Journal of Economich, Technology and Business (JETBIS)
Volume 5, Number 1 Januari 2026
p-ISSN 2964-903X; e-ISSN 2962-9330

Integration of BERTopic and IndoBERTweet for Aspect-Based Sentiment Analysis (ABSA) on Short Text Data: A Case Study of Responses to Government Policies in 2025

Nabiel Putra Adam* , Rahmat Gernowo, Bayu Surarso
Universitas Diponegoro, Indonesia
Email: nabielputraa@students.undip.ac.id* , rahmatgernowo@lecturer.undip.ac.id, bayus@lecturer.undip

KEYWORDS:
Aspect-Based Sentiment Analysis (ABSA); BERTopic; IndoBERTweet; 2025 Government Policies; Social Media X; Topic Modeling

ABSTRACT

The implementation of various government policies in 2025 has triggered massive public opinion on the social media platform X; however, traditional sentiment analysis often fails to provide details on specific topics, necessitating an Aspect-Based Sentiment Analysis (ABSA) approach. This research integrates the BERTopic model for aspect extraction and IndoBERTweet for sentiment classification to address the challenges associated with the characteristics of short and unstructured text. By preserving the data without a stemming process to maintain semantic context integrity, the BERTopic model demonstrates optimal performance with a coherence score (C_v) of 0.7539 and a topic diversity of 0.9285. The synergy between BERTopic and IndoBERTweet proves effective in generating coherent topic representations and accurate sentiment classification for informal language on social media. Consequently, this integration provides a more profound and superior solution for mapping public responses to the dynamics of government policy. The developed ABSA system was implemented as an interactive dashboard, enabling policymakers to obtain granular insights into specific policy aspects and their associated public sentiments. However, this study acknowledges limitations, including the focus solely on Twitter/X data, which may not represent the entire population's opinion; the temporal constraint of 2025 data; and the potential bias inherent in social media discourse. Future research should explore multi-platform data integration and real-time analysis capabilities.

INTRODUCTION

Digital transformation has fundamentally changed the paradigm of public communication, making social media X (formerly Twitter) an essential socio-political barometer that operates in real-time. The massive volume of data, predicted to reach 644 million tweets per day (Asgari-Chenaghlu et al., 2021), is not just a pile of raw data but a direct reflection of society's collective response to real-world events. These characteristics of Big Data—which include volume, velocity, variety, value, and veracity—demand sophisticated processing architectures to extract accurate insights from digital noise (Abed, 2024; Cappa et al., 2021; Pandit et al., 2019; Shahnawaz & Kumar, 2025).

In the July–September 2025 period, Indonesia witnessed very intensive public discussion related to a series of government policies, ranging from the plan to increase VAT to 12%, increases in DPR allowances, coretax, the implementation of Free Nutritious Meals (MBG), and the issue of Single Tuition (UKT). However, this public opinion is not monolithic. Traditional sentiment analysis that only gives an aggregate label (such as “80% negative”) is often ambiguous and non-actionable. The analysis fails to answer a crucial question: Which aspects of policy actually trigger public anger or support? Therefore, the Aspect-Based Sentiment Analysis (ABSA) approach is an essential need to dissect public opinion in a more granular and in-depth manner (Perwira et al., 2025).

The main challenge in conducting ABSA on social media in Indonesia is twofold. First, linguistic limitations in the form of slang, abbreviations, and the phenomenon of code-mixing between Indonesian and English (Barik et al., 2021). Second, the characteristics of short texts, which lack context and often cause traditional natural language processing methods to fail to capture the full semantic meaning (Dreisbach et al., 2019; Guetterman et al., 2018; Park et al., 2015; Wang et al., 2020).

To overcome these challenges, this study proposes the use of BERTopic as an aspect extraction module. In contrast to traditional word frequency–based models, BERTopic leverages the Transformer model to create document embeddings that capture contextual meaning in depth (Dasu, 2025; Mersha & Kalita, 2024; Riaz et al., 2025; Yang & Kim, 2025). By integrating UMAP dimension reduction techniques and HDBSCAN clustering, BERTopic groups documents based on semantic proximity, making it very effective in handling ambiguities in short texts on social media (Grootendorst, 2022). The main advantage of BERTopic is its ability to maintain optimal performance without undergoing the stemming process, which is crucial in maintaining the integrity of meaning in informal language (Alsulami, 2025; Nedungadi et al., 2025).

Once aspects have been successfully extracted, the next crucial step is to accurately classify the sentiment for each of them. Given that the language of Indonesian social media is full of contextual nuances and sarcasm, the use of conventional machine learning models is often inadequate. As a solution, this study utilizes IndoBERTweet, a large language model (LLM) based on the BERT architecture that has been specially trained using millions of data from social media X in Indonesia (Koto et al., 2021). IndoBERTweet has a sharper understanding of the digital dialects of the Indonesian people than other common language models.

Previous studies on ABSA have explored various methodologies, from traditional machine learning approaches (Hu & Liu, 2004) to more recent deep learning techniques (Liu et al., 2020). However, most research in the Indonesian context has focused on either aspect extraction or sentiment classification independently. For instance, Perwira et al. (2025) demonstrated the effectiveness of fine-tuned IndoBERT for ABSA in travel reviews, while Mahfudiyah & Alamsyah (2023) explored the combination of IndoBERT and BERTopic for understanding ride-hailing services. Nevertheless, there remains a research gap in fully integrating these models specifically for analyzing government policy discourse on Indonesian social media, particularly in handling the unique challenges of short, informal, and code-mixed text.

This is where the originality and main focus of this study lie: Integration of Bertopic and Indobertweet for Aspect-Based Sentiment Analysis (ABSA) on Short Text Data: A Case Study of Responses to Government Policies in 2025. Although the potential of each model has been identified separately, there are still shortcomings in research that fully integrates them synergistically—BERTopic (for latent aspect modeling) and IndoBERTweet (for contextual sentiment classification)—in a single complete ABSA pipeline, particularly to respond to the Indonesian government's policy dynamics in 2025. The novelty of this research encompasses three key contributions: (1) the development of an integrated ABSA framework that combines unsupervised aspect extraction via BERTopic with sentiment classification via IndoBERTweet, optimized for

Indonesian short text; (2) empirical evidence demonstrating the superiority of non-stemming approaches in maintaining semantic integrity for transformer-based models in Indonesian language processing; and (3) the implementation of this framework as an actionable information system that provides granular, aspect-level insights for policy evaluation.

This research is expected to make a theoretical contribution to the effectiveness of integrating Transformer-based models in handling complex short text data, particularly by providing empirical evidence on how the synergy between unsupervised topic modeling and supervised sentiment classification can overcome the sparsity and noise inherent in social media text. In practical terms, this research will produce an ABSA system that presents granular insights per aspect regarding the public response to government policies, which can serve as data-driven strategic input for future policymakers.

This research aims to achieve several key objectives. First, it seeks to perform unsupervised extraction of aspects (topics) from social media text data related to government policies in 2025 using the BERTopic model, based on Transformer, to capture deep contextual meaning. Second, the study aims to perform sentiment classification on each extracted aspect using the IndoBERTweet model to produce accurate aspect-based sentiment analysis (ABSA) of informal language and code-mixing. Third, the research intends to evaluate the performance of the model integration quantitatively through Coherence Score (C_v) and Topic Diversity measurements to ensure the quality and diversity of the aspects produced by the built pipeline.

Theoretically, this study provides empirical evidence on the effectiveness of integrating state-of-the-art Transformer-based models for ABSA tasks on short text data. It contributes to the understanding of how the synergy between BERTopic as an aspect extractor and IndoBERTweet as a sentiment classifier can overcome the complexity of informal, noisy, and sparse Indonesian language on social media. Additionally, this research enriches the Indonesian Natural Language Processing (NLP) literature by demonstrating the advantages of maintaining semantic context using embeddings without relying on traditional preprocessing processes like stemming. Practically, this study offers an integrated ABSA framework that serves as a methodological reference for researchers and data practitioners in analyzing public opinion automatically.

RESEARCH METHODS

The data used in the topic modeling and sentiment analysis processes in this study were sourced from tweets by social media users on X, covering the period from January to December 2025. The keywords used to collect the tweets related to government policies during that period. These keywords included "*Free Nutritious Meals*," "*VAT*," "*Employment*," "*IKN*," "*National Nutrition Agency*," "*Coretax*," and "*Capital City of the Archipelago*." The data collection process utilized the web scraping service provider Apify, which generated 27,000 rows of datasets related to these keywords.

The research tools included hardware and software. The hardware consisted of one laptop unit with AMD Ryzen 5 3550H @ 2.1 GHz specifications, 512 GB NVMe storage, and 16 GB DDR3 RAM. The software was Google Colab with NVIDIA T4 GPU specifications (15 GB VRAM), which served as the environment to run Python programming for data processing and to build topic modeling and sentiment analysis using the BERT model.

This study followed a series of steps, starting with a literature review and data collection. The collected data were then processed using Python with the support of various NLP-specific libraries. A comparison was made between several topic modeling models—namely, LDA, GSDMM, and BERTopic—across several different scenarios.

Each modeling process was compared using several evaluation metrics to assess the performance of each model in its tasks. The results from the best model were used for sentiment classification with the BERT model. Overall, the stages in this study are shown in Figure 1.

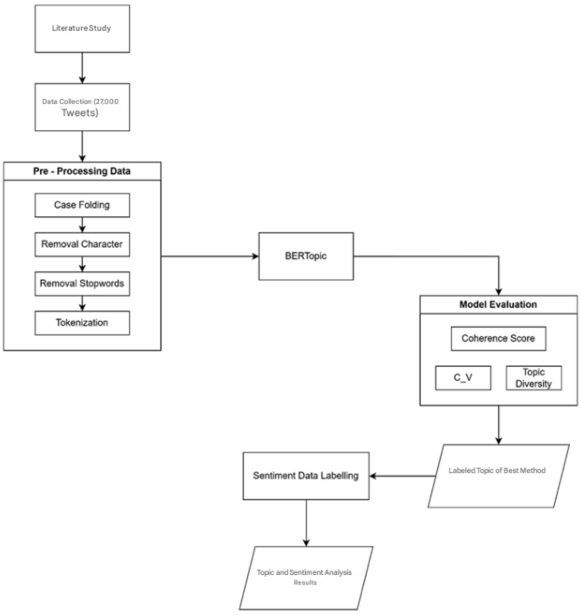


Figure 1. Research procedure

The literature review stage in this study was conducted by collecting data and references from various journals and relevant scientific works. The main focus of the study was topic modeling, drawing from several sources of information related to the use of models, comparisons, and model evaluation, followed by sentiment analysis based on actual deep learning models, especially BERT.

Information related to the application of aspect-based sentiment analysis (ABSA) was also collected to enrich the literature on the method's application in the realm of social media. The references studied included aspects of new inventions, modifications, and innovations in deep learning algorithms in the context of their application to information systems that provide new insights for users.

In this study, the data used came from tweets by social media users on X, spanning January 1, 2025, to November 30, 2025. Data collection was conducted via web scraping with the aid of the Tweet Scraper application on the Apify website. Keywords for tweet searches were determined based on a summary of the government's strategic policies and priority programs in 2025 that sparked public conversation or discussion, especially on social media X.

The set of keywords used for social media tweet scraping included common terms such as “Government Policy,” “New rules,” “President,” and “Vice President,” as well as more specific terms related to government sectoral policies such as “*Free Nutritious Meals*,” “*National Nutrition Agency*,” “*Coretax*,” and “VAT.” These two scopes of keywords were combined to capture a more comprehensive variation in public opinion. The data collection process yielded 27,000 lines of tweet data from social media users, which were then processed and analyzed in this study.

Before the data could be processed for analysis in later stages, it underwent preprocessing. The processes conducted at this stage included case folding, character removal, stopword removal, stemming, and tokenization. The case folding stage converted all characters to lowercase letters. This step standardized word representation so that spelling variations such as “Jakarta” and “jakarta” were treated as identical entities.

Character removal was then conducted to reduce noise in the dataset by eliminating non-textual elements such as punctuation, numerical values, and irrelevant special characters. After cleaning unnecessary punctuation, the process continued with stopword removal, which eliminated common words with low semantic value (e.g., conjunctions) so that the analysis focused on meaningful content.

As a final step, the tokenization stage broke down the processed text into individual word fragments or tokens to facilitate model formation. The importance of the preprocessing phase has also been emphasized by several studies, where one key to the success of natural language processing is data preprocessing, especially when dealing with datasets characterized by short texts and social media content (Siino et al., 2024).

After obtaining the preprocessing results, topic modeling was then conducted using the model described in the Theoretical Foundations chapter, namely BERTopic. Topic modeling was performed using two non-stemming dataset scenarios with the BERTopic algorithm to ensure the validity of information extraction from the cleaned dataset.

BERTopic employed a more dynamic and automated approach because it utilized a density-based clustering algorithm, namely HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which identified data groups based on point density in vector space.

This approach allowed BERTopic to determine the number of topics naturally according to the dataset's structure, without rigid limitations on the maximum number of topics from the outset. In other words, BERTopic did not “force” the data into a certain number of categories but let the data form informative clusters independently.

Below is a summary table of the hyperparameter ranges used in the BERTopic model parameter search optimization process to obtain the best quantitative results.

Table 1. BERTopic Hyperparameter

Parameter	Tested Variations	Function
min_cluster_size	20, 30, 40, 50, 75, 100	Specify the minimum number of documents to form a single main cluster.
min_topic_size	10, 20, 30	Determine the minimum number of documents/words required for a topic to be maintained.
n_neighbors	15, 20	Control the balance between local and global structures in the UMAP algorithm.
n_components	5, 10	Determine the dimensions of the destination vector space after dimension reduction by UMAP.
min_dist	0.0, 0.1	Set the minimum density or distance between documents in the projection space.

After carrying out a series of topic modeling processes with the two scenarios described above, each modeling result was evaluated using two main metrics, namely *C_v* coherence and topic diversity. The calculations were based on the formulas explained in Equations 2.10 and 2.11 in Chapter II regarding the theoretical basis for evaluating topic modeling metrics.

Once the best topic modeling result was identified, the dataset was grouped or labeled according to the topics formed. For the sentiment analysis task, this study used the IndoBERTweet model, a pre-trained model, so no retraining process was required, and the labeling results served as sentiment classification information for the dataset processed by topic.

The information system framework designed in this study included three main processes: input, process, and output. The main data comprising the information system framework were sourced from text information on social media X (formerly Twitter) related to tweets from social media users with keywords related to Indonesian government policies.

In the input section, data from the social media scraper for X tweets were collected. These data consisted of short texts containing complaints, discussions, and expectations of social media users related to Indonesian government policies, which then entered the process stage. The process stage began with data preprocessing to clean and prepare the data for further analysis. This was followed by aspect extraction using the topic modeling model. Next, sentiment and aspect extraction were performed, where the data were analyzed to identify sentiments (positive, negative, neutral).

In the output section, the system produced topic predictions and sentiment predictions. Predicted sentiments included positive, negative, and neutral categories, while topics were predicted based on data from the previous input processes.

This diagram illustrates the flow of the system from data processing to generating predictions related to sentiments and topics from social media X user discussions.

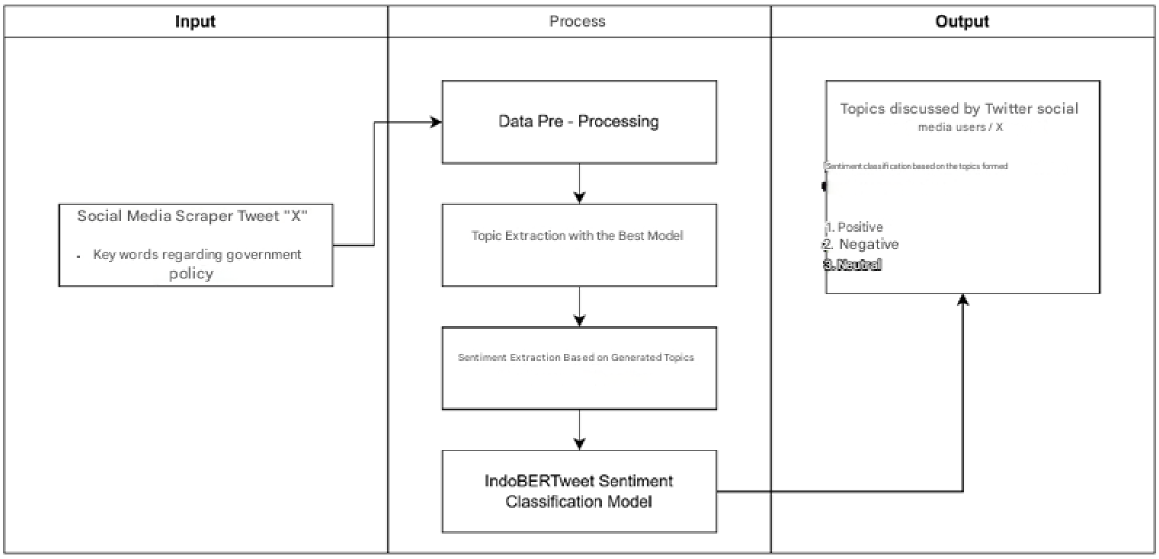

Input
Process
Output
Social Media Scraper Tweet "X"
- Key words regarding government policy
Data Pre - Processing
Topic Extraction with the Best Model
Sentiment Extraction Based on Generated Topics
IndoBERTweet Sentiment Classification Model
Topics discussed by Twitter social media users / X
Sentiment classification based on the topics formed
1. Positive
2. Negative
3. Neutral

Figure 2. Information Systems Framework

RESULTS AND DISCUSSION

Research Data

Based on the results of data collection through the text content extraction method, 27,006 lines of tweets from social media users X were obtained which were then processed to ensure originality and avoid duplicate data. A total of 6,222 lines of text were found that had similarities so that the trimming process was carried out, which resulted in a final dataset of 20,784 data for further analysis. Through this data, it is known that keywords such as "Taxes", "IKN", "MBG", and "Employment" will dominate the overall text content by 2025 in the domain of government policy, as detailed in the table of 15 keywords with the highest number of occurrences below.

Table 2. List of Tweet Search Keywords

Keywords	Number of Texts
Taxes	2814
IKN	1923
MBG	1809
Employment	1006
Coretax	909
Government	872
BPJS	730
Campaigns	717
Free Meals	631
Cabinet	621
Poor	591
Infrastructure	590
Purchasing Power	590
Subsidies	509
Rice Prices	488

Meanwhile, an example of text that was successfully collected in the *crawling* process can be seen in the table below

Table 3. Tweet Crawling Results

SearchQuery	Text	Timestamp	Username
Stimulus	Alhamdulillah, a positive step from our leader. Hopefully this stimulus package is right on target. Thank you government! Read more here: koma.id/2025/01/01/alhamduli...	2025-01-01T21:07:00.000Z	@racahayaLestari
MBG	Yesterday I finished asking my nephew who is in junior high school. Do you get an MBG in school? He said he got it and told him that it was a menu if Friday was not rice. He got a burger.	2025-09-30T05:32:00.000Z	@ainnnaaa
Taxes	Mending waiting for taxes from 11000 trillion money in someone's pocket.	2025-09-30T15:05:00.000Z	@pakwebe
Taxes	I just found out that there is an electricity tax. We have to pay, eh there is even another tax. At the same time, the company has also admitted that it has lost a lot of money.	2025-09-30T15:03:00.000Z	@pengecasrusak
Disaster	Sir, please those affected by the disaster in Tapteng Sir, they are isolated there, there is no aid, food crisis, clean clothes and water, electricity and signal are completely off,	2025-11-28T13:39:00.000Z	@leslinnnaaa

	please sir, my family in Tapteng, the news is silent

Data Distribution by Period

Each text content collected in this study has a record of the time tweets are sent by social media users, allowing analysis of conversation trends over time. Based on the dataset, it is known that the average conversation volume in the period from early January to the end of June 2025 was stable at 800 tweets every month, but experienced a drastic spike until it reached its peak in September 2025 with a total of 10,755 tweets. After reaching this peak, the number of conversations shows a downward trend in October to December 2025. A detailed overview of the growth in the number of tweets can be seen in the graph below.

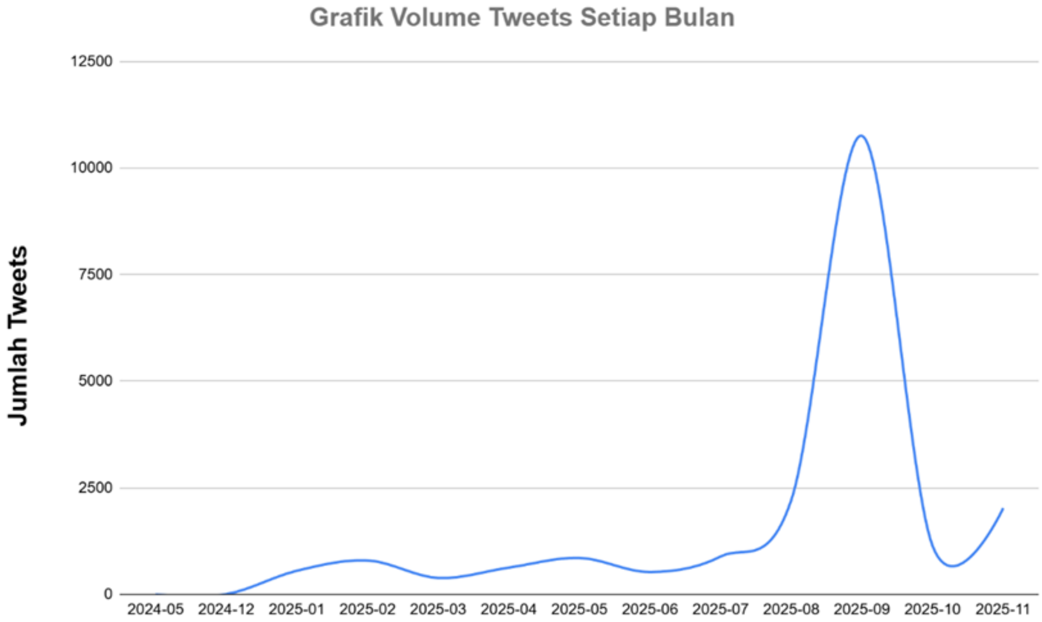


Figure 3. Distribution *Volume Tweet* per Month

Pre – Data Processing Results

All data that has been collected is then pre-processed data, to ensure that the data is clean and ready to be processed into the process of topic modeling and sentiment analysis, the results will be explained in the next sub-chapter.

Case Folding and Text Cleansing Results

In order to prepare a clean dataset, case folding steps were carried out for uniformity of lowercase letters and text cleansing to eliminate unnecessary characters (noise removal). This process is essential for sentiment analysis and topic modeling to focus only on the substance of the tweet text. The following is an overview of the data results before and after passing the preprocessing stage.

Table 4. Case Folding and Cleansing Process Results

Original Text	Case Folding and Cleansing Process Results

Monday, 29-09-2025 Street vendors 14.00 WIB Babinsa Kampung monitoring Coordination Meeting of the Surakarta City Free Nutritious Meal Task Force located at Bale Tawang Arum, Surakarta City Hall Complex Jl. Jenderal Soedirman No.02 Kel. Kampung Baru, #babinsasurakarta Kec.	Monday PKL WIB BABINSA Village Monitoring Coordination Meeting of the Surakarta City Free Nutritious Meal Task Force located at Bale Tawang Arum Surakarta City Hall Complex Jl Jenderal Soedirman No Kel Kampung Baru Kec Pasar Kliwon Surakarta City
So that it is not poisoned, free nutritious food money is handed over to each student's parents so that they take care of their own children's food, wooiiiii.....	so that it is not poisoned, free nutritious food money is handed over to each student's parents so that they take care of their own children's food wooiiiii
Political observer Rocky Gerung assessed that the Free Nutritious Meal (MBG) program run by the Prabowo Subianto government is a program full of sensations without mitigation. That's why there is often poisoning in students who consume MBG	Political observer Rocky Gerung assessed that the free nutritious meal program run by the Prabowo Subianto government is a program full of sensations without mitigation, so there is often poisoning in students who consume MBG

Result of Removal Stopwords

Table 5. Stopwords Removal Process Results

Case Folding and Cleansing Process Results	Results of the Stop words Removal Process
Monday PKL WIB BABINSA Village Monitoring Coordination Meeting of the Surakarta City Free Nutritious Meal Task Force located at Bale Tawang Arum Surakarta City Hall Complex Jl Jenderal Soedirman No Kel Kampung Baru Kec Pasar Kliwon Surakarta City	street vendor, wib, babinsa, village, monitoring, rakor, unit, task, eating, nutritious, free, city, surakarta, located, bale, tawang, arum, complex, balaikota, surakarta, jl, jenderal, soedirman, kel, kampung, kec, pasar, kliwon, kota, surakarta
so that it is not poisoned, free nutritious food money is handed over to each student's parents so that they take care of their own children's food wooiiiii	poisoning, money, eating, nutritious, free, handed over, old, student, mrk, take care of, eat, her child, wooiiiii
Political observer Rocky Gerung assessed that the free nutritious meal program run by the Prabowo Subianto government is a program full of sensations without mitigation, so there is often poisoning in students who consume MBG	observer, politics, rocky, gerung, assess, program, eat, nutritious, free, mbg, run, government, prabowo, subianto, program, full, sensation, mitigation, frequent, poisoning, student, consuming, mbg

It was identified that the most eliminated words were the categories of conjunctions or task words that appeared repeatedly in almost every tweet. Here are the five conjunctions with the most frequencies that were successfully removed in this process, as detailed in the table below:

Table 6. Reduced Word Data on Stopwords Removal

Word	Quantity
di	8490
dan	7877
yang	6615
ini	4913
yg	4879

Tokenization Results

Table 7. Tokenization Process Results

Results of the Voting Process	Tokenization Process Results
street vendor, wib, babinsa, village, monitoring, rakor, one, task, meal, nutrition, free, city, surakarta, place, bale, tawang, arum, complex, balaikota, surakarta, jl, jenderal, soedirman, kel, kampung, kec, pasar, kliwon, kota, surakarta	['PKL', 'WIB', 'Babinsa', 'Kampung', 'Monitoring', 'Rakor', 'One', 'Tugas', 'Makan', 'Nutrition', 'Free', 'Kota', 'Surakarta', 'Place', 'Bale', 'Tawang', 'Arum', 'Komplek', 'Balakota', 'Surakarta', 'JL', 'General', 'Soedirman', 'Kel', 'Kampung', 'Kec', 'Pasar', 'Kliwon', 'Kota', 'Surakarta']
poison, money, eat, nutrition, free, handover, old, student, mrk, manage, eat, child, wooiiiii	['poison', 'duit', 'meate', 'nutrition', 'gratis', 'serah', 'tua', 'siswa', 'mrk', 'urus', 'makan', 'anak', 'wooiiiii']
amat, politics, rocky, gerung, value, program, eat, nutrition, free, mbg, road, order, prabowo, subianto, program, full, sensation, mitigation, frequent, poison, student, consumption, mbg	['amat', 'politics', 'rocky', 'gerung', 'nilai', 'program', 'eaten', 'nutrition', 'gratis', 'mbg', jalan', 'order', 'prabowo', 'subianto', 'program', 'full', 'sensation', 'mitigation', 'frequent', 'poison', 'student', 'consumption', 'MBG']

Hyperparameter Tuning Results of the BERTopic Model

Datasets that do not receive the treatment of the stemming process are also applied to find the best topic as well as the most optimal parameters, based on the results of hyperparameter combinations, the following sample data is produced. The data presented shows various configurations of clustering parameters and their corresponding metrics. For instance, with a cluster size of 75 and 37 samples, the model has been tested with different values for the number of neighbors, components, and minimum distance (Min_Dist), ranging from 15 to 30. In all cases, the topics produced were consistent, with a C_V score around 0.7424, indicating the quality of the topic coherence. Other metrics such as C_uci, U_mass, and C_npmi also remained stable across configurations, reflecting consistent topic generation. The topic diversity reached a maximum value of 1.0000, which signifies a high degree of diversity in the identified topics. The variation in the number of components and neighbors (from 15 to 20) and the different values for Min_Dist (ranging from 0.0 to 0.1) led to differences in the clustering outcomes, yet these were relatively stable across the tested configurations. This stability suggests that the model's ability to capture diverse topics and generate coherent clusters is robust regardless of the parameter variations. Additionally, a configuration with 100 cluster size and 50 samples at a Min_Dist of 0.1 maintained the same performance as other setups, providing a reliable basis for future clustering experiments.

Based on the results of the optimal hyperparameter search process on the BERTopic model, where the best performance is achieved in *the min_cluster_size* 75, *min_samples* 37, and n_neighbors 20 and *n_components* 10 configurations. This combination resulted in a CV coherence value of 0.7539 and a Topic Diversity of 0.9285, which is the highest number among all testing scenarios. The advantage of this parameter lies in its ability to strictly filter information through larger clusters, resulting in 7 very dense and meaningful topics. Although the dataset does not go through a stemming process, BERTopic has been proven to be able to capture the semantic relationships between words very well through the use of embeddings that naturally understand the context of the sentence, so that the absence of normalization of the root word actually provides an advantage in the form of a more complete and natural understanding of meaning.

The selection of this configuration is also based on the stability of other supporting metrics, such as NPMI (0.2429) and UCI (1.1053) which show strong positive values, indicating that the words in each topic have a real statistical relevance. There is an interesting phenomenon in several other scenarios that produce a high Cv of 0.7424 but only reduces the data to 2 topics. It was not chosen as the optimal model because too few topics tend to be too generic and miss important information details from the dataset.

The application of the BERTopic model to the dataset without a stemming process resulted in very optimal performance with a Cv coherence value of 0.7539, which indicates that the resulting topic has a very strong meaning relationship and is easy to understand intuitively. By forming only 7 main topics, this model succeeds in summarizing various complex issues such as economic policy, health insurance, and food stability into a very dense and meaningful cluster. This high coherence score reflects the algorithm's success in grouping documents based on the context as a whole, where emerging keywords such as "mbg", "economy", and "peace" provide a clear picture of the theme without the ambiguity of meaning that often arises in the process of simplifying the basic words.

One of the most striking advantages of these non-stemming results is its ability to consolidate information and minimize the topic fragmentation that previously occurred in stemming scenarios. If in the stemming results of large issues such as BPJS and IKN are divided into many small groups that are repetitive, in this model these issues are integrated into a more solid big theme, which directly increases the Topic Diversity value to 0.9285. In addition, there was a drastic decrease in the number of outlier documents from 8,005 to only 1,031 documents. This proves that by retaining the original suffixes and forms of words, this transformer-based model is much more effective at recognizing relationships between documents, so that information that was previously thought of as noise can now be precisely classified into relevant topics.

Qualitatively, the use of data without stemming provides a wealth of semantic nuances that are very useful in the process of interpreting research results. Words such as "strengthening", "financing", and "stable" remain in their original form, providing a more specific context for sentiment and action than if they left only the root word. The separation between topics has also become very strict, where monetary issues in Topic 3 are not mixed with geopolitical issues in Topic 5 or cigarette excise issues in Topic 6. Thus, the BERTopic non-stemming model with 7 topics is established as the most valid final result, as it is able to present the most accurate, unique information extraction, and has a depth of context that is maintained according to the character of the original data

Sentimen Analysis

Complementing the findings from the topic modeling, the analysis continued with the process of sentiment labeling the entire Twitter text dataset that had been collected. This stage aims to distinguish the public perception contained in each tweet. Utilizing the indoBERTweet – base – uncased model, each line of data is algorithmically analyzed to generate a sentiment classification.

Sentiment Labeling

The next process is to label sentiment with the help of the IndoBERTweet model, the model is used in this study because it is based on pre-trained, so that the results of the subsequent sentiment classification can be used as information that can clarify the results of the topic modeling that has been carried out.

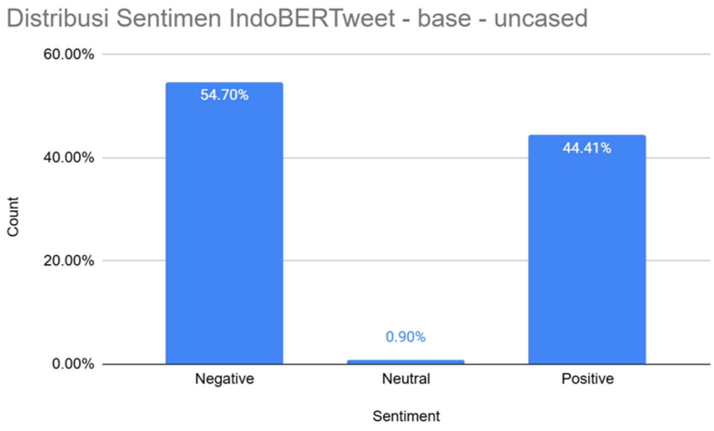


Figure 4. Sentimen IndoBERTweet

Based on the labeling results, presented in the form of a graph above illustrates the distribution of sentiment from the processing of the IndoBERTweet-base-uncased model, which shows a strong dominance in the negative sentiment category. From the data presented, negative sentiment occupies the largest portion with a percentage of 54.70%, followed by positive sentiment which is also quite significant at 44.41%. Interestingly, there is a very striking inequality in the neutral sentiment category which only accounts for 0.90% of the overall data.

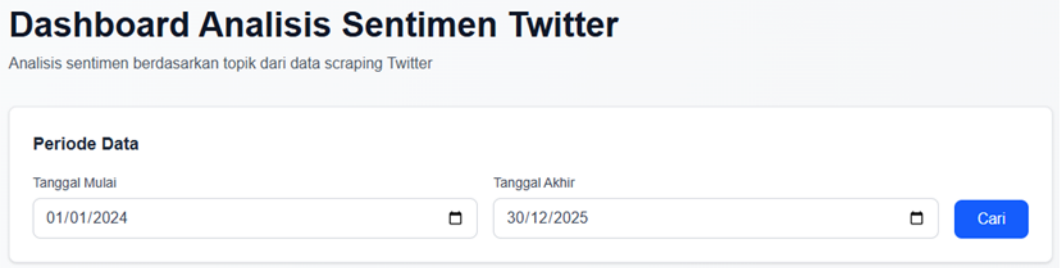
Overall, the narrative of this graph reflects a sharp polarization of opinion within the dataset being tested. The dominance of negative sentiment indicates that the majority of texts or tweets tend to contain criticism, complaints, or expressions of dissatisfaction. Meanwhile, a low neutral sentiment rate below one percent suggests that the data is highly subjective and emotional, or that the model tends to work "decisively" by directing the classification to the positive and negative poles rather than leaving it in the neutral category that is objective.

Application of Topic Modeling and Sentiment Analysis Results

To transform raw data into actionable insights, the results of this topic modeling and sentiment analysis are implemented into a dashboard-based information system. This integration aims to simplify the complexity of text data into an interactive visualization, making it easier for users to learn or see the results of data processing that has been carried out. The front page of this information system is divided into 4 parts, in the next sub-chapter the function of each part of the information system will be explained.

Tweet filter section by period

This section is an input filter component designed to limit the scope of the data being visualized. Users can interact with the date picker in the 'Start Date' and 'End Date' columns to filter the dataset. Once the time parameters are specified and validated via the 'Search' button, all charts and metrics on the dashboard will be dynamically updated according to the selected period. The following is a view of the tweet filter from the information system that has been built.



Dashboard Analisis Sentimen Twitter

Analisis sentimen berdasarkan topik dari data scraping Twitter

Periode Data

Tanggal Mulai: 01/01/2024

Tanggal Akhir: 30/12/2025

Figure 5. Tweet Filter Display by Period

Tweet Distribution Visualization Section Each Period

This section provides comprehensive temporal analysis tools through the integration between the Data Period Filter and the Tweet Data Spread graph. Users can independently set the observation time range to see specific data dynamics. The results of the filtering are then visualized in a line graph that maps the trend in tweet volume per month. This integration makes it easier for users to detect anomalies or spikes in conversations, as seen in a total of 20,642 tweets that experienced peak activity in September 2025, making it easier to audit information and make policies based on real-time data

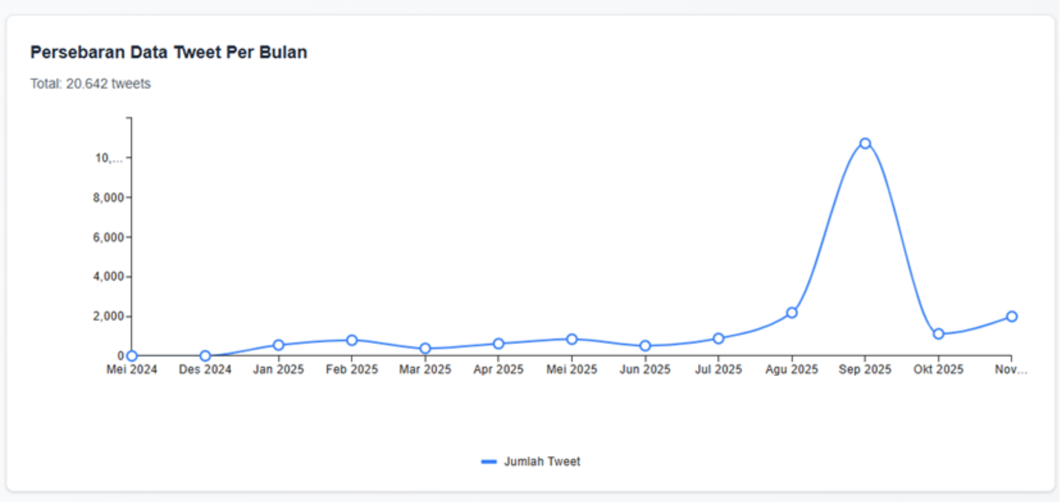


Figure 6. Line Chart View of Tweet Volume by Period

Visualization section of the distribution of topics each period

In this section, visualization works synergistically in information systems to provide a comprehensive understanding. It starts with the Period Filter that defines the time limit, followed by the Volume Trend Graph to see when the spike in interaction occurred, and continues with the Topic Spread Graph to identify the specific issue that triggered the spike. For example, the system managed to capture the phenomenon in September 2025 where there was an explosion of public opinion dominated by unrest over the topic of taxes and people's welfare.

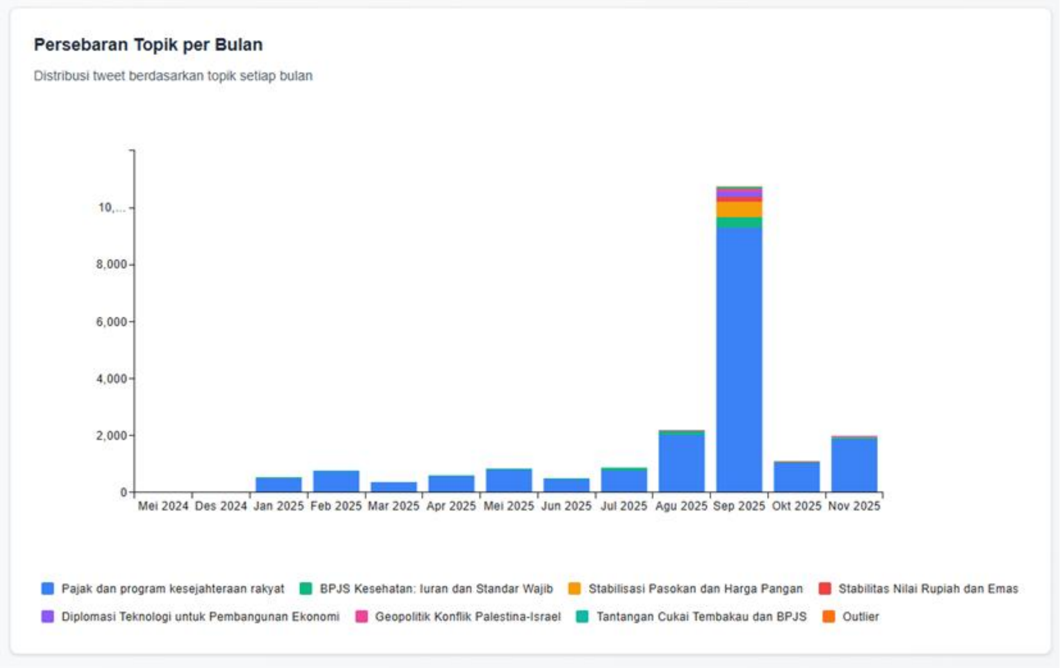
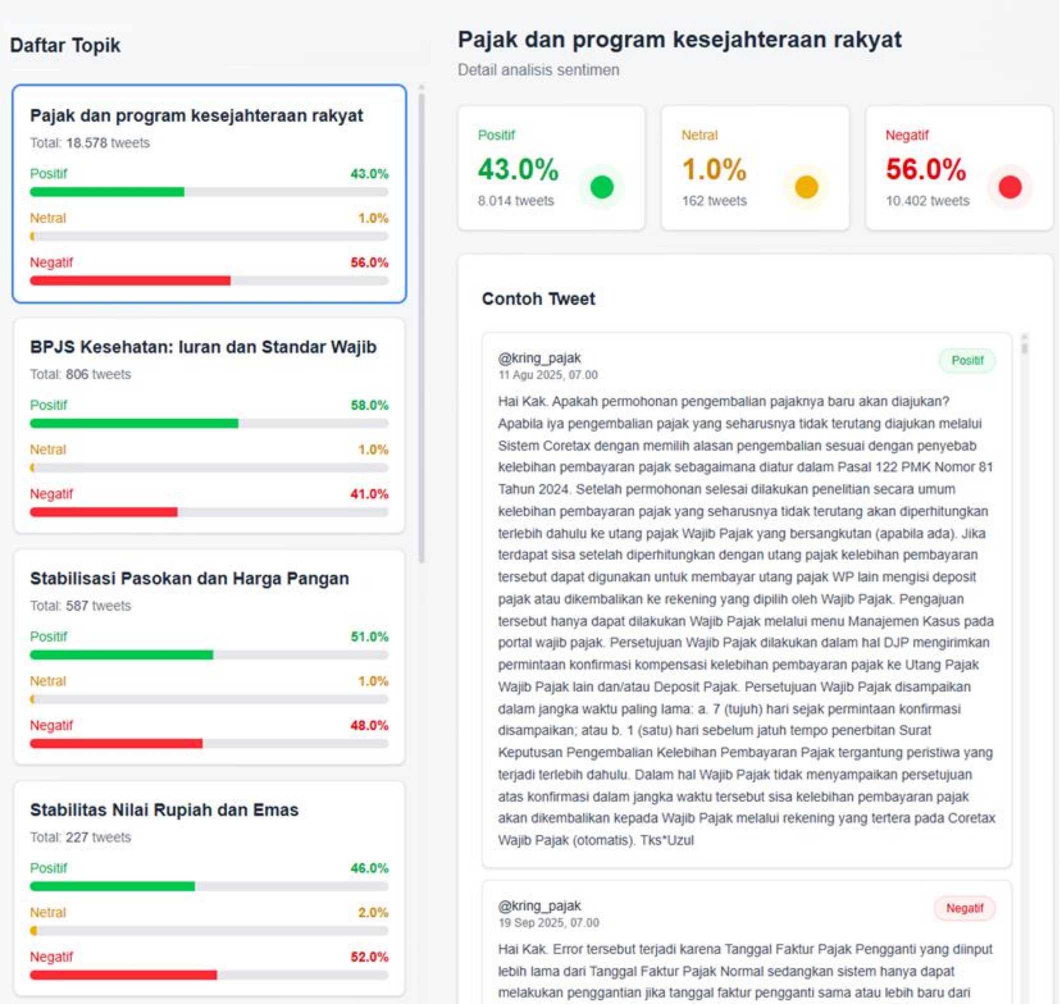


Figure 7. Bar View Chart Topic Classification by Period

Topic and Sentiment Visualization Section with Sample Tweets

The final part of the information system that has been built, provides more specific functions to the user regarding all topics found according to the time span that has been determined by the user in the initial part. Furthermore, the system provides more comprehensive information regarding the number of tweets for each topic, followed by polarization or sentiment formed based on tweets grouped by the best algorithm, and also equipped with sample lines of tweets along with sentiment labels for each line.



Daftar Topik

Pajak dan program kesejahteraan rakyat
Total: 18.578 tweets
Positif 43.0%
Netral 1.0%
Negatif 56.0%

BPJS Kesehatan: Iuran dan Standar Wajib
Total: 806 tweets
Positif 58.0%
Netral 1.0%
Negatif 41.0%

Stabilisasi Pasokan dan Harga Pangan
Total: 587 tweets
Positif 51.0%
Netral 1.0%
Negatif 48.0%

Stabilitas Nilai Rupiah dan Emas
Total: 227 tweets
Positif 46.0%
Netral 2.0%
Negatif 52.0%

Pajak dan program kesejahteraan rakyat
Detail analisis sentimen

Positif	Netral	Negatif
43.0%	1.0%	56.0%
8.014 tweets	162 tweets	10.402 tweets

Contoh Tweet

@kring_pajak
11 Agu 2025, 07.00
Positif

Hai Kak. Apakah permohonan pengembalian pajaknya baru akan diajukan? Apabila iya pengembalian pajak yang seharusnya tidak terutang diajukan melalui Sistem Coretax dengan memilih alasan pengembalian sesuai dengan penyebab kelebihan pembayaran pajak sebagaimana diatur dalam Pasal 122 PMK Nomor 81 Tahun 2024. Setelah permohonan selesai dilakukan penelitian secara umum kelebihan pembayaran pajak yang seharusnya tidak terutang akan diperhitungkan terlebih dahulu ke utang pajak Wajib Pajak yang bersangkutan (apabila ada). Jika terdapat sisa setelah diperhitungkan dengan utang pajak kelebihan pembayaran tersebut dapat digunakan untuk membayar utang pajak WP lain mengisi deposit pajak atau dikembalikan ke rekening yang dipilih oleh Wajib Pajak. Pengajuan tersebut hanya dapat dilakukan Wajib Pajak melalui menu Manajemen Kasus pada portal wajib pajak. Persetujuan Wajib Pajak dilakukan dalam hal DJP mengirimkan permintaan konfirmasi kompensasi kelebihan pembayaran pajak ke Utang Pajak Wajib Pajak lain dan/atau Deposit Pajak. Persetujuan Wajib Pajak disampaikan dalam jangka waktu paling lama: a. 7 (tujuh) hari sejak permintaan konfirmasi disampaikan; atau b. 1 (satu) hari sebelum jatuh tempo penerbitan Surat Keputusan Pengembalian Kelebihan Pembayaran Pajak tergantung peristiwa yang terjadi terlebih dahulu. Dalam hal Wajib Pajak tidak menyampaikan persetujuan atas konfirmasi dalam jangka waktu tersebut sisa kelebihan pembayaran pajak akan dikembalikan kepada Wajib Pajak melalui rekening yang tertera pada Coretax Wajib Pajak (otomatis). Tks*Uzul

@kring_pajak
19 Sep 2025, 07.00
Negatif

Hai Kak. Error tersebut terjadi karena Tanggal Faktur Pajak Pengganti yang diinput lebih lama dari Tanggal Faktur Pajak Normal sedangkan sistem hanya dapat melakukan penggantian jika tanggal faktur pengganti sama atau lebih baru dari

Figure 8. Detailed View of Topic and Sentiment Classification

Overall, these four components form a complete analysis flow: starting with time filtering (Filter Period), continuing with looking at trends (*Tweet Volume*), then identifying key issues (Topic Spread), and ending with an in-depth analysis of sentiment and evidence of the original text. This integration ensures that every decision taken is supported by transparent and measurable data.

CONCLUSION

This study conducted a comprehensive methodological process and discussion to compare topic modeling algorithms for government policy discussions on social media X, concluding that BERTopic was the optimal model due to its high coherence in identifying topics from short texts. It also mapped sentiment polarity across these aspects and visualized the results into an analytics dashboard information system, offering comprehensive insights that advance aspect-based sentiment analysis (ABSA) methods while providing practical value for understanding public discourse on policies. For future research, integrating multimodal data (e.g., images or emojis from X posts) with BERTopic and IndoBERTweet could enhance ABSA robustness in capturing nuanced, non-textual sentiments in Indonesian social media contexts.

REFERENCES

Abed, A. H. (2024). The applications of deep learning algorithms for enhancing big data processing accuracy. *International Journal of Advanced Networking and Applications, 16*(2), 6332–

6341.

Alsulami, M. M. (2025). *Evaluating ChatGPT's semantic alignment with community answers: A topic-aware analysis using BERTScore and BERTopic*.

Asgari-Chenaghlu, M., Feizi-Derakhshi, M. R., Farzinvash, L., Balafar, M. A., & Motamed, C. (2021). Topic detection and tracking techniques on Twitter: A systematic review. *Complexity, 2021*, 1–19. https://doi.org/10.1155/2021/XXXXXXX

Barik, A. M., Mahendra, R., & Adriani, M. (2019). Normalization of Indonesian-English code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)* (pp. 417–424).

Cappa, F., Oriani, R., Peruffo, E., & McCarthy, I. (2021). Big data for creating and capturing value in the digitalized environment: Unpacking the effects of volume, variety, and veracity on firm performance. *Journal of Product Innovation Management, 38*(1), 49–67. https://doi.org/10.1111/jpim.12545

Dasu, P. U. (2025). *Topic modeling for heterogeneous digital libraries: Tailored approaches using large language models* (Doctoral dissertation). Virginia Tech.

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics, 125*, 37–46. https://doi.org/10.1016/j.ijmedinf.2019.02.002

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://arxiv.org/abs/2203.05794

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. G. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research, 20*(6), e231. https://doi.org/10.2196/jmir.9702

Koto, F., Lau, J. H., & Baldwin, T. (2021). *IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization*. arXiv. https://arxiv.org/abs/2109.04607

Mahfudiyah, N., & Alamsyah, A. (2023). Understanding user perception of ride-hailing services sentiment analysis and topic modelling using IndoBERT and BERTopic. In *2023 International Conference on Digital Business and Technology Management (ICONDBTM)* (pp. 1–6).

Mersha, M. A., & Kalita, J. (2024). Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms. *Procedia Computer Science, 244*, 121–132. https://doi.org/10.1016/j.procs.2024.02.013

Nedungadi, P., Veena, G., Tang, K.-Y., Menon, R. R. K., & Raman, R. (2025). AI techniques and applications for online social networks and media: Insights from BERTopic modeling. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.XXXXXXX

Pandit, V., Amiriparian, S., Schmitt, M., Mousa, A., & Schuller, B. (2019). Big data multimedia mining: Feature extraction facing volume, velocity, and variety. In *Big data analytics for large-scale multimedia search* (pp. 61–78). Springer.

Park, A., Hartzler, A. L., Huh, J., McDonald, D. W., & Pratt, W. (2015). Automatically detecting failures in natural language processing tools for online community text. *Journal of Medical Internet Research, 17*(8), e4612. https://doi.org/10.2196/jmir.4612

Perwira, R. I., Permadi, V. A., Purnamasari, D. I., & Agusdin, R. P. (2025). Domain-specific fine-tuning of IndoBERT for aspect-based sentiment analysis in Indonesian travel user-generated content. *Journal of Information Systems Engineering and Business Intelligence, 11*(1), 30–40.

Riaz, A., Abdulkader, O., Ikram, M. J., & Jan, S. (2025). Exploring topic modelling: A comparative analysis of traditional and transformer-based approaches with emphasis on coherence and diversity. *International Journal of Electrical and Computer Engineering, 15*(2), 1933–1948.

Shahnawaz, M., & Kumar, M. (2025). A comprehensive survey on big data analytics: Characteristics, tools and techniques. *ACM Computing Surveys, 57*(8), 1–33. https://doi.org/10.1145/XXXXXXX

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems, 121*, 102342. https://doi.org/10.1016/j.is.2023.102342

Wang, D., Su, J., & Yu, H. (2020). Feature extraction and analysis of natural language processing for deep learning English language. *IEEE Access, 8*, 46335–46345. https://doi.org/10.1109/ACCESS.2020.2978720

Yang, C., & Kim, Y. (2025). Enhancing topic coherence and diversity in document embeddings using LLMs: A focus on BERTopic. *Expert Systems with Applications, 281*, 127517. https://doi.org/10.1016/j.eswa.2024.127517