Journal of Emerging Information Systems and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

Topic Modeling on Social Media and News Website Data Concerning the Electronic-Based Government System in Indonesia

Nanda Eka Hermawanto[1], Wiyli Yustanti[2]
[1,2]*State University of Surabaya, Surabaya, Indonesia*
nanda.19085@mhs.unesa.ac.id, wiyliyustanti@unesa.ac.id

ABSTRACT

The rapid growth of social media and news websites has generated a massive amount of unstructured text data on the internet, including public opinions regarding the Electronic-Based Government System (SPBE) in Indonesia. To discover hidden information and topics within this textual dataset, an effective analysis technique such as topic modeling is required. This study aims to map the discussion topics on social media (Twitter, Tiktok) and news websites related to SPBE, and to analyze the implementation of topic modeling using the Latent Dirichlet Allocation (LDA) method. This research employs a computational approach by applying the LDA algorithm to a text corpus consisting of 3,327 datasets. Model quality evaluation was conducted through experiments across 12 modeling scenarios using the Coherence Score metric. The results indicate that Scenario 12 is the most optimal model, achieving the highest Coherence Score of -1.234, which generated 5 main topics. This peak performance was reached under the configuration of K=5 topics, 100 iterations, an Alpha value of 0.001, and a Beta value of 0.001. The top five words with the highest weights dominating the dataset are "aplikasi" (145 times), "spbe" (95 times), "setuju" (51 times), "sistem" (51 times), and "elektronik" (47 times). In conclusion, this study proves that the LDA method is capable of coherently extracting latent topic structures, revealing that public focus is not only centered on the technical form of the SPBE application, but also highlights the system evaluation process and its impact on specific sectors such as educators (teachers).

Keyword: Topic Modeling, Latent Dirichlet Allocation (LDA), SPBE, Social Media, Coherence Score.

Article Info:
Article history:
Received July 10, 2026
Revised August 11, 2026
Accepted September 15, 2026

Corresponding Author
Nanda Eka Hermawanto
State University of Surabaya, Surabaya, Indonesia
nanda.19085@mhs.unesa.ac.id

1. INTRODUCTION

Information and Communication Technology (ICT) plays an important role in society. The development of ICT can improve the effectiveness and efficiency of government systems, which leads to better public services. The Electronic-Based Government System (SPBE), also known as e-Government, is the government's effort to utilize information and communication technology to support government systems and improve the effectiveness of government performance. The implementation of e-Government in a country will certainly encounter many obstacles. If a country cannot overcome these barriers, it may lead to implementation failure [1]. SPBE is expected to increase the productivity and efficiency of bureaucratic processes, as well as support the economic growth process. The concept of SPBE itself is a stage to improve

interactions between the government and the community (Government to Citizen), the government and the business sector (Government to Business Enterprise), and relationships among government entities (Inter-Agency Relationship) [2].

Social media is an online medium where its users can easily participate, share, and create content, including blogs, social networks, wikis, forums, and virtual worlds. Blogs, social networks, and wikis are the most common forms of social media used by people worldwide. Another perspective states that social media is an online medium that supports social interaction, utilizing web-based technologies to turn communication into interactive dialogues [3]. According to Andika [4], social media is a platform that allows its users to represent themselves by sharing, interacting, communicating with other users, and forming virtual social bonds using the internet.

With the increasing amount of unstructured text on the internet, a tool is needed to help discover hidden topics within a collection of texts. To address this issue, researchers have developed a technique called topic modeling. Topic modeling is the clustering of text data based on specific topics. It aims to obtain the distribution of words that make up a topic and documents associated with that specific topic. Topic modeling generally consists of two stages. The first stage is conducting topic modeling based on the addition and subtraction of the number of topics. The second stage is performing topic modeling based on the number of iterations [5].

This research uses the LDA method to determine topic trends on social media platforms such as Twitter and TikTok, as well as news sites related to SPBE. Latent Dirichlet Allocation (LDA) is a method for detecting topics in a dataset and performing topic modeling. Latent Dirichlet Allocation (LDA) is currently the most popular topic modeling and analysis method, widely used for analyzing massive documents. LDA is frequently utilized today because it can cluster, summarize, connect, and process data by assigning weights to each document, which subsequently generates a list of topics [6].

2. METHODS

The data used in this study are primary data sourced from the opinions and comments of the Indonesian public on social media platforms, specifically Twitter and TikTok. The data extraction process was conducted using a crawling technique with the search keyword "spbe" (Electronic-Based Government System). The data collection period was focused on the timeframe between 2023 and 2024. All extracted data were subsequently stored in a .csv format. The initial crawling results generated several metadata attributes, such as id, full_text, image_url, location, retweet_count, and created_at. However, because this research focuses on topic modeling, a feature selection stage was conducted to reduce the data, utilizing only the main text column (full_text). Each row in the full_text column represents a single document. Prior to the modeling stage using the LDA algorithm, the raw data underwent a text preprocessing phase to remove noise, such as URLs, username mentions, and other irrelevant characters.

2.1 Selection

At this stage, data selection for the topic modeling process was conducted. The utilized data consisted of public comments regarding the Electronic-Based Government System (SPBE), sourced from social media platforms and online news articles. The data acquisition process was carried out through a crawling technique using a custom Python-based application.

This application was programmed to automatically extract relevant texts from the designated sources, which were then directly stored into a database to serve as the primary dataset.

2.2 Preprocessing

At this stage, a data cleaning procedure was performed. Because the appropriate cleaning method depends strictly on the data format, and this study utilized text data, the purification process was executed through a rigorous text preprocessing pipeline. This pipeline consisted of several sequential stages. First, cleansing was performed to filter the documents by removing duplicate terms and non-alphabetical characters such as numbers, symbols, and URLs. Next, tokenizing separated the text strings into individual tokens. This was followed by case folding, which converted all uppercase letters into lowercase to ensure word uniformity. Subsequently, stopword removal filtered out irrelevant words using a predefined stoplist. Finally, stemming was applied to reduce the remaining words to their base or root forms, eliminating any affixes commonly found in social media and news texts.

2.3 Transformation

Following the text preprocessing phase, the transformation stage involved extracting essential information from the refined data. The primary objective of this procedure was to enhance both the effectiveness and the overall accuracy of the subsequent data processing.

2.4 Data Mining

In this phase, the appropriate data mining technique was selected. Following the completion of various preprocessing procedures to extract the required data components, the study proceeded to the training process of the Latent Dirichlet Allocation (LDA) model.

2.5 Evaluation

In this final phase, the evaluation and interpretation of the topics generated by the LDA model were conducted. The outcomes of this algorithmic topic modeling were visualized through business trend charts and word distribution graphs to provide deeper insights. Furthermore, the generated topic model was tested and validated using the real-world news data that had been previously collected.

2.6 Knowledge

In the final step, the extracted information was visualized to enhance comprehension. Specifically, word clouds were utilized to provide an intuitive visual interpretation of the topic modeling results. This visualization approach serves as a highly valuable tool for data scientists and researchers in the field of Natural Language Processing (NLP) to effectively communicate complex textual patterns.

3. RESULTS AND DISCUSSION

3.1 Result

3.1.1 Selection

Data acquisition was conducted through a crawling technique utilizing Tweet Harvest for Twitter and Apify for TikTok. For the Twitter dataset, Indonesian-language tweets were collected between July 2023 and October 2023 using the keywords “spbe” and “sistem pemerintahan berbasis elektronik”, yielding an initial total of 309 records. After filtering out duplicate and null entries, the dataset was reduced to 241 valid records. Conversely, the

TikTok dataset consisted of Indonesian-language comments extracted from a post by the user @cecesunandar163, generating 3,150 raw data points. Following a similar data cleaning process to eliminate duplicates and missing values, a total of 2,856 valid records were retained for further analysis.

3.1.1.1 Preprocessing

The sequential steps involved in the preprocessing phase are detailed as follows. Figure 1 illustrates the flowchart of this text preprocessing pipeline.

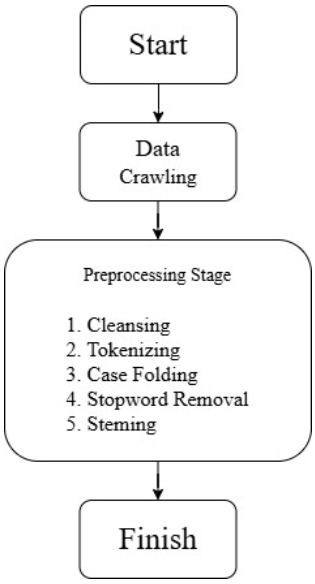


Figure 1 Preprocessing Flow

3.1.1.1.1 Cleansing

During this initial phase, the text was cleansed of irrelevant elements that could potentially disrupt the analysis process. This procedure involved the rigorous removal of non-alphabetical characters, including punctuation marks, numbers, symbols, URLs, and other artifacts that do not contribute to the contextual understanding of the documents. The primary objective of this cleansing step was to generate highly refined and structured text data, ensuring consistent processing in the subsequent stages. Furthermore, this process effectively minimized data noise, which could otherwise degrade the performance and accuracy of the topic modeling algorithm.

Table 1 Cleansing Comparison

Before Cleansing	After Cleansing
Siang tadi Menteri PANRB Abdullah Azwar Anas menerima audiensi Menteri KLHK @SitiNurbayaLHK. Pertemuan ini membahas terkait reformasi birokrasi kelembagaan SPBE dan SDM aparatur di lingkungan Kementerian LHK. __ #KemenPANRB #ASNBerAKHLAK	Siang tadi Menteri PANRB Abdullah Azwar Anas menerima audiensi Menteri KLHK Pertemuan ini membahas terkait reformasi birokrasi kelembagaan SPBE dan SDM aparatur di lingkungan Kementerian LHK

3.1.1.1.2 Tokenizing

During the tokenizing stage, the text documents were segmented into individual word units or tokens. This procedure aimed to separate the text based on specific delimiters, primarily spaces, since punctuation marks had already been eliminated during the prior cleansing phase. Tokenization served as the fundamental basis for all subsequent analytical steps, given that most text mining algorithms operate at the word level rather than the sentence level. Consequently, this step enabled each distinct word to be isolated and treated as an independent feature within the data representation.

Table 2 Tokenizing Comparison

Before Tokenizing	After Tokenizing
Siang tadi Menteri PANRB Abdullah Azwar Anas menerima audiensi Menteri KLHK Pertemuan ini membahas terkait reformasi birokrasi kelembagaan SPBE dan SDM aparatur di lingkungan Kementerian LHK	Siang, tadi, Menteri, PANRB, Abdullah, Azwar, Anas, menerima, audiensi, Menteri, KLHK, Pertemuan, ini, membahas, terkait, reformasi, birokrasi, kelembagaan, SPBE, dan, SDM, aparatur, di, lingkungan, Kementerian, LHK

3.1.1.1.3 Case Folding

The case folding stage was applied to convert all uppercase characters into lowercase letters. In text processing, this procedure is highly crucial, particularly to ensure that identical words are treated uniformly, regardless of whether they are written in uppercase or lowercase formats.

Table 3 Case Folding Comparison

Before Case Folding	After Case Folding
Siang, tadi, Menteri, PANRB, Abdullah, Azwar, Anas, menerima, audiensi, Menteri, KLHK, Pertemuan, ini, membahas, terkait, reformasi, birokrasi, kelembagaan, SPBE, dan, SDM, aparatur, di, lingkungan, Kementerian, LHK	siang, tadi, menteri, panrb, abdullah, azwar, anas, menerima, audiensi, menteri, klhk, pertemuan, ini, membahas, terkait, reformasi, birokrasi, kelembagaan, spbe, dan, sdm, aparatur, di, lingkungan, kementerian, lhk

3.1.1.1.4 Stopword Removal

In this stage, stopword removal was executed to eliminate non-essential words from the text. The primary objective of this filtration process was to reduce the volume of uninformative terms, ensuring that only relevant and meaningful words were retained for the analysis. Furthermore, this procedure significantly enhanced computational efficiency during the modeling phase and improved the overall quality of feature representation within the documents.

Table 4 Stopword Removal Comparison

Before Stopword Removal	After Stopword Removal
siang, tadi, menteri, panrb, abdullah, azwar, anas, menerima, audiensi, menteri, klhk, pertemuan, ini, membahas, terkait, reformasi, birokrasi, kelembagaan, spbe, dan, sdm, aparatur, di, lingkungan, kementerian, lhk	siang, tadi, menteri, panrb, abdullah, azwar, anas, menerima, audiensi, menteri, klhk, pertemuan, membahas, terkait, reformasi, birokrasi, kelembagaan, spbe, sdm, aparatur, lingkungan, kementerian, lhk

3.1.1.1.5 Stemming

In this study, the stemming process was executed using an Indonesian language stemmer. The primary objective of this procedure was to reduce morphological variations of words with similar meanings, ensuring that terms derived from the same root were consistently represented as a single feature. This stage was crucial for reducing data dimensionality and enhancing the overall quality of the topic modeling in the subsequent phase.

Table 5 Stemming Comparison

Before Stemming	After Stemming
siang, tadi, menteri, panrb, abdullah, azwar, anas, menerima, audiensi, menteri, klhk, pertemuan, membahas, terkait, reformasi, birokrasi, kelembagaan, spbe, sdm, aparatur, lingkungan, kementerian, lhk	siang, tadi, menteri, panrb, abdullah, azwar, anas, terima, audiensi, menteri, klhk, temu, bahas, kait, reformasi, birokrasi, lembaga, spbe, sdm, aparatur, lingkungan, menteri, lhk

3.1.2 Transformation

The purpose of using bigram and trigram models is to discover and understand the meaning formed by the combination of two or three words. In text analysis, this approach is crucial because the meaning of an expression or phrase can often only be fully understood when the words are viewed together, rather than individually.

Table 6 Bigram and Trigram Results

Clean Text	Bigram	Trigram
siang, tadi, menteri, panrb, abdullah, azwar, anas, menerima, audiensi, menteri, klhk, pertemuan, membahas, terkait, reformasi, birokrasi, kelembagaan, spbe, sdm, aparatur, lingkungan, kementerian, lhk	siang_tadi, tadi_menteri, menteri_panrb, panrb_abdullah, abdullah_azwar, azwar_anas, anas_menerima, menerima_audiensi, audiensi_menteri, menteri_klhk, klhk_pertemuan, pertemuan_membahas, membahas_terkait, terkait_reformasi, reformasi_birokrasi, birokrasi_kelembagaan, kelembagaan_spbe, spbe_sdm, sdm_aparatur, aparatur_lingkungan, lingkungan_kementerian, kementerian_lhk	siang_tadi_menteri, tadi_menteri_panrb, menteri_panrb_abdullah, panrb_abdullah_azwar, abdullah_azwar_anas, azwar_anas_menerima, anas_menerima_audiensi, menerima_audiensi_menteri, audiensi_menteri_klhk, menteri_klhk_pertemuan, klhk_pertemuan_membahas, pertemuan_membahas_terkait, membahas_terkait_reformasi, terkait_reformasi_birokrasi, reformasi_birokrasi_kelembagaan, birokrasi_kelembagaan_spbe, kelembagaan_spbe_sdm, spbe_sdm_aparatur, sdm_aparatur_lingkungan

3.1.3 Data Mining

Following the text preprocessing phase, topic modeling was conducted utilizing the Latent Dirichlet Allocation (LDA) algorithm. As illustrated in Figure 2, the modeling workflow was executed using several operators, specifically Read CSV for data import, Select Attributes for feature selection, and Extract Topics from Data for the LDA execution. To configure the LDA model, specific hyperparameters were established. The number of topics (K) was set to 10, and the optimization process was run for 100 iterations. These iterations allowed the LDA algorithm to continuously update and stabilize the word-topic distributions until an optimal and representative topic structure was achieved. Furthermore, the number of top words per topic was limited to 5, providing a concise lexical foundation to interpret the underlying theme of each cluster. Finally, the built-in Indonesian stopword dictionary provided by the software was applied directly within the operator configuration to automatically filter out any remaining irrelevant terms during the extraction process.

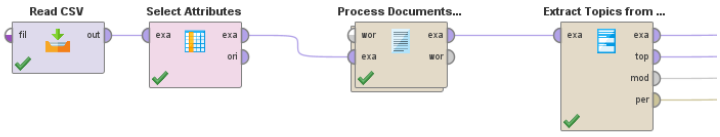


Figure 2 Data Mining Process

The generated topic models were evaluated using coherence scores to systematically determine the optimal number of topics (K). This metric was utilized to identify the specific number of topics that not only yielded the highest quantitative score but also maintained logical and semantic interpretability. Based on the parameter configurations and the coherence score calculations across different modeling iterations, the evaluation results are presented in Table 7:

Table 7 Experimental Comparison of LDA Model Parameters

Scenario	Parameter Input Values			Coherence Score	Num Topics
	Iteration	Alpha	Beta		
1	10	0,1	0,1	-2,115	3
2	10	0,1	0,001	-5,851	5
3	10	0,001	0,1	-1,878	3
4	10	0,001	0,001	-2,769	5
5	50	0,1	0,1	-2,562	3
6	50	0,1	0,001	-3,132	5
7	50	0,001	0,1	-1,422	3
8	50	0,001	0,001	-3,798	5
9	100	0,1	0,1	-2,581	3
10	100	0,1	0,001	-3,165	5
11	100	0,001	0,1	-1,496	3
12	100	0,001	0,001	-1,234	5

As presented in Table 6, the comparative analysis of various LDA topic modeling scenarios highlight the optimal configuration derived from the highest coherence score. Scenario 12—configured with 100 iterations, an alpha value of 0.001, and a beta value of 0.001—yielded the highest coherence score of -1.234. This configuration outperformed all other tested scenarios, indicating an optimal extraction of 5 distinct topics. Consequently, the parameter settings from Scenario 12 were adopted as the baseline for the final LDA implementation. Furthermore, the LDA execution successfully extracted the most dominant terms across the 3,327 dataset samples by calculating the aggregate weight of each term. The top five terms with the highest prevalence were aplikasi (145 occurrences), spbe (95 occurrences), setuju (51 occurrences), sistem (51 occurrences), and elektronik (47 occurrences). A detailed weight distribution of the top 10 terms is illustrated in Table 8

Table 8 Distribution of the Top 10 Terms

topic.id	word	weight
1	aplikasi	145
9	spbe	95
5	setuju	51
9	sistem	51
9	elektronik	47
9	pemerintahan	46
9	berbasis	44
5	guru	36
9	evaluasi	31
9	penilaian	30

To enhance visual clarity and provide a more intuitive representation of the data, the most frequently occurring terms from the aforementioned table were also visualized using a word cloud, as illustrated in Figure 3.



Figure 1 Wordcloud Visualization

3.2 Discussion

3.2.1 Results

The analytical results demonstrate that the quality of topic formulation within the LDA model is highly contingent upon the tuning of hyperparameters governing the document-topic (Alpha) and topic-word (Beta) distributions. Based on the Coherence Score evaluation across 12 modeling scenarios, Scenario 12 emerged as the most optimal configuration, yielding the highest coherence score of -1.234. This peak performance was

attained using 5 topics (K = 5) and 100 iterations—which facilitated optimal data convergence—alongside Alpha and Beta values of 0.001. These specific hyperparameter values indicate that the documents and topics are composed of highly specific and densely concentrated terms. Furthermore, a deeper analysis of the semantic weights in this scenario (highlighting terms such as pemerintahan [government], guru [teacher], evaluasi [evaluation], and penilaian [assessment]) proves that the LDA model can coherently map public sentiment. It reveals that societal attention extends beyond the mere technical manifestation of the applications, actively scrutinizing the system's evaluation processes and its tangible impacts on specific public sectors, notably educators.

CONCLUSION

Based on these findings, text mining has proven to be an effective solution for extracting valuable insights from the vast amount of information and discourse embedded within large-scale document collections. Through the application of topic modeling, this study successfully identified the prevalent themes discussed by the Indonesian public on social media regarding the Electronic-Based Government System (SPBE). These results can serve as a prioritized framework for formulating future government programs and policies. Consequently, policymakers can strategically identify specific areas requiring enhancement and mitigate program redundancies, thereby ensuring that future digital initiatives are more comprehensive, diversified, and impactful.

REFERENCES

[1] Arief, Assaf, and M. Yunus Abbas. "Kajian Literatur (Systematic Literature Review): Kendala Penerapan Sistem Pemerintahan Berbasis Elektronik (SPBE)." PROtek: Jurnal Ilmiah Teknik Elektro 8.1 (2021): 1-6.

[2] Pratiwi, Hemy Dwi, Widhy Hayuhardhika Nugraha Putra, and Admaja Dwi Herlambang. "Evaluasi Penerapan E-Government Di Pemerintah Kabupaten Sumbawa Besar Menggunakan Kerangka Kerja Sistem Pemerintahan Berbasis Elektronik (SPBE)." Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer 4.11 (2020): 4116-4124.

[3] Liedfray, Tongkotow, Fonny J. Waani, and Jouke J. Lasut. "Peran Media Sosial Dalam Mempererat Interaksi Antar Keluarga Di Desa Esandom Kecamatan Tombatu Timur Kabupaten Minahasa Tenggara." Jurnal Ilmiah Society 2.1 (2022).

[4] Andika, L. A., Azizah, P. A. N., & Respatiwulan, R. (2019). Analisis sentimen masyarakat terhadap hasil quick count pemilihan presiden indonesia 2019 pada media sosial twitter menggunakan metode naive bayes classifier. Indonesian Journal of Applied Statistics, 2(1), 34-41.

[5] Alfanzar, A. I., Khalid, K., & Rozas, I. S. (2020). Topic modelling skripsi menggunakan metode latent diriclhet allocation. JSiI (Jurnal Sistem Informasi), 7(1), 7-13.

[6] Nurdin, Khoirunnisa Fi, Taufik Edy Sutanto, and Ary Santoso. "Analisa Pemodelan Topik Berita Daring Menggunakan Semi-Supervised dan Fully Unsupervised Latent Dirichlet Allocation." Jurnal Pendidikan Tambusai 7.2 (2023): 8043-8055.