Jurnal CoreIT, Vol.11, No.2, December 2025
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Evaluation of the Latent Dirichlet Allocation for Modeling News Topics of Nusantara Capital City

Luh Gede Surya Kartika[1*], Anggara Putu Dharma Putra[2], Komang Rinartha[3], Megawati[4]
[1]Dept. of Informatics, Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, Indonesia
[2]Dept. of Hindu Communication, Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, Indonesia
[3]Dept. of Computer System, STIKOM Bali, Indonesia
[4]Dept. of Information System, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
suryakartika@uhnsugriwa.ac.id*, ankfdd@uhnsugriwa.ac.id, komangrinartha@gmail.com, megawati@uin-suska.ac.id

Abstract. Research regarding topic modeling on the coverage of the Nusantara Capital City (IKN) in national mass media remains limited. This study aims to not only model IKN-related topics but also rigorously evaluate the Latent Dirichlet Allocation (LDA) model to ensure its robustness for future implementation. The dataset comprises 1,498 news articles gathered from prominent Indonesian online media, specifically Detik (1,050 articles) and Kompas (448 articles). The methodology involves experimental variations of LDA parameters, including document volume, maximum features, and topic count, utilizing the Scikit-learn library. The results indicate that an increase in data volume and feature dimensions significantly correlates with longer computation times and a higher number of epochs required for convergence. Furthermore, the expansion of variables and data volume resulted in more negative log-likelihood values and increased perplexity, suggesting that model complexity challenges predictive precision. A convergence threshold of 0.01 was applied to optimize the training cessation point. While this study establishes a baseline for static topic modeling, future research implies the necessity of Dynamic Topic Modeling (DTM) to capture the temporal evolution of topics, a dimension not addressed by the standard LDA model.

Keywords: IKN, LDA, log-likelihood, model evaluation, perplexity

Received November 2024 / **Revised** October 2025 / **Accepted** December 2025

INTRODUCTION

The ratification of the Draft Law on the Nusantara Capital City (IKN) into Law (UU) by the House of Representatives of the Republic of Indonesia and the government was carried out on January 18th, 2022. The day is an important milestone in the development of the Indonesian nation. Thus, Indonesia will have a new Nusantara Capital City to replace the city of Jakarta. The relocation of the IKN has several interests both in terms of economy [1], [2], traffics [3], overpopulation, and administrative efficiency [4], as well as other complex interests [4], [5]. In terms of economic importance, in accordance with Indonesia's vision 2045, the nation's economy will be part of the world's top five in 2045. The estimated GDP per capita in 2045 is US$ 23,119. Thus, Indonesia needs economic transformation to achieve Indonesia's Vision 2045. Economic transformation requires support for the utilization of human resources, infrastructure, regulatory simplification, and bureaucratic reform starting from 2020-2024 so that industrial down streaming is supported. The relocation of the IKN is expected to be able to support and encourage the economic transformation.

Previous studies have extensively examined IKN. These studies generally discuss the sentiment analysis of IKN using various models such as Support Vector Machine [6], Orange application [7], [8], Naïve Bayes [9], [10], Long Short-Term Memory (LSTM) [11], [12] and word2vec [11], Feature Selection [13] and other research. Research related to topic modeling was [8] which found that there were topics such as hajj funds, the topic of the younger generation, the topic of local wisdom, the topic of tourism, and environmental topics related to the transfer of the IKN on Twitter. Nonetheless, acquiring research on the examination of news topic modeling pertaining to the Nusantara capital city in national mass media coverage becomes challenging. According to [14] the national mass media has a central role in conveying information and shaping public opinion regarding the relocation project of the archipelago's capital city. This study not only covered topic modeling, but also evaluated the model for topic analysis to ensure its appropriate use by future researchers or the government.

DOI: 10.24014/coreit.v11i2.33397

Topic modeling is a process that aims to identify and group topics in a large set of text documents automatically. This is important because the increasingly abundant text data often contains scattered information and is difficult to interpret manually. With topic modeling, users can obtain a summary of the topics or themes contained in the document, so that the data analysis process becomes more efficient [15] The most used algorithm for topic modeling is Latent Dirichlet Allocation (LDA), which works by grouping words in documents that have statistical similarities to form a specific topic [16]. LDA assumes that each document is a mixture of several topics, and each topic is a distribution of specific words. In the LDA model, each word in the document is selected based on a topic-dependent multinomial distribution, which allows this model to identify patterns and correlations between words [17]. Although LDA has been shown to be effective in finding latent topics in text data, this model has its drawbacks. Because the training is unsupervised, the results obtained need to be evaluated to ensure the quality of the topics produced [18].

Evaluation of topic models such as LDAs is carried out to determine whether the model is reliable and provides meaningful representation of topics. There are several evaluation methods used, including: (a) Intrinsic Evaluation: Includes an assessment of topic coherence or internal quality of the topic produced. Topic coherence can measure how consistent words are in a topic, usually through techniques such as the eyeball model or the top N words [19]. (b) Extrinsic Evaluation: Focuses on applying the model to a specific task, such as text classification, to assess how well the model solves the task [20].

The state of the art in topic modeling using LDA lies in optimizing model performance through careful adjustment of hyperparameters and computational strategies. Modern implementations, such as those provided in Scikit-learn, apply Bayesian variational inference [21], particularly the Online Variational Bayes algorithm which enables efficient training on large document corpora by processing data in mini-batches and supporting parallel computation. Performance evaluation focuses on monitoring log-likelihood convergence and perplexity as key indicators of model quality. A well-trained model achieves stable log-likelihood values (indicating convergence) and low perplexity (reflecting strong generalization and predictive power). Moreover, research emphasizes finding an optimal balance between dataset size, the number of features (max-features), and topic components (n_components) to improve interpretability and computational efficiency.

Recent advancements also highlight the integration of quantitative metrics (e.g., log-likelihood, perplexity) with qualitative human judgment, evaluating topic coherence through top-word inspection, to validate topic relevance and interpretability. This combination of probabilistic optimization, scalable computation, and hybrid evaluation methods represents the current state of the art in LDA-based topic modeling. Topics in LDA topic modeling are unique because they are based on probabilistic representations. Since the LDA model is unsupervised and still equipped with hyperparameters, it is necessary to evaluate that the model has produced the right results. In this study, in addition to looking at the quality of coherence, it is also important to pay attention to other metrics, such as metrics that measure convergence speed as well as parameters that affect computational time. Understanding these factors helps in refining the model and optimizing the use of topic modeling in various applications, such as sentiment analysis, document grouping, and text classification. This study aims to explore the evaluation metrics for LDA, especially related to topic coherence and computational time, which have important implications in the development of automated text analysis systems.

METHODS

This study applied the LDA model with libraries from Scikit-learn. The organization produces libraries that are the result of community projects, which are made up of a group of people from all over the world. The library of Scikit-learn used in this study is mainly decomposition, especially LDA. The library uses the Bayesian Online Variational algorithm [22]. The model was evaluated using log-likelihood and perplexity. The output is in the form of topics in the document, and the words that make up each topic.

The research conducted as follows: LDA models can be run by using several libraries such as *LDA* from Scikit-learn to create and train LDA models. The *Count Vectorizer* library from Scikit-learn to convert text documents into word matrices. The *matplotlib* library is for visualization of results such as log-likelihood and perplexity, and the *NumPy* library is for numerical calculations. At the document preprocessing stage, using a collection of news containing the word Ibu Kota Nusantara (IKN). This collection of news was obtained from online news in Indonesia, namely Detik and Kompas. The vectorization stage uses the *CountVectorizer* function to transform the news document into a term-document matrix representation,

DOI: 10.24014/coreit.v11i2.33397

where each column represents a word, and each row represents a document. There are several parameters that are set (fine tuning) to affect the results, namely ***max_features, max_df,* and *min_df*** [23].

Max_features parameter was used to limit the number of unique words to be used, which helps limit the complexity of the model. In this study, several experiments were carried out on the size of the *max_features* parameters. A size of *max_features* start from 1000 to 3000, then means that the *CountVectorizer* will only retain the 1000 to 3,000 features of the most frequently occurring words (words or terms) based on their frequency of occurrence throughout the corpus [22]. These features will be selected from all the words found in the document. By limiting the number of features to specific number of words, it reduces the complexity and dimension of the text representation that the LDA model will use. The model will only use the words that are considered the most important or appear most frequently in the corpus, which are most likely more relevant to determine the topic than the words that appear very rarely. The *max_df* parameter was set to 0.95, which means that words that appear in more than 95% of documents will be ignored [22] because they are considered uninformative (e.g., common words such as "and", "in", "which", "to" and so on). The *min_df* parameter was set at 2. Words that appear in fewer than two documents will be ignored or removed from the feature set. In other words, words that appear in only one document will be removed, as they are considered too specific to that document and do not contribute enough to the overall analysis.

The initialization of the LDA model was carried out through several stages. Once the document is converted into a word matrix, the LDA model was initialized using the *LatentDirichletAllocation* library from Scikit-learn. The main parameters in the initialization stage were *n_components, learning_method, and max_iter*. n_components parameter is used to determine the number of topics you want to find. In this study, n_components set at 10. So, the result of the model is 10 news topics. The learning_method parameter in this study was set to be *online*. This means that the LDA model uses Bayesian Variational Online algorithms to process data in small batches. *max_iter* parameter is used to determine the number of iterations to train the model. This research used a *max-iter* of 7400. The hyperparameters in this study, namely *alpha* and *beta*, were set according to the default value of Scikit-learn, which is *1/n_components*. *Alpha* is the parameter of the previous Dirichlet form for the distribution of topics per document, while *beta* will control the distribution of words per topic [22] .

The training stage of the LDA model was carried out by running the training process by calling the function *lda.fit()* on the word matrix (the result of *CountVectorizer*). During the training, log-likelihood and perplexity calculations were also performed to evaluate the model's performance. Log-likelihood is a measure of how well the model describes the data. The higher the log-likelihood, the better the model is at capturing the data structure. Perplexity is another metric that is often used to evaluate the quality of LDA models, and it usually decreases as log-likelihood increases. Lower perplexity values indicate better models [24] When log-likelihood converges, the perplexity will also stop changing or reach the optimal value. Both are indicators that the model has reached an equilibrium point so it does not require longer training. After the model is trained, the process of extracting the topics and words that appear most often in each topic is carried out. Each topic consists of a collection of words with a certain probability of occurrence.

RESULT AND DISCUSSION

Model Evaluation based on Log-likelihood dan Perplexity

Log-likelihood is a measure used in statistics to assess how well a probabilistic model predicts the data observed. It is based on the likelihood or match model, which measures the probability that the observed data (e.g., News in LDA) is generated by the model used. In other words, likelihood measures how likely it is that the data is given certain model parameters. The log-likelihood in LDA is formulated as the sum of the log-likelihood of each document. For each document, likelihood was calculated based on the distribution of topics per document and the distribution of words per topic learned by the model.
In simple terms, the log-likelihood is calculated using the following equation [25]:

$$Log_{likelihood} = \sum_{d=1}^{D} \log P(W_d | \Phi, \alpha) \tag{1}$$

Equation (1) describe as D is the number of documents in the corpus. W_d is a collection of words in a document d. Furthermore, θ_d is the topic distribution for d (topic distribution) documents. The Φ parameter is the word distribution for each topic. Finally $P(W_d / \Phi, \alpha)$ is the probability that the document d is generated from the distribution of the topic θ_d and the distribution of Φ words.

Tabel 1 shows the results of the evaluation of the document. The evaluation was carried out through several experiments on news documents about the IKN in two national mass media, Detik and Kompas. The experiments were carried out by trying several combinations of *max_features* and *n-components*. This experiment is to see the computational time to achieve the convergent log-likelihood condition, the number of iterations to the convergent log-likelihood of the convergence, the magnitude of the log-likelihood and perplexity in the last iteration. To further explain the ability of the LDA model to find the right topic, experiments were carried out up to a maximum of 7400 iterations.

Table 1. The experiment of number of documents, max-features, n-components

News Source	Number of documents	Max-features	n-components	Compute time	Number epoch of convergence	Log-Likelihood at last iteration	Perplexity at last iteration
Detik	1050	1000	10	54 min 59 sec	2748	-989389,36453	409,540963
Detik	1050	2000	10	46 min 31 sec	2246	-1264029,57154	700,304048
Detik	1050	3000	10	1 hr 40 min	3786	-1417555,14344	936,524625
Kompas	448	1000	10	19 min 28 sec	2096	-423467,26324	653,712141
Kompas	448	1500	10	16 min 16 sec	1811	-498486,85924	989,198198
Kompas	448	2000	40	41 min 2 sec	7389	-589774,22501	991,137198

Based on Table 1, an analysis was obtained regarding the changes in the *max-features* parameter and the Number of Data/Documents that affect the computing time, convergence, log-likelihood, and perplexity in the model.

First, the influence of the number of data/documents: documents with a total of 1050 take longer compared to data or documents totaling 448. This is natural because the more data is processed, the longer the computing time will take. The number of epochs needed to achieve convergence tends to be greater in data totaling 1050 compared to data totaling 448. Log-likelihood values for larger data (1050) tend to be more negative, indicating that the model is trying harder to adapt to more complex data. Meanwhile, perplexity values also tended to be lower at 1050 data than at 448 data, suggesting that the model was better able to predict larger data than smaller data.

Second, the effect of *Max-Features* on the same amount of data is 1050: With a fixed number of data (1050), increasing the number of max-features from 1000 to 3000 increases the compute time. In this study, from 54 minutes 59 seconds (1000 features) to 1 hour 40 minutes (3000 features). This indicates that the more features are considered, the more complex the model will be and the longer it will take to calculate it. The number of epochs required for convergence varies, but generally increases with the increasing number of features. This indicates that the model needs more iterations to adapt to more complex features. Log-likelihood tends to be increasingly negative with an increase in the number of max-features. This can mean that the model has more parameters to optimize so it tends to give more negative probabilities. Meanwhile, the perplexity value is also increasing, which suggests that with more features, it is increasingly difficult for the model to make accurate predictions.

Third, the effect of *Max-Features* on smaller data, i.e. 448: on smaller data (448), the increase in *max-features* from 1000 to 2000 still shows an increase in compute time, although not as large as the 1050 data. The increase from 1000 to 2000 max-features increases the compute time from 19 minutes 28 seconds to 41 minutes 2 seconds. The number of epochs for convergence increased quite significantly from 2096 (1000 features) to 7389 (2000 features), suggesting that the model required more iterations to adapt to a larger number of features. Log-likelihood is getting more negative with the increase in max-features, suggesting that the addition of features adds to the complexity of the model. The perplexity value also increases with the increase in max-features, which indicates that the accuracy of the model's predictions decreases. Figure 1 and Figure 2 show the log-likelihood and perpelexity sequentially in the news documents about the IKN in the Detik media.

DOI: 10.24014/coreit.v11i2.33397

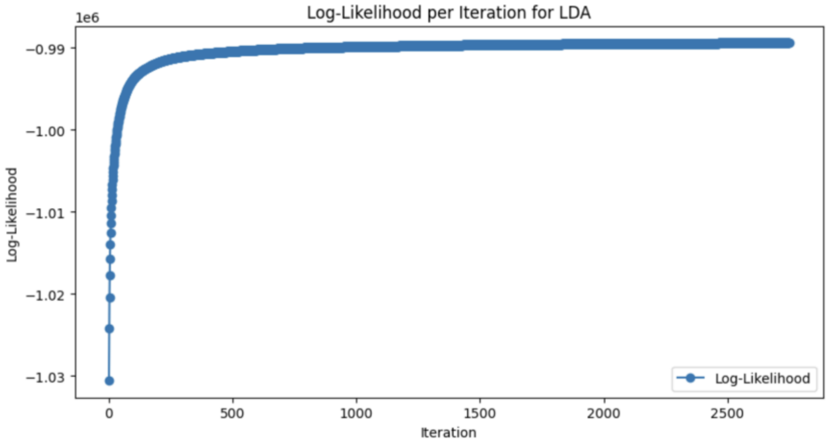

Figure 1. Log-Likelihood on data from Detik.com

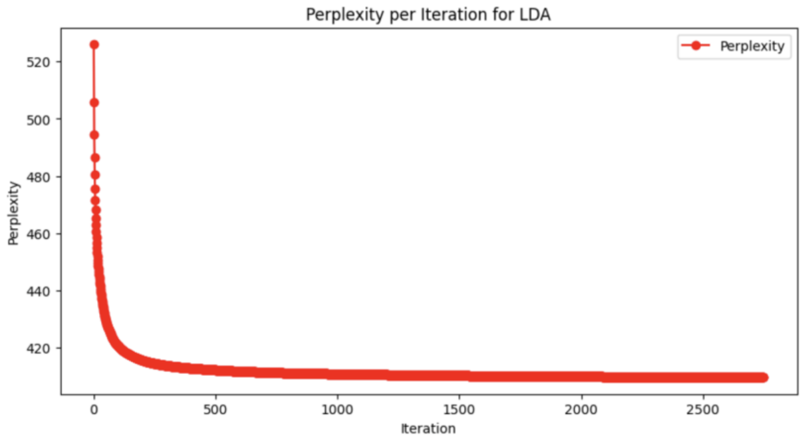

Figure 2. Perplexity on data from Detik.com

An experiment was also carried out on IKN news data in Kompas.com with the number of documents or news about IKN as many as 448. The parameter condition of the number of topics is 10, the number of features is 1500, then the log-likelihood and perplexity are obtained as shown in Figure 3 and Figure 4.

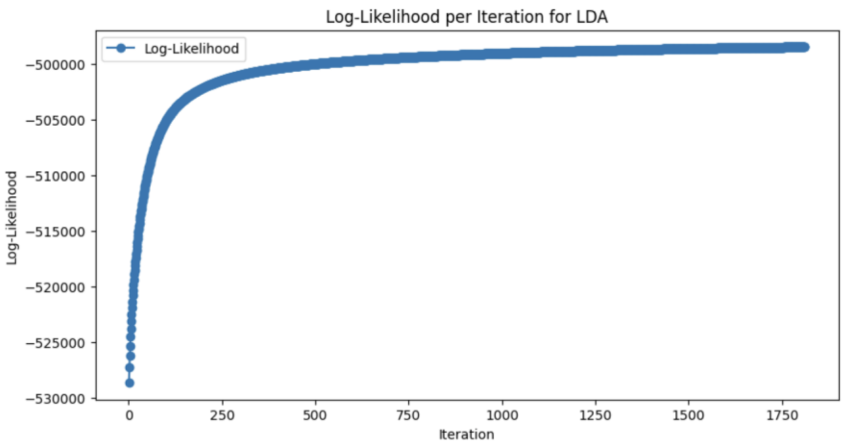

Figure 3. Log-Likelihood on data from Kompas.com

DOI: 10.24014/coreit.v11i2.33397

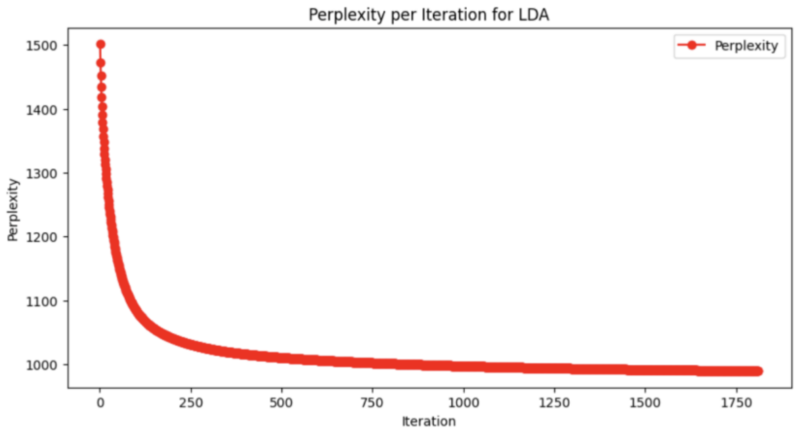

Figure 4. Perplexity on data from Kompas.com

The experiment led to discussion that are convergent log-likelihood is a state in which log-likelihood (i.e., a measure of how well a probabilistic model explains data) stops changing significantly after a certain number of iterations in the model training process. In the context of a model such as LDA, log-likelihood is measured at each iteration, and the goal is to maximize this value. Figure 1, Figure 2, Figure 3, Figure 4 meet the expectation that when the log-likelihood increases (towards zero), then the perplexity value decreases, which indicates that the model is getting better at predicting the words in the document. A high log-likelihood (closer to zero) generally indicates that the model has a good ability to adjust the data it has been trained on, while a low perplexity indicates that the model has good generalization ability for new data. Perplexity is often used in the evaluation of LDA models on test data to see if the model is overfitting or not. If the perplexity is high, it means that the model is having trouble predicting new data, which could indicate overfitting.

Log-likelihood convergence occurs when the log-likelihood value is stable, i.e. the difference between the log-likelihood in the *n*th iteration and the *n*+1 iteration is small (smaller than the predefined threshold, i.e. 1e-2). At this point, it can be said that the models have converged. This indicates that the model has reached an optimal point, where updates to the model's parameters (such as topic distribution in the document and word distribution in topic) no longer result in a significant improvement in the model's fit to the data. Convergence shows that the model has learned from the data well [26]. When log-likelihood is already convergent, stopping further iterations is usually done, as continuing the iteration will not provide a significant benefit.

This study used Skicit-Learn, especially the *LatentDirichletAllocation* library. According to [22] the LDA model uses Bayesian variational optimization algorithms. The training process involved iterations where the model's parameters were updated at each step. Bayes' Online Variational algorithm makes it possible to break the process into small sizes (batches) [27], as well as perform parallel data processing [21]. This algorithm uses stochastic optimization, which is an optimization approach that uses only a portion of the data (or mini batch) of the entire document corpus in each iteration. Instead of processing the entire corpus at once, the algorithm only takes a small sample or subset of the document at each step. This method helps reduce computational time and allows for faster model updates.

Model Evaluation based on human judgement
Often the easiest way to evaluate is to just look at the output of the model [28] Evaluating the quality of the model by examining the top words of each topic is the simplest way to see the quality of the model. Table 2 shows the distribution of words on each topic generated. The parameters in the results are *max_features*=1000 and *n_components*=10 with the total amount of detik.com data amounting to 1050.

Table 2. The comparison of topics generated in the first and last epoch on documents from Detik

First Epoch	Last Epoch (2748th)
Topic 0: ikn pembangunan gibran rp triliun jokowi kota indonesia presiden ibu	Topic 0: gibran prabowo ikn ganjar pembangunan cawapres program **pak** mahfud presiden
Topic 1: kota ikn bandara makassar tanah ibu masyarakat baik operasional **orang**	Topic 1: indonesia ekonomi jokowi ibu tahun masyarakat kota investasi kerja presiden

First Epoch	Last Epoch (2748th)
Topic 2: ikn jokowi nusantara groundbreaking proyek hotel investasi kota investor ibu	Topic 2: ikn nusantara jokowi hotel investor proyek groundbreaking investasi ibu pembangunan
Topic 3: indonesia tahun hari ekonomi ibu nasional kota rp **ma** sebesar	Topic 3: hari rp tahun kota jakarta biaya bangunan indonesia juta harga
Topic 4: ikn kota pembangunan nusantara bangunan ibu rumah dibangun indonesia gedung	Topic 4: ikn kota nusantara pembangunan lingkungan ibu kawasan bangunan otorita oikn
Topic 5: ikn asn kementerian kota pemerintah nusantara negara tahun ibu pindah	Topic 5: ikn asn pemerintah jakarta rp tahun triliun negara kota kementerian
Topic 6: pks ibu ganjar **mengatakan tetap** ikn kota jakarta prabowo nusantara	Topic 6: pks ibu kota jakarta undang **mengatakan tetap** ikn sohibul kampanye
Topic 7: jalan hari ikn tol kota listrik balikpapan kendaraan kawasan bus	Topic 7: ikn balikpapan kota kalimantan air jalan tol bandara km timur
Topic 8: kota ikn jakarta anies undang **baru** ibu indonesia pemerataan imin	Topic 8: kota ikn anies indonesia baru imin pembangunan cak membangun pemerataan
Topic 9: desember pembangunan ikn rusun negara juli masjid tower ibu pekerjaan	Topic 9: ikn pembangunan masjid kantor infrastruktur negara jokowi rumah pekerjaan Menteri

Table 2 shows that in the first iteration, the focus of the topic was more general, leading to the infrastructure and physical development aspects of the IKN. The role of political figures is not very prominent. In the 2748th iteration, some topics underwent an evolution with a greater emphasis on political issues and broader economic impacts. There are more mentions of political figures such as Gibran, Prabowo, Ganjar, and Mahfud, as well as details on investment and development costs. Overall, the 2748th iteration appears to be more specific, with more mature topics focused on more complex issues, such as politics, economics, and development costs. This shows that the model appears to have successfully constructed a topic that has a specific distribution of words on that topic.

Table 3 shows the results of topics and the distribution of topical words in documents/news in Kompas.com. The experimental parameters on the document are *max_features*=1000 and *n_components*=10 and the data totals 448 documents.

Table 3. The comparison of topics generated in the first and last epoch on documents from Kompas

First Epoch	Last Epoch (2096th)
Topic 0: tol jalan persen seksi infrastruktur kementerian danis segmen pembangunan kariangau	Topic 0: tol jalan danis pembangunan seksi proyek infrastruktur segmen kementerian kkt
Topic 1: pembangunan nusantara tahap infrastruktur kawasan presiden istana ibu air pekerjaan	Topic 1: pembangunan presiden bandara tahap istana persen proyek vvip nusantara infrastruktur
Topic 2: kalimantan tahun balikpapan lebaran april pertahanan kendaraan persen nusantara timur	Topic 2: balikpapan kendaraan tol samarinda mudik penumpang bandara lintas hari jalan
Topic 3: nusantara ibu oikn balikpapan pembangunan bambang otorita indonesia hotel baru	Topic 3: nusantara pembangunan oikn ibu indonesia otorita teknologi bambang lingkungan tahun
Topic 4: balikpapan persen kaltim kalimantan samarinda provinsi tni timur daerah tahun	Topic 4: balikpapan persen kaltim kalimantan **sebesar** pertumbuhan harga timur tahun inflasi
Topic 5: pembangunan ppu masyarakat warga kabupaten pemerintah nusantara tahun oikn ibu	Topic 5: ppu masyarakat warga kabupaten pembangunan sepaku oikn penajam tanah pemerintah
Topic 6: rp proyek triliun **sebesar** pt nusantara pembangunan miliar tahun ibu	Topic 6: rp triliun investasi miliar nusantara groundbreaking proyek tahun agung Bambang
Topic 7: investasi proyek agung industri groundbreaking suara timur teknologi balikpapan nusantara	Topic 7: proyek investasi balikpapan kalimantan timur industri tahun air bkpm rp
Topic 8: asn kementerian tower pembangunan rusun tanah investasi bkpm rp unit	Topic 8: asn jakarta nusantara ibu presiden menteri rusun jokowi tower jadi
Topic 9: rumah pembangunan nusantara menteri ibu rtjm unit persen pekerja konstruksi	Topic 9: rumah pembangunan konstruksi unit rtjm menteri pekerjaan pupr jabatan nusantara

The results of human observation show that the distribution of words on each topic in Table 3 shows that the results of iteration 1 and the last iteration (2096th) show a relatively relevant distribution of words. Each topic undergoes changes that reduce the appearance of common words, narrowing the topic to be more relevant to a specific context, such as transportation, economic growth, and the role of government in regional development. Although LDA has long been established as a standard method for topic analysis [16], [29], it exhibits several limitations compared to more recent approaches such as BERTopic and neural topic models. LDA relies on a bag-of-words representation that ignores word order and semantic context, making it less capable of capturing deeper conceptual meanings or cross-topic semantic relationships. In the context of media or policy text analysis, this limitation may reduce the model's ability to interpret complex narratives or framing structures.

In contrast, BERTopic[30] integrates sentence embeddings derived from transformer-based models (e.g., BERT or RoBERTa) with clustering algorithms, allowing for the generation of topics that are more contextual and semantically coherent. This approach has been shown to more effectively capture nuanced meanings and linguistic variations, particularly in heterogeneous corpora such as news texts. Meanwhile, neural topic models developed using variational autoencoders [31] or deep generative frameworks can learn more complex and non-linear topic distributions, thereby addressing the independence assumptions inherent in LDA.

From an evaluation perspective, metrics such as perplexity and coherence score should not be interpreted purely as computational measures. A lower perplexity value indicates a better model fit to the data, but it does not necessarily correspond to topic interpretability [19]. Similarly, a higher coherence score suggests stronger internal consistency among words but may not reflect semantic relevance within the research context. Therefore, topic model evaluation should include both quantitative and qualitative dimensions, considering factors such as the interpretability of topics by domain experts and their alignment with the policy or communication issues under study [32].

Considering these limitations, LDA remains a transparent and interpretable baseline method. However, its analytical results should ideally be complemented by semantically informed or neural approaches to achieve a deeper and more meaningful understanding of media discourse structures.

CONCLUSION

This study demonstrates that both the number of documents and the max-features parameter significantly affect the performance of the LDA model. Increasing the number of documents leads to longer computation times and more iterations to reach convergence due to higher data complexity. Although the log-likelihood values become more negative for larger datasets, the lower perplexity indicates better predictive capability. Similarly, increasing the number of features raises computation time and the number of epochs required for convergence. The increasingly negative log-likelihood and higher perplexity values suggest that excessive features add complexity and reduce model accuracy. Hence, an appropriate balance between dataset size and feature count is crucial for optimal performance.

Convergence of log-likelihood, occurring when changes between iterations fall below a defined threshold 0.01, indicates that the model has reached stability and no longer benefits from further updates. The consistent pattern between higher log-likelihood (closer to zero) and lower perplexity confirms that the model effectively learns word distributions and generalizes well to unseen data. Qualitative evaluation based on human judgment also shows that the model progressively produces more specific and coherent topics across iterations, shifting from general discussions toward focused issues such as politics, economics, and infrastructure.

Overall, the findings highlight the importance of tuning data size, feature selection, and training parameters to build an efficient and well-converged LDA model with stable log-likelihood, low perplexity, and meaningful topic representations.

ACKNOWLEDGEMENTS

Gratitude to UHN I Gusti Bagus Sugriwa Denpasar, especially the Unit of Research and Community Service, the Hindu Communication Science Study Program, and the Informatics Study Program, for the support and permission to conduct this research. This research was carried out in accordance with the Research Implementation Assignment Letter Number: 1551/Uhn.01/1/04/2024 Fiscal Year 2024. This research was funded by DIPA UHN I Gusti Bagus Sugriwa Denpasar Number: SP DIPA – 025.07.2.552762/2024 dated November 24, 2023.

REFERENCES

[1] S. D. Saputra, T. G. J., and M. Halkis, “Analisis Strategi Pemindahan Ibu Kota Negara Indonesia Ditinjau dari Perspektif Ekonomi Pertahanan,” *192 | Jurnal Ekonomi Pertahanan* |, vol. 7, no. 2, 2021.

[2] R. R. A. Hasibuan and S. Aisa, "Dampak dan Resiko Perpindahan Ibu Kota Terhadap Ekonomi di Indonesia," *AT-TAWASSUTH: Jurnal Ekonomi Islam*, vol. 5, no. 1, 2020, doi: 10.30829/ajei.v5i1.7947.

[3] R. Cybriwsky and L. R. Ford, "City profile Jakarta," *Cities*, vol. 18, no. 3, 2001, doi: 10.1016/S0264-2751(01)00004-X.

[4] Y. S. Amelinda, R. A. Wulandari, and A. Asyary, "The effects of climate factors, population density, and vector density on the incidence of dengue hemorrhagic fever in South Jakarta Administrative City 2016-2020: an ecological study," *Acta Biomedica*, vol. 93, no. 6, 2022, doi: 10.23750/abm.v93i6.13503.

[5] R. Setiadi, J. Baumeister, P. Burton, and J. Nalau, "Extending Urban Development on Water: Jakarta Case Study," *Environment and Urbanization ASIA*, vol. 11, no. 2, 2020, doi: 10.1177/0975425320938539.

[6] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[7] H. Dhery, A. Assyam, and F. N. Hasan, "Analisis Sentimen Twitter Terhadap Perpindahan Ibu Kota Negara Ke IKN Nusantara Menggunakan Orange Data Mining," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, 2023.

[8] Y. Sunesti and A. K. Putri, "Narasi Ibu Kota Negara Baru di Twitter: dari Isu Kearifan Lokal Hinggi Isu Anak Muda," in *Konferensi Nasional Sosiologi IX APSSI 2022 Balikpapan*, 2022, pp. 53–60.

[9] S. D. Prasetyo, S. S. Hilabi, and F. Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *Jurnal KomtekInfo*, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[10] S. Anggraeni and S. D. Saraswati, "Klasifikasi Sentimen Terhadap Ibu Kota Nusantara (IKN) pada Media Sosial Menggunakan Naive Bayes," *Teknika*, vol. 16, no. 2, 2022.

[11] F. Nurdiyansyah and L. U. Pratama, "Analisis sentimen perpindahan ibu kota negara pada aplikasi Tiktok menggunakan metode LSTM," *Teknosains: Media Informasi dan Teknologi*, vol. 17, no. 3, pp. 382–387, 2023.

[12] Y. Ardian Pradana, I. Cholissodin, and D. Kurnianingtyas, "Analisis Sentimen Pemindahan Ibu Kota Indonesia pada Media Sosial Twitter menggunakan Metode LSTM dan Word2Vec," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 5, pp. 2389–2397, 2023, [Online]. Available: http://j-ptiik.ub.ac.id

[13] F. Zamachsari, G. V. Saragih, Susafa'ati, and W. Gata, "Analisis Sentimen Pemindahan Ibu Kota Negara dengan Feature Selection," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 3, 2020.

[14] B. Dame Laoera and T. O. Wibowo, "Indonesian online news and digital culture: a media ecology perspective," *Jurnal Studi Komunikasi (Indonesian Journal of Communications Studies)*, vol. 7, no. 2, pp. 355–368, Jul. 2023, doi: 10.25139/jsk.v7i2.6190.

[15] Y. Matira and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," *Estimasi: Journal of Statistics and Its Application*, vol. 4, no. 1, pp. 2721–379, 2023, doi: 10.20956/ejsa.vi.24843.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4–5, 2003, doi: 10.7551/mitpress/1120.003.0082.

[17] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," in *Handbook of Latent Semantic Analysis*, Laurence Erlbaum, 2007.

[18] D. Newman, E V Bonilla, and W. Buntine, "Improving Topic Coherence with Regularized Topic Models. Improving Topic Coherence with Regularized Topic Models," in *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 2011. [Online]. Available: https://www.researchgate.net/publication/260639294

[19] I. C. Chang, T. K. Yu, Y. J. Chang, and T. Y. Yu, "Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals," *Sustainability (Switzerland)*, vol. 13, no. 19, 2021, doi: 10.3390/su131910856.

[20] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.

[21] H. Valpola and A. Honkela, "On-Line Variational Bayesian Learning," 2003. [Online]. Available: http://www.cis.hut.fi/projects/ica/bayes/

[22] Scikit Learn, "LatentDirichletAllocation," Scikit Learn. Accessed: Oct. 08, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html#sklearn.decomposition.LatentDirichletAllocation

[23] O. M. AL-Janabi, N. H. Ahamed Hassain Malim, and Y. N. Cheah, "Unsupervised model for aspect categorization and implicit aspect extraction," *Knowl Inf Syst*, vol. 64, no. 6, pp. 1625–1651, Jun. 2022, doi: 10.1007/s10115-022-01678-5.

[24] Y. O. Santoso and S. A. Nugroho, "Pengelompokkan Jurnal Ilmiah Berdasarkan Judul Menggunakan LDA," vol. 3, no. 1, pp. 31–42, 2019.

[25] Quentin Pleple, "Perplexity To Evaluate Topic Models," https://qpleple.com/perplexity-to-evaluate-topic-models/.

[26] P. A. Telnoni, Suryatiningsih, and E. Rosely, "Pelabelan Data Dengan Latent Dirichlet Allocation dan K-Means Clustering pada Data Twitter Menggunakan Bahasa Indonesia," *Jurnal Elektro dan Telekomunikasi Terapan*, vol. 7, no. 2, pp. 885–892, Mar. 2020, doi: 10.25124/jett.v7i2.3442.

[27] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic Variational Inference," *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, 2013.

[28] Matti Lyra, "Evaluating Topic Models." Accessed: Nov. 01, 2024. [Online]. Available: https://mattilyra.github.io/2017/07/30/evaluating-topic-models.html

[29] J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of Topic Models," *Foundations and Trends® in Information Retrieval*, vol. 11, no. 2–3, pp. 143–296, 2017, doi: 10.1561/1500000030.

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

[31] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," *Trans Assoc Comput Linguist*, vol. 8, pp. 439–453, Jul. 2020, doi: 10.1162/tacl_a_00325.

[32] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017, pp. 165–174. doi: 10.1109/DSAA.2017.61.