SJCS Scientific Journal of Computer Science
futuristech PT. Teknologi Futuristik Indonesia
Date of publication May 10, 2024, date of current version May 10, 2024.
Digital Object Identifier **10.64539/sjcs.v1i1.2025.19**

Article

Identification of Dominant Topics in Public Discussions on IKN Using Latent Dirichlet Allocation (LDA) and BERTopic

Ariska Fitriyana Ningrum[1,*], Florence Jean B. Talirongan[2], Diana May Glaiza G. Tangaro[3]

[1] Department of Data Science, Universitas Muhammadiyah Semarang, Semarang, Indonesia; ariskafitriyana@unimus.ac.id
[2] College of Computer Studies, Misamis University, Ozamiz City, Misamis Occidental, Philippines; florencejean.talirongan@mu.edu.ph
[3] TK Elevator WLL, Fereej Bin Mahmoud, Shoumoukh Tower, Doha, Qatar; dianahmayglaiza@gmail.com
* Correspondence

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Abstract: This study aims to analyze public opinion related to the relocation of Indonesia's National Capital City (IKN) through topic modeling on Twitter data. The two main approaches used are Latent Dirichlet Allocation (LDA) based on Bag of Words and BERTopic based on Transformer language model. LDA was chosen for its ability to identify topic distribution in large text collections, while BERTopic was used to overcome the limitations of LDA in capturing semantic meaning in short and informal texts such as tweets. The analysis was conducted on a collection of tweets discussing the relocation of IKN, with the aim of uncovering the main themes and public perceptions. The result of LDA showed three main topics in the public discussion, namely (1) political debate and nationalism related to the relocation, (2) policy implementation and project execution, and (3) economic justification and challenges facing Jakarta. Mean-while, BERTopic identified topics with more contextual representations, including aspects of investment, economic impact construction progress, and public perception. Dominant topics include urban relocation, investment in IKN, and socio-economic impacts. The novelty of study lies in the comparison of two topic modeling approaches in the context of social media sentiment analysis related to major public policy issues. These findings not only enrich the understanding of the narratives that develop in society, but also provide important insights for policy makers in responding to public opinion more appropriately and contextually.

Keywords: Topic Modeling; LDA; BERTopic; Capital Relocation; Sentiment Analysis

1. Introduction

Latent Dirichlet Allocation (LDA) is one of the popular methods for topic modeling used in many previous studies to extract information from large text collections [1]. LDA was first introduced by Blei et al. (2003) and is a probabilistic-based generative approach that assumes that each document consists of various topics, and each topic has a certain word distribution [2]. Previous research shows that LDA is effective in finding latent structures in text, especially in the analysis of scientific literature, social media, and customer reviews [3]. Although LDA can identify topics automatically, this method has a weakness in determining the optimal number of topics and sometimes produces topics that are less interpretable [4].

As a more modern alternative, BERTopic was introduced by Grootendorst (2020) as a Transformer-based topic modeling method that utilizes embedding from language models such as BERT [5]. BERTopic incorporates clustering techniques, such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), to group text based on semantic representations [6]. Previous research shows that BERTopic excels in understanding the context of words in more complex texts compared to LDA [7]. In addition, this method is more flexible in handling natural language as it utilizes a richer and more contextual vector representation [8]. Compared to LDA, research shows that BERTopic is superior in capturing semantic relationships between words in a document [5]. Some studies confirm that BERTopic can provide a better interpretation of topics because the Transformer model allows for a deeper understanding of word meanings in various contexts [9]. In addition, BERTopic has the advantage of handling short texts, such as tweets or social media comments, which are often challenging for LDA due to the limitations of word distribution-based approaches [5]). However, BERTopic requires more computational resources than LDA, which makes it less efficient for the analysis of very large datasets [10].

In recent topic modeling research, the combination of LDA and BERTopic approaches has also been explored to improve the accuracy and interpretation of results [11]. Some studies use LDA as a baseline to compare BERTopic's performance in various domains, such as news analysis, public policy, and customer surveys. With the development of Transformer-based language models, methods such as BERTopic are increasingly in demand in text analysis due to their ability to capture context and relationships between words better than classic approaches such as LDA [12]. The analysis of text from Twitter regarding positive and negative sentiments towards the relocation of Indonesia's national capital is becoming an increasingly important topic in text mining research and public opinion analysis [13]. Twitter as a social media platform allows users to express their opinions in real-time, so the data obtained is very diverse and reflects various points of view of the community [14]. In the context of relocating IKN, positive sentiments may reflect support for the policy, such as the hope for equitable development and reducing the burden on Jakarta, while negative sentiments may reflect concerns about development costs, environmental impacts, or population displacement [15].

Therefore, an appropriate method is needed to extract, analyze, and categorize the main topics of public conversations regarding this policy. LDA is one of the suitable methods for text mining analysis on Twitter data related to IKN because of its ability to identify topic distribution in large text collections [3]. With LDA, tweets containing positive and negative sentiments can be grouped based on words that often appear in one topic, making it easier to understand the narratives that develop in the [1]. In addition, LDA can help in quantifying the dominance of certain topics in public discourse, such as whether people discuss more economic, environmental, or social aspects related to the relocation of IKN [4].

However, LDA has limitations in capturing deeper semantic relationships between words, especially in short texts such as tweets. As a solution to the limitations of LDA, BERTopic can be used to improve the quality of topic analysis related to public sentiment towards the transfer of IKN [5]. Transformer-based BERTopic is able to capture the contextual meaning of tweets, so it can categorize opinions with a deeper understanding of the semantic relationships between words [7]. For example, tweets with words like "sustainable development" or "foreign investment" can be grouped under the same topic despite using different terminology. BERTopic is also superior in handling the dynamics of informal language and slang that is often used on Twitter, so it can provide more accurate results in the analysis of public opinion regarding the relocation of IKN [16].

In research related to sentiment analysis on social media, the selection of topic modeling methods is crucial to understand public opinion more deeply [17]. LDA has long been used to identify topic patterns in large texts, but its probabilistic distribution-based approach is often less effective in capturing semantic meaning in short texts such as tweet. In contrast, Transformer-based BERTopic can overcome these limitations by generating a more contextualized representation of the text and better understanding the relationship between words within a topic [18]. Therefore, a comparison between LDA and BERTopic is important to evaluate the effectiveness of each method in analyzing public opinion on the IKN move, given the dynamic and informal nature of Twitter data.

2. Research Methodology

2.1. Data Collection

Data was collected from kaggle.com media sources that discuss the Capital City of the Archipelago (IKN). Scraping technique is used to collect comments on the tweeter platform regarding the Capital City of the Archipelago. The data used amounted to 1464 comments containing positive and negative comments. The following table 1 is the form of data used in this study.

2.2 Data Preprocessing

a. Text Cleaning
Text cleaning aims to remove irrelevant elements in the text, such as HTML tags, mentions (@username), hashtags (#hashtag), and URLs. Text cleaning can improve accuracy in text classification, especially in statistical-based models such as Naïve Bayes and SVM.

b. Tokenization
Tokenization is the process of separating text into word units or tokens. Research by Jurafsky & Martin (2021) in Speech and Language Processing explains that tokenization is very important in Natural Language Processing (NLP), especially for deep learning-based models.

c. Stemming & Lemmatization
The stemming and lemmatization process aims to convert words into their basic form for easier analysis in natural language processing (NLP). Stemming works by removing word endings without considering the context, so the result is sometimes not a standard word. In contrast, lemmatization converts words to their base form by considering

grammar and meaning, thus more accurately preserving the meaning of the word. Research by Porter (1980) created the Porter Stemmer, a stemming algorithm that is still widely used today in NLP. Meanwhile, research by Minhas et al. (2019) found that lemmatization is superior to stemming, especially in NLP tasks that require context understanding, because lemmatization can better preserve word meaning.

d. Stopword Removal
Stopword removal is the process of removing common words that have no significant meaning in text analysis, thus helping to improve efficiency in natural language processing (NLP). Research by Manning et al. (2008) in Introduction to Information Retrieval shows that removing stopwords can speed up text processing without reducing the quality of analysis. Stopwords are generally derived from standardized lists, such as the NLTK stopwords list for English or ID-Stopwords specifically developed for Indonesian.

e. Vectorization (TF-IDF & Word Embedding)
Vectorization converts text into numerical form so that it can be processed by machine learning models. One commonly used technique is TF-IDF (Term Frequency-Inverse Document Frequency), which serves to assess the importance of a word in a document. Research by Ramos (2003) shows that TF-IDF is effective in information filtering and text classification. In addition, there are Word Embedding methods, such as Word2Vec, GloVe, and BERT, which are able to capture semantic relationships between words. Mikolov et al. (2013) introduced Word2Vec, which enables the representation of words in numerical vectors based on their context. Meanwhile, BERT (Devlin et al., 2019) produces a more sophisticated word representation by considering context from both directions of the text (bidirectional), making it more effective in understanding the meaning of words in a sentence.

2.3. Topic Modeling with LDA Method

Latent Dirichlet Allocation (LDA) is one of the most popular topic models for extracting information from large amounts of text documents. LDA was first introduced by Blei et al. (2003) as a probabilistic method that bases each document as a distribution of multiple topics, while each topic is considered as a distribution of words. This approach allows the model to uncover latent structures in document collections without the need for manual annotation. Since its introduction, LDA has been widely used in various fields, such as social media analysis, recommendation systems, and information retrieval [1].

One of the main advantages of LDA is its ability to handle large amounts of unstructured text. Research conducted by Griffiths & Steyvers (2004) showed that LDA can be effectively used to find hidden patterns in collections of historical documents. In addition, Li et al. (2016) also showed that this model can be used in opinion analysis to understand users' perceptions of a product or service. In practical applications, selecting the optimal number of topics is a challenge, so several approaches such as perplexity and coherence score are often used to determine the best parameters [19]. In addition, several studies have developed variations of the LDA model to improve accuracy and efficiency. For example, the Dynamic Topic Model (DTM) developed by Blei & Lafferty (2006) enables the analysis of topic changes over time, which is particularly useful in the study of social media and news trends. Meanwhile, Supervised LDA (sLDA) introduced by Blei & McAuliffe (2007) integrates additional variables to improve prediction in text classification tasks. Other variations such as Hierarchical LDA (hLDA) have also been introduced to capture hierarchical relationships between topics [20].

Despite its advantages in topic modeling, LDA also has some limitations. One of them is the dependence on the number of topics to be determined before modeling is performed, which may affect the interpretation results (Chang et al., 2009). In addition, this model assumes independence between words within a topic, which does not always reflect the actual relationship in natural language. Some studies have proposed neural network-based approaches, such as the Neural Topic Model (NTM), which can overcome some of the limitations of LDA by utilizing a richer representation of words [21]. Overall, LDA remains one of the main approaches in text topic analysis due to its ability to probabilistically capture the latent structure of documents. Various model developments and optimization techniques continue to be carried out to improve the performance of LDA in various applications. In the future, integration with deep learning-based methods and exploration in the field of big data are predicted to further increase the effectiveness of this model in various text analysis scenarios [22].

2.4. Topic Modeling with BERTopic Method

BERTopic is a topic modeling model that combines text representation using Transformer with clustering techniques to extract topics from text data. This model utilizes the text representation generated by the BERT model to capture the contextual meaning of the words in the document. Furthermore, dimensionality reduction techniques such as Kernel Principal Component Analysis (KernelPCA) are used to reduce data complexity, and clustering algorithms such as K-means are applied to identify topics hidden in the text corpus [23].

Table 1. Research Variables.

Tweet	Sentiment
@jokowi saya sangat setuju pak bahkan lebih setuju ketika bapak mengalokasikan dulu angggaran untuk pemindahan ibu kota sehingga anggaran nya di fokus kan untuk penanganan covid 19 ini. lebih baik selamatkan dulu nyawa masyarakat di banding fokus terhadap	positive
@hnurwahid @FPKSDPRRI Saya setuju ibu kota pindah tapi ga sekarang	positive
@MardaniAliSera @FPKSDPRRI Saya dan mayoritas rakyat NKRI setuju pindah Ibu Kota Negara... Yang tidak setuju bodo amat..	positive
cocok ibu kota pindah ke kalimantan apalagi gubernurnya ga bisa kerja. dukung @jokowi	positive
@geedeulbeyou1 Jadi kepada lo yang gak setuju ibu kota di pindah lo itu maunya apa sih. Lo kan tau kalau jakarta itu sudah punya beban yang berat untuk menampung semua kegiatan negara dan bisnis. Lagi pula para pakar pun memprediksi kalau beberapa tahun	positive

Table 2. Result of Topic Modeling with LDA.

Topic ID	Word of Representative and Probability
Topic 0	“0.035*"pindah_kota" + 0.025*"jawa_timur" + 0.025*"tolak_pindah" + 0.016*"konsep_ramah" + 0.014*"tuju" + 0.013*"yg" + 0.013*"tolak" + 0.011*"rakyat" + 0.011*"nusantara" + 0.011*"konsep”
Topic 1	“0.043*"kotaduniauntuksemua_iknsejahterakanindonesia" + 0.018*"pindah_kota" + 0.018*"kotaduniauntuksemua" + 0.017*"nusantara" + 0.016*"masyarakat_jawa" + 0.015*"iknsejahterakanindonesia" + 0.014*"indonesia" + 0.014*"fgelael_maryshelparaiso" + 0.013*"jokowi" + 0.013*"jagat_nusantara"')
Topic 2	“0.032*"nusantara" + 0.022*"manfaat" + 0.018*"tuju" + 0.016*"ekonomi" + 0.014*"timur_dukung" + 0.013*"mantap" + 0.011*"oicoffe_maryshelparaiso" + 0.011*"representasi_bangsa" + 0.011*"kalimantan" + 0.011*"negara"

In studies comparing various topic modeling approaches, BERTopic showed superior performance over traditional methods such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). For example, a study analyzing Twitter posts found that BERTopic and NMF were more effective in handling short and unstructured text data than LDA [24]. In addition, another study comparing BERTopic with LDA, NMF, and Top2Vec in the context of banking showed that BERTopic produced more coherent topics with a coherence score of 0.8463 [23]. The application of BERTopic has also proven effective in various domains. In research on customer experience, BERTopic was used to extract key topics from existing literature, which included service experience, store brand marketing, online shopping, and artificial intelligence. Moreover, in an extractive text summarization task, a hybrid model combining BERTopic and BERT successfully improved the summarization performance with higher ROUGE scores than the standard BERT model. These findings emphasize the flexibility and effectiveness of BERTopic in extracting meaningful insights from various types of text data [25].

3. Result and Discussion

3.1. Topic Modeling Results with LDA

In this research, BoW embedding is used, which is built from text features after pre-processing. The results obtained include the identified topics and their representative words, as well as the grouping of documents into each topic. The following are the results of applying topic modeling with LDA using BoW embedding, which are presented in the Table 1.

Table 1 shows the list of topics identified through the topic modeling process using LDA with BoW representation. In this study, the number of topics was set to 3 (num_topics=3). Each topic is represented by 10 words with the highest probability that reflect the main theme of each group. For example, Topic 0 relates to the issue of moving cities, specifically related to East Java and the concept of environmentally friendly. Words such as “move_city,” “jawa_timur,” and “tolak_pindah” indicate a discussion about moving the capital city and the pros and cons that come with it.

Meanwhile, Topic 1 highlights campaigns and discourses related to the Capital City of the Archipelago (IKN), with words such as “kotaduniauntuksemua,” “iknsejahterakanindonesia,” and “jagat_nusantara,” which indicate the narrative of IKN development as a new growth center.

Furthermore, Topic 2 focuses on the economic benefits and representation of the nation in the context of the capital move, as reflected by words such as “nusantara”, “economy”, “kalimantan”, and “negara”.

Overall, these results provide insights into public perceptions and key themes in discussions about the capital city move, which can be used for further analysis in various aspects of social, economic, and public policy.

Broadly speaking, these results show that in the topic analysis of tweets related to the relocation of the capital city, Topic 0 is the most dominant with 607 tweets, followed by Topic 1 with 453 tweets, and Topic 2 which has

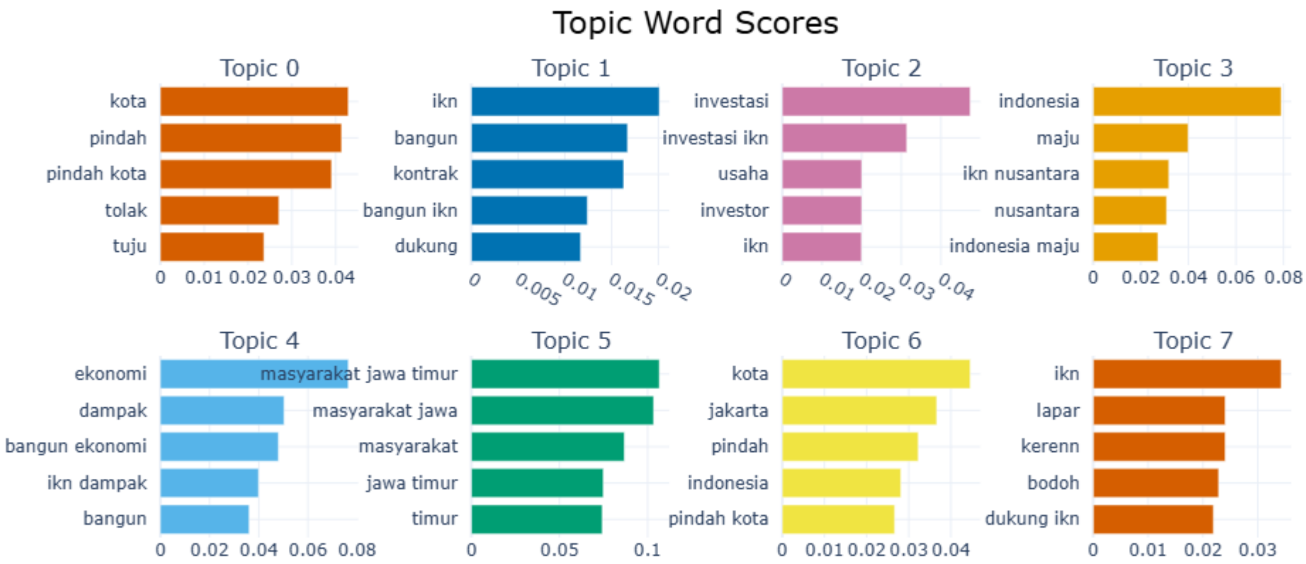

Figure 1. Topic Modeling Results with BERTopic: Topic List

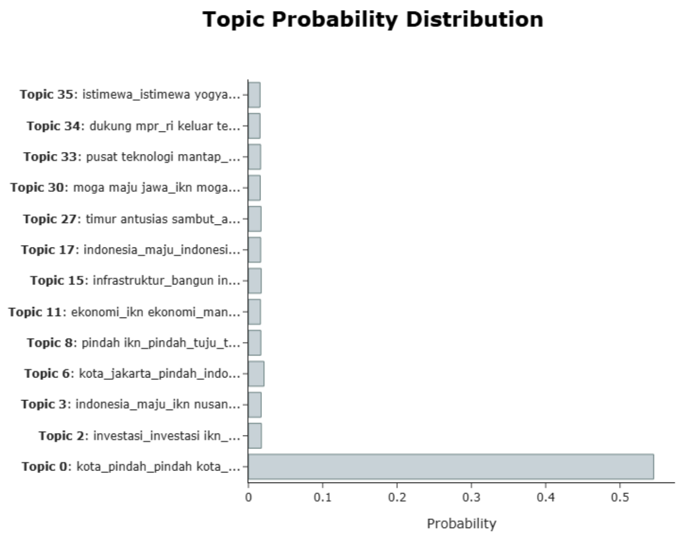

Figure 2. Topic Probability Distribution

404 tweets. This indicates that discussions on Topic 0 appear more frequently than the other two topics. Looking at the previous example, Topic 0 most likely relates to the political and nationalistic aspects of the capital relocation debate, while Topic 1 focuses more on the policy and project implementation, and Topic 2 highlights the reasons for the move such as the problems in Jakarta. With this distribution, it can be concluded that the political debate and national identity related to moving the capital city attracts more attention than the technical aspects or environmental and economic reasons.

Figure 1 shows the results of topic modeling using BERTopic, where each topic has a set of keywords that reflect the main themes in the discussion about the Capital City of the Archipelago (IKN). Topic 0 (city transfer) has a probability of about 0.04, reflecting the conversation about moving the capital city with key words such as "city" and "move". Topic 1 (construction of IKN) with a probability of around 0.02 highlights the construction and contracting aspects of the IKN project. Topic 2 (investment in IKN) with the highest probability of around 0.04 shows the importance of investment and investor involvement in this project. Topic 3 (Indonesia's progress and IKN) also has a probability of around 0.06, emphasizing the narrative that the construction of IKN is a step of national progress. Topic 4 (economic impact of IKN) with a probability of around 0.08 highlights the economic effects, while Topic 5 (perception of East Java community) has a probability of around 0.05, indicating community involvement in this discussion. Topic 6 (impact on Jakarta and Indonesia) with a probability of around 0.04 focuses on the impact of moving the capital city on Jakarta. Finally, Topic 7 (mixed sentiments about IKN) with a probability of around 0.03 shows the diverse opinions of the community, from support to criticism of the project. Overall, the conversation about IKN covers various aspects ranging from development, investment, socio-economic impact, to public perception and subjective sentiments.

Figure 2 shows the distribution of topic probabilities in the BERTopic modeling results, with Topic 0 (city move) having the highest probability of around 0.5. This indicates that most of the discussion is related to the capital move, which likely reflects the pros and cons as well as the impacts. Topic 2 (investment in IKN) has the second highest probability, at around 0.1, indicating that investment and investor engagement is also an important focus in this discourse. Meanwhile, other topics such as Topic 3 (progress of Indonesia and IKN), Topic 6 (impact on Jakarta and Indonesia), and Topic 8 (displacement sentiment) have much lower probabilities, below 0.05, indicating that although they are discussed, they are not at the center of the conversation. The majority of other topics, including Topic 35 (special Yogyakarta) and Topic 34 (support for MPR RI), have very small probabilities, close to zero, meaning that there is relatively little discussion of these aspects in the analyzed data.

Table 3. Topic Modeling Results with LDA: Document Clustering.

Topic	Topic Probability	Text
1	0.7806	['jokowi', 'tuju', 'tuju', 'alokasi', 'angggaran', 'pindah', 'kota', 'anggar', 'nya', 'fokus', 'tangan', 'covid', 'selamat', 'nyawa', 'masyarakat', 'banding', 'fokus', 'proyek', 'pindah', 'kota', 'pindah_kota', 'pindah_kota', 'pindah_kota', 'pindah_kota', 'pindah_kota', 'pindah_kota']
2	0.6934	['hnurwahid', 'fpksdprri', 'tuju', 'kota', 'pindah', 'ga']
0	0.7395	['mardanialisera', 'fpksdprri', 'mayoritas', 'rakyat', 'nkri', 'tuju', 'pindah', 'kota', 'negara', 'tuju', 'bodo', 'pindah_kota', 'pindah_kota', 'pindah_kota']
1	0.7073	['cocok', 'kota', 'pindah', 'kalimantan', 'gubernur', 'ga', 'kerja', 'dukung', 'jokowi']
2	0.7896	['lo', 'gak', 'tuju', 'kota', 'pindah', 'lo', 'mau', 'sih', 'lo', 'tau', 'jakarta', 'beban', 'berat', 'tampung', 'giat', 'negara', 'bisnis', 'pakar', 'prediksi', 'jakarta', 'tenggelam']
0	0.6728	['hukumdan', 'alhamdulillah', 'gerinda', 'tuju', 'pindah', 'kota', 'pindah_kota', 'pindah_kota', 'pindah_kota']

4. Conclusion

This study employs two distinct topic modeling approaches—Bag-of-Words (BoW)-based Latent Dirichlet Allocation (LDA) and BERTopic—to analyze public discussions related to the relocation of Indonesia's capital city. Through LDA, three prominent themes emerge from the dataset: (1) political and nationalistic debates that reflect the ideological tensions and identity issues tied to the capital move, (2) policy implementation and project execution involving administrative decisions, infrastructure development, and intergovernmental coordination, and (3) economic justifications and the multifaceted challenges Jakarta currently faces, including congestion, pollution, and environmental degradation. The topic distribution reveals that conversations centering on politics and national identity dominate public discourse, emphasizing how deeply intertwined the capital relocation is with broader questions of governance and nationalism.

In contrast, BERTopic—by incorporating transformer-based embeddings and clustering techniques provides a more nuanced and granular exploration of the data. This method captures additional themes such as investment strategies, economic impact assessments, infrastructure development timelines, construction progress updates, and public sentiment toward the new capital. Among the highest-probability topics are the relocation itself, investment in Ibu Kota Nusantara (IKN), and the anticipated socio-economic outcomes, indicating that many discussions focus on the project's feasibility and its broader implications for national development.

Together, the findings from LDA and BERTopic offer complementary insights into the central narratives and concerns expressed by the public. These insights are valuable for policymakers, researchers, and stakeholders interested in understanding societal perceptions, guiding communication strategies, and informing evidence-based decision-making related to the ongoing development of the new capital city.

5. Conflicts of Interest

The authors declare no conflicts of interest.

6. References

[1] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," 2003.

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc Natl Acad Sci U S A*, vol. 101, no. SUPPL. 1, pp. 5228–5235, Apr. 2004, doi: 10.1073/pnas.0307752101.

[3] H. Jelodar *et al.*, "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey," 2017, doi: 10.48550/arXiv.1711.04305.

[4] R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining Topic Over Time (TOT), Dynamic Topic Models (DTM), Multiscale Topic Tomography, Dynamic Topic Correlation Detection, Detecting Topic Evolution in scientific literatures, etc. Keywords-Topic Modeling; Methods of Topic Modeling; Latent Semantic Analysis (LSA); Probabilistic Latent Semantic Analysis (PLSA); Latent Dirichlet Allocation (LDA); Correlated Topic Model (CTM); Topic Evolution Model," 2015. [Online]. Available: www.ijacsa.thesai.org

[5] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05794

[6] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.

[7] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” [Online]. Available: https://github.com/UKPLab/

[8] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” [Online]. Available: https://github.com/tensorflow/tensor2tensor

[9] M. A. Mersha, M. Gemeda Yigezu, and J. Kalita, “Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 121–132. doi: 10.1016/j.procs.2024.10.185.

[10] C. Y. Sy, L. L. Maceda, N. M. Flores, and M. B. Abisado, “International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Unsupervised Machine Learning Approaches in NLP: A Comparative Study of Topic Modeling with BERTopic and LDA.” [Online]. Available: www.ijisae.org

[11] F. Bianchi, S. Terragni, and D. Hovy, “Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence.” [Online]. Available: https://github.com/UKPLab/

[12] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, “Overview of the Transformer-based Models for NLP Tasks,” in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 179–183. doi: 10.15439/2020F20.

[13] V. C. Storey and D. E. O’Leary, “Text Analysis of Evolving Emotions and Sentiments in COVID-19 Twitter Communication,” *Cognit Comput*, vol. 16, no. 4, pp. 1834–1857, Jul. 2024, doi: 10.1007/s12559-022-10025-3.

[14] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[15] Y. Ardian Pradana, I. Cholissodin, and D. Kurnianingtyas, “Analisis Sentimen Pemindahan Ibu Kota Indonesia pada Media Sosial Twitter menggunakan Metode LSTM dan Word2Vec,” 2023. [Online]. Available: http://j-ptiik.ub.ac.id

[16] Y. Liu, “Comparison of LDA and BERTopic in News Topic Modeling: A Case Study of The New York Times’ Reports on China,” *Article in Pacific International Journal*, 2024, doi: 10.55014/pij.v7i3.616.

[17] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, “SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis,” in *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, Oct. 2020, pp. 105–114. doi: 10.1145/3340531.3412003.

[18] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05794

[19] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, “Exploring Topic Coherence over many models and many topics,” Association for Computational Linguistics, 2012. [Online]. Available: http://mallet.cs.umass.edu/

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical Dirichlet processes,” *J Am Stat Assoc*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006, doi: 10.1198/016214506000000302.

[21] Y. Miao, E. Grefenstette, and P. Blunsom, “Discovering Discrete Latent Topics with Neural Variational Inference,” 2017.

[22] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.

[23] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, “Comparison of Topic Modelling Approaches in the Banking Context,” *Applied Sciences (Switzerland)*, vol. 13, no. 2, Jan. 2023, doi: 10.3390/app13020797.

[24] R. Egger and J. Yu, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Frontiers in Sociology*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.

[25] Maryanto, Philips, and A. S. Girsang, “Hybrid model for extractive single document summarization: utilizing BERTopic and BERT model,” *IAES International Journal of Artificial Intelligence*, vol. 13, no. 2, pp. 1723–1731, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1723-1731.