartner user reviews obtained through the Google Play Store. This research was conducted to find opinions in the form of complaints or user satisfaction while using the GoPartner application through topics generated by each method with a different approach. The coherence value and quality of topics generated by the three topic modeling methods used will measure how well the model used in generating topics that have easy-to-understand meaning to gain deep insight into GoPartner application user reviews.

II. METHOD

A. Data Collection

The first stage in this research is data collection, in the form of user reviews of GoPartner application on the Google Play Store [14]. This stage is the first step in the research process which aims to obtain data and information that is relevant and under the research objectives. Reviews are collected using Google-Play-Scraper library through Google Colaboratory by entering the application ID and review time range for February 29 to August 30, 2024. At this stage, 15000 reviews were generated and stored in Excel.

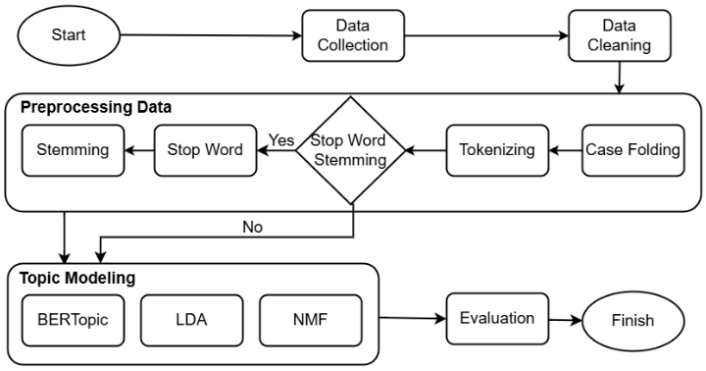

Figure 1. Research flow [17]

B. Data Cleaning

In the data cleaning stage, cleaning removes mentions, numbers, URLs, and non-alphanumeric characters, selects the review column to be analyzed, and corrects writing and word usage errors. This stage is carried out because the data obtained previously contains a lot of unwanted data [18]. At this stage, only user reviews in the form of complaints and user satisfaction related to the GoPartner application will be selected for topic modeling. This stage aims to obtain overview data that provides relevant information for topic modeling. After cleaning and selecting data, the collected data which initially amounted to 15000 was reduced to 11417.

C. Preprocessing Data

In the preprocessing stage, the original sentences will be prepared to be cleaner and ready to be used in the model and analysis. This is a very important stage to get more accurate

analysis results by ensuring data quality and consistency. There are four stages in data preprocessing, namely case folding; a process of converting all letters in the sentence into lowercase, tokenizing; converting sentences into word units, stopwords; removing words that do not add value in the analysis, and stemming; converting words to their basic form by removing affixes.

TABLE I
PREPROCESSING DATA

Process	Before	After
Case Folding	Masih banyak kekurangan di aplikasinya Sering terjadi gangguan dan sering mengganggu driver saat menjalankan tugas	masih banyak kekurangan di aplikasinya sering terjadi gangguan dan sering mengganggu driver saat menjalankan tugas
Tokenizing	masih banyak kekurangan di aplikasinya sering terjadi gangguan dan sering mengganggu driver saat menjalankan tugas	['masih', 'banyak', 'kekurangan', 'di', 'aplikasinya', 'sering', 'terjadi', 'gangguan', 'dan', 'sering', 'mengganggu', 'driver', 'saat', 'menjalankan', 'tugas']
Stop Word	['masih', 'banyak', 'kekurangan', 'di', 'aplikasinya', 'sering', 'terjadi', 'gangguan', 'dan', 'sering', 'mengganggu', 'driver', 'saat', 'menjalankan', 'tugas']	['banyak', 'kekurangan', 'aplikasinya', 'sering', 'terjadi', 'gangguan', 'sering', 'mengganggu', 'driver', 'menjalankan', 'tugas']
Stemming	['banyak', 'kekurangan', 'aplikasinya', 'sering', 'terjadi', 'gangguan', 'sering', 'mengganggu', 'driver', 'menjalankan', 'tugas']	['banyak', 'kurang', 'aplikasi', 'sering', 'jadi', 'ganggu', 'sering', 'ganggu', 'driver', 'jalan', 'tugas']

D. Topic Modeling

After preprocessing data, the next stage is topic modeling. Three techniques will be used BERTopic, LDA, and NMF. This stage is conducted using tokenizing data and stemming data.

BERTopic is a topic modeling technique that utilizes Bidirectional Encoder Representations from Transformers (BERT) and Term Frequency-Inverse Document Frequency (TF-IDF) to generate easily interpretable topics through clusters and important words that describe the topic [19]. BERTopic consists of several steps: document embedding, dimensionality reduction, clustering, and topic representation. The implementation of BERTopic is performed using the BERTopic library. In this research, the algorithms used in dimensionality reduction and clustering are Kernel Principal Component Analysis (KPCA) and K-means. The selection of these algorithms is based on previous research by Ogunleye et al. (2023) where the use of these algorithms produces the highest coherence score of other algorithms.

```
kpca = KernelPCA(n_components = 5, kernel = 'rbf',
                 gamma=15, random_state=42)

cluster_model = KMeans(n_clusters=8, n_init=10,
                       max_iter=300, init='k-means++')

ctfidf_model = ClassTfidfTransformer()
topic_model = BERTopic(umap_model=kpca,
                       hdbscan_model=cluster_model,
                       ctfidf_model=ctfidf_model,
                       language="multilingual")
```

Figure 2. BERTopic parameter setup

Latent Dirichlet Allocation (LDA), one of the most popular topic modeling techniques, uses a generative probabilistic model that finds latent topics in a collection of documents by studying the relationship between words, documents, and topics [20]. LDA topic modeling not only analyzes sentences that contain keywords or phrases related to Nature of Science (NOS) but also includes sentences that convey participants' views on NOS even though they do not contain specific keywords or phrases related to NOS [21]. The LDA model is constructed using gensim library. In this research, to get the highest coherence score from LDA, a grid search is used to determine the number of topics and other parameters that produce the highest coherence score.

```
# Grid search parameters
num_topics_list = [2, 4, 6, 8, 10]
passes_list = [10, 20, 30, 50]
alpha_list = ['symmetric', 'asymmetric', 'auto']
```

Figure 3. Grid search parameter for LDA

Non-negative Matrix Factorization (NMF) is a decompositional, non-probabilistic algorithm that uses matrix factorization [10]. NMF consists of two matrices W and H, where W is the topics found and H is the coefficient (weight) for those topics [22]. Each W column represents the strength of association of each word in the dictionary with the topic being learned and each H column represents the relevance of each topic for a given document in the corpus [23]. Just like LDA, NMF also uses grid search to determine the combination of the number of topics and parameters that produces the highest coherence score. In this research, scikit-learn was used to implement NMF.

```
# Grid search parameters
num_topics_list = [2, 4, 6, 8, 10]
max_iter_list = [100, 200, 300]
init_list = ['random', 'nndsvd']
solver_list = ['cd', 'mu']
```

Figure 4. Grid search parameter for NMF

E. Evaluation

This research uses a coherence score to measure the semantic quality of the topics generated by the model. The coherence model used is C_v. C_v measure evaluates the similiarity and consistency of each word contained in the topic [24]. C_v evaluates based on a sliding window, a set of top word segmentation, indirect confirmation using Normalized Point-wise Mutual Information (NPMI), and cosine similarity [22]. The coherence score was calculated using the gensim library.

```
analyzer = vectorizer.build_analyzer()
tokens = [analyzer(doc) for doc in cleaned_docs]

dictionary = Dictionary(tokens)
corpus = [dictionary.doc2bow(token) for token in tokens]

topic_words = []
for topic in range(len(set(topics))):
    topic_terms = topic_model.get_topic(topic)
    if topic_terms:
        topic_words.append([word for word, _ in topic_terms])

# Evaluasi koherensi
coherence_model = CoherenceModel(topics=topic_words,
                                 texts=tokens,
                                 corpus=corpus,
                                 dictionary=dictionary,
                                 coherence='c_v')
coherence = coherence_model.get_coherence()
```

Figure 5. Coherence mode

III. Result and Discussion

After performing topic modeling using BERTopic on tokenizing and stemming data, the coherence value is obtained in Table 2.

TABLE II
COHERENCE SCORE OF BERTOPIC

Model	Data	Coherence Score
BERTopic	Tokenizing	0.6904775079537018
	Stemming	0.6045666979982685

In the application of grid search in LDA and NMF, the parameter combination with the best coherence value is obtained as shown in Table 3.

TABLE III
RESULT OF GRID SEARCH

Model	Data	Result
LDA	Tokenizing	Best Num Topics: 4 Best Passes: 20 Best Alpha: auto Best Coherence Score: 0.6887674992369215
	Stemming	Best Num Topics: 6 Best Passes: 20 Best Alpha: auto Best Coherence Score: 0.6262973633020125
NMF	Tokenizing	Best Num Topics: 2 Best Init: random Best Solver: cd Best Max Iter: 100 Best Coherence Score: 0.6449577821754222
	Stemming	Best Num Topics: 2 Best Init: random Best Solver: mu Best Max Iter: 100 Best Coherence Score: 0.5566103489860162

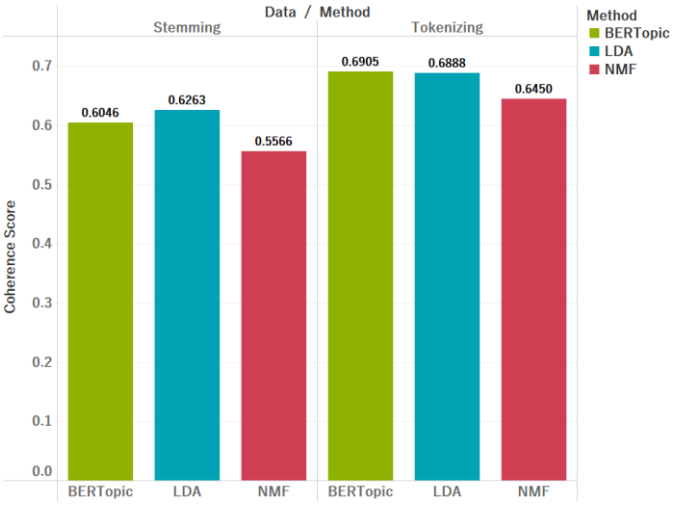

Figure 6. Coherence score comparison

After modeling the topics, Figure 6 compares the coherence scores produced by each model. The figure shows that LDA achieves the highest coherence score in topic modeling using stemming data. BERTopic gets a greater score than LDA in topic modeling using tokenizing data. NMF has the lowest coherence score in topic modeling with both types of data.

TABLE IV
LDA TOPICS WITH TOKENIZING DATA

LDA - Tokenizing	Terms
Topic 1	tidak, orderan, makin, yang, malah, ada, dapat, order, di, update
Topic 2	yang, driver, dan, gojek, tidak, aplikasi, di, untuk, dengan, mitra
Topic 3	driver, tidak, jauh, kilometer, yang, jarak, jemput, argo, ribu, di
Topic 4	saya, tidak, sudah, bisa, ada, yang, gojek, di, driver, akun

LDA with tokenizing data has a coherence value of 0.6888 and produces the terms presented in Table 4. The resulting terms contain several words that do not represent the topic discussed, such as 'yang', 'dan', and 'di'. This happens because the data used did not go through the stopword stage to eliminate words that are considered unimportant. Topic 1 identified a problem with reduced orders after users updated the application. Topic 2 revealed problems experienced by driver-partners when using the app. Topic 3 has the easiest

terms to understand, which discusses the long pick-up distance. Topic 4 identified discussions about gojek driver accounts that have problems such as inaccessibility.

TABLE V
LDA TOPICS WITH STEMMING DATA

LDA - Stemming	Terms
Topic 1	jauh, jemput, kilometer, driver, jarak, customer, jalan, antar, ambil, maps
Topic 2	mitra, driver, gojek, kalian, kerja, anak, sangat, sejahtera, potong, besar
Topic 3	moga, jauh, bagus, maps, sameday, baik, jalan, gacor, banyak, gila
Topic 4	akun, daftar, jadi, gojek, aplikasi, bantu, tiba, padahal, selalu, baik
Topic 5	order, jam, masuk, hari, order, versi, kasih, order, baru, pagi
Topic 6	driver, order, makin, buat, gojek, aplikasi, sistem, sama, banyak, malah

The terms generated from LDA with stemming data presented in Table 5 can effectively represent the discussed topics and encompass many topics. In topic 1, the words that appear are related to pick-up and drop-off distances of customers, topic 2 is related to the welfare of drivers with large deductions, topic 3 discusses sameday service, maps, and orders, topic 4 is related to gojek account registration, topic 5 about the application version related to order entry hours, and topic 6 is about gojek applications and systems.

The knowledge discovery drawn from the topics generated by LDA using data tokenizing and stemming covers several key issues involving operational aspects such as orders, driver account issues, features, services, and driver welfare.

TABLE VI
NMF TOPICS WITH TOKENIZING DATA

NMF - Tokenizing	Terms
Topic 1	yang, tidak, driver, orderan, saya, di ada, sudah, gojek, dan
Topic 2	makin, makin, sini, ke, aplikasi, parah, malah, jelas, parah, update

TABLE VII
NMF TOPICS WITH STEMMING DATA

NMF - Stemming	Terms
Topic 1	makin, makin, sini, order, parah, aplikasi, parah, malah, susah, jelas
Topic 2	order, driver, gojek, aplikasi, kasih, buat, jadi, sama, jauh, jemput

NMF with tokenizing and stemming data gets the lowest coherence score of the other two methods. Terms generated by the NMF method can be seen in Table 6 and Table 7. The quality of topics generated by NMF is not as high as that produced by LDA. For example, in topic 2 in Table 6 and topic 1 in Table 7, the words ‘makin’ and ‘parah’ appear twice in the same topic. The resulting terms make it somewhat challenging to ascertain the discussed topics, as they include several words that lack significant meaning.

TABLE VIII
BERTOPIC WITH TOKENIZING DATA

BERTopic - Tokenizing	Terms
Topic 1	tidak, yang, driver, orderan, gojek, aplikasi, di, saya, sudah, ada
Topic 2	membantu, sangat
Topic 3	jelas, aplikasi, tidak
Topic 4	bangus, sangat
Topic 5	bobrok, sistem
Topic 6	baik, sangat
Topic 7	susah, dapat, orderan
Topic 8	terbantu, sangat

The coherence score in Figure 6 shows that BERTopic with tokenizing data has the highest coherence score with 0.6905. In Table 8, the terms generated can be represented that topic 1 is related to orders in the gojek application, and topic 2 to topic 8 terms only consist of two or three words directly representing the topic discussed. As in topics 3, 5, and 7 the terms that appear indicate complaints about the clarity of the system and the difficulty of getting orders. While on topics 2, 4, 6, and 8 terms that appear indicate the usefulness of the GoPartner application is very helpful and good.

TABLE IX
BERTOPIC WITH STEMMING DATA

BERTopic - Stemming	Terms
Topic 1	order, driver, aplikasi, gojek, makin, buat, sistem, jadi, kasih, akun
Topic 2	sangat, ekonomi, bobrok, bantu, jelas, bagus, baik, sistem, makin, aplikasi
Topic 3	bantu, sangat
Topic 4	susah, order
Topic 5	jelas, aplikasi
Topic 6	prioritas, kembali, sistem
Topic 7	bagus, sangat
Topic 8	sangat, guna, muas

Terms generated in BERTopic with stemming data are almost the same as tokenizing data. The difference lies in topic 2 which represents the relationship of the application system with the economy, topic 6 is related to the request to restore the priority system. Other topics are similar to tokenizing data, except that words such as ‘sangat’, ‘tidak’, and ‘dapat’ are removed during the stopword removal stage, while certain words like ‘bantu’ and ‘muas’ result from the stemming stage.

The knowledge discovery that can be extracted from the topics generated by BERTopic using data tokenizing and stemming includes several main issues such as the difficulty of getting orders that affect the economy or income of drivers. Aside from these complaints, the GoPartner app remains a helpful app for some drivers.

IV. CONCLUSION

From the application of topic modeling on GoPartner application user reviews using three methods that have different approaches, namely LDA, NMF, and BERTopic, it was found that the application of BERTopic using the same setup parameters as in the research conducted by Ogunleye et al. (2023) on tokenizing data produced the highest coherence score with eight number of topics. LDA gets the highest coherence score in topic modeling using stemming data compared to other methods. This indicates that BERTopic excels in modeling topics using original sentences and LDA demonstrates superiority in generating topics from stemming data. While NMF gets the lowest coherence score and the given topic is difficult to understand because there is no strong similarity between words, the research conducted by Egger & Yu (2022) also explained that this is one of the disadvantages of NMF.

In the topics generated by LDA, NMF, and BERTopic there are several complaints expressed by GoPartner application users such as the distance of pick-up, account, registration, difficulty getting orders, and priority system, In addition to complaints, some GoPartner applications users also feel helped by this application.

The results of knowledge discovery that can be taken from the research results are application users who are gojek drivers feel operational constraints when accessing and using the application. The main obstacle faced by drivers is the distance of picking up customers who are too far from the driver's position, making some drivers feel that the fare listed in the gojek application is not appropriate because the pick-up of customers whose distance is sometimes farther than the delivery distance to the destination location, besides that the account registration process also needs to be considered for the convenience of drivers in accessing the application and running orders. Apart from these complaints, the drivers also appreciate the benefits received from this application. This shows that although there are still some obstacles, this application still makes a positive contribution. However, the benefits provided will be even greater if these constraints can be followed up for the convenience of drivers and customers.

Based on the results of this research, there are several suggestions for further research, such as applying other topics modeling methods by combining optimization techniques and using more data to improve the quality of topics and coherence score.

REFERENCES

[1] S. Iqbal and Z. A. Bhatti, "A qualitative exploration of teachers' perspective on smartphones usage in higher education in developing countries," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, Dec. 2020, doi: 10.1186/s41239-020-00203-4.

[2] Y. Liu, L. Liu, H. Liu, and S. Gao, "Combining goal model with reviews for supporting the evolution of apps," *IET Softw.*, vol. 14, no. 1, pp. 39–49, Feb. 2020, doi: 10.1049/iet-sen.2018.5192.

[3] M. R. Maarif, "Summarizing Online Customer Review using Topic Modeling and Sentiment Analysis," 2022. doi: 10.14421/jiska.2022.7.3.177-191.

[4] D. Atzeni, D. Bacciu, D. Mazzei, and G. Prencipe, "A Systematic Review of Wi-Fi and Machine Learning Integration with Topic Modeling Techniques," *Sensors*, vol. 22, no. 13. MDPI, Jul. 01, 2022. doi: 10.3390/s22134925.

[5] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/s41870-023-01268-w.

[6] F. Alqurashi and I. Ahmad, "A data-driven multi-perspective approach to cybersecurity knowledge discovery through topic modelling," *Alexandria Eng. J.*, vol. 107, pp. 374–389, Nov. 2024, doi: 10.1016/j.aej.2024.07.044.

[7] W. Ning, J. Liu, and H. Xiong, "Knowledge discovery using an enhanced latent Dirichlet allocation-based clustering method for solving on-site assembly problems," *Robot. Comput. Integr. Manuf.*, vol. 73, Feb. 2022, doi: 10.1016/j.rcim.2021.102246.

[8] L. Mora, X. Wu, and A. Panori, "Mind the gap: Developments in autonomous driving research and the sustainability challenge," *Journal of Cleaner Production*, vol. 275. Elsevier Ltd, Dec. 01, 2020. doi: 10.1016/j.jclepro.2020.124087.

[9] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc. Sci. Res.*, vol. 110, Feb. 2023, doi: 10.1016/j.ssresearch.2022.102817.

[10] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Front. Sociol.*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.

[11] S. Ying, "Guests' Aesthetic experience with lifestyle hotels: An application of LDA topic modelling analysis," *Heliyon*, vol. 10, no. 16, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35894.

[12] C. Meaney *et al.*, "Non-negative matrix factorization temporal topic models and clinical text data identify COVID-19 pandemic effects on primary healthcare and community health in Toronto, Canada," *J. Biomed. Inform.*, vol. 128, Apr. 2022, doi: 10.1016/j.jbi.2022.104034.

[13] A. Kumar, A. Karamchandani, and S. Singh, "Topic Modeling of Neuropsychiatric Diseases Related to Gut Microbiota and Gut Brain Axis Using Artificial Intelligence Based BERTopic Model on PubMed Abstracts," *Neurosci. Informatics*, p. 100175, Dec. 2024, doi: 10.1016/j.neuri.2024.100175.

[14] S. Mutmainah, D. H. Fudholi, and S. Hidayat, "Analisis Sentimen dan Pemodelan Topik Aplikasi Telemedicine Pada Google Play Menggunakan BiLSTM dan LDA," *J. MEDIA Inform. BUDIDARMA*, vol. 7, no. 1, p. 312, Jan. 2023, doi: 10.30865/mib.v7i1.5486.

[15] M. R. Fahlevvi and Azhari, "Topic Modeling on Online News.Portal Using Latent Dirichlet Allocation (LDA)," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 16, no. 4, p. 335, Oct. 2022, doi: 10.22146/ijccs.74383.

[16] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of Topic Modelling Approaches in the Banking Context," *Appl. Sci.*, vol. 13, no. 2, Jan. 2023, doi: 10.3390/app13020797.

[17] T. Ramamoorthy, V. Kulothungan, and B. Mappillairaju, "Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India," *Front. Artif. Intell.*, vol. 7, 2024, doi: 10.3389/frai.2024.1329185.

[18] S. E. Uthirapathy and D. Sandanam, "Topic Modelling and Opinion Analysis on Climate Change Twitter Data Using LDA and BERT Model.," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 908–917. doi: 10.1016/j.procs.2023.01.071.

[19] L. B. Hutama and D. Suhartono, "Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic," *Inform.*, vol. 46, no. 8, pp. 81–90, 2022, doi: 10.31449/inf.v46i8.4336.

[20] S. J. Blair, Y. Bi, and M. D. Mulvenna, "Aggregated topic models for increasing social media topic coherence," *Appl. Intell.*, vol. 50, no. 1, pp. 138–156, Jan. 2020, doi: 10.1007/s10489-019-01438-z.

[21] M. Wang, S. Gao, W. Gui, J. Ye, and S. Mi, "Investigation of Pre-service Teachers' Conceptions of the Nature of Science Based on the LDA Model," *Sci. Educ.*, vol. 32, no. 3, pp. 589–615, Jun. 2023,

doi: 10.1007/s11191-022-00332-4.

[22] Zoya, S. Latif, F. Shafait, and R. Latif, "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling," *IEEE Access*, vol. 9, pp. 127531–127547, 2021, doi: 10.1109/ACCESS.2021.3112620.

[23] P. Li *et al.*, "Guided Semi-Supervised Non-Negative Matrix Factorization," *Algorithms*, vol. 15, no. 5, May 2022, doi: 10.3390/a15050136.

[24] T. Gokcimen and B. Das, "Exploring climate change discourse on social media and blogs using a topic modeling analysis," *Heliyon*, vol. 10, no. 11, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32464.