Implementation of LSA for Topic Modeling on Tweets with the Keyword ‘Kemenkeu’

Shofiyatul Khariroh[1], Farrikh Al Zami[2]*, Heni Indrayani[3], Ika Novita Dewi[4], Aris Marjuni[5], Mira Riezky Adriani[6], Moh Hadi Subowo[7]

[1245]Information Systems, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[3]Communication Sciences, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[6]Kementerian Keuangan Republik Indonesia
[7]UIN Walisongo, Semarang, Indonesia
[1]112202106667@mhs.dinus.ac.id , [2]alzami@dsn.dinus.ac.id , [3]heni.indrayani@dsn.dinus.ac.id,
[4]ikadewi@dsn.dinus.ac.id, [5]aris.marjuni@dsn.dinus.ac.id, [6]mira.adriani@kemenkeu.go.id,
[7]hadi.subowo@walisongo.ac.id

Submitted :Dec 8, 2024 | **Accepted** : Dec 31, 2024 | **Published** : Jan 8, 2025

Abstract: This research explores public discourse on financial policies by analyzing tweets mentioning the keyword 'Kemenkeu' (Ministry of Finance). Using Latent Semantic Analysis (LSA), the study examined 10,099 tweets to uncover key topics that reflect public sentiment toward the Ministry’s policies. Preprocessing steps, such as stopword removal and stemming with Sastrawi, were essential to ensure the effectiveness of the analysis. The results revealed three main topics: Finance and Budget, Salaries and Employee Welfare, and Excise and Customs Regulations. These insights provide a better understanding of public opinion on financial issues and highlight the importance of proper text preprocessing in topic modeling. This approach demonstrates how LSA can be used as a tool for analyzing large-scale social media data, offering valuable input for policymakers. Future research could expand on this by using more advanced models or larger datasets to gain deeper insights.

Keywords: Topic Modeling; Latent Semantic Analysis; TruncatedSVD; Ministry of Finance; Sentiment Analysis;

INTRODUCTION

In this era of rapid technological development, almost all levels of society depend on and make social media the main source of information and a place to interact. One platform that social media users widely use is Twitter (X). Twitter allows users to post tweets of opinions and news and become a real-time discussion place. Launching from dataindonesia.id and based on We Are Social and Meltwater reports, Indonesia ranks as the fourth largest Twitter (X) user in the world, with a total of 24.85 million users as of April 2024 (Stevany, 2024). This proves that Twitter is one of the popular platforms Indonesians use to express public opinion on various topics, including economic and financial policies. One of the topics that is often discussed on Twitter is the Ministry of Finance of the Republic of Indonesia (Kemenkeu). The Ministry of Finance has a vital role in managing the country's fiscal and economic policies, and it can be concluded that the Ministry of Finance is one of the essential institutions in the Indonesian government. The Ministry of Finance's primary concern is financial and budgetary control, especially since traditional parliamentary attention is only on financial and budgetary control (Hepworth, 2024). The flurry of discussion related to the Ministry of Finance means that the public is scrutinizing the image of this institution. Tweets of public opinion or diverse responses from the public to policies issued by an institution are called sentiments. On social media platforms, sentiment refers to user-shared beliefs, feelings, and attitudes about various entities categorized as positive, negative, or neutral (Parveen et al., 2023). The results of sentiment analysis can be used to identify a government institution's image in the public's eyes as the discussion is being conducted. This sentiment is crucial because it can affect policy implementation and be a consideration when the institution decides on new policies (Wang et al., 2020).

The large amount of data generated from these online discussions, where the scope of people's conversations is vast and not about a single topic, creates a challenge to get a deep understanding of the main topics that people are discussing related to the Ministry of Finance. One approach that can be used effectively based on large text data sets is topic modeling. In text mining and natural language processing (NLP), topic modeling is a technique for locating latent structures in text data to extract and learn topics, facilitating the comprehension and organization

*name of corresponding author

of massive text corpora (Chen et al., 2023). NLP is critical in helping to add structure and resolve ambiguities in language for text analysis and speech recognition (Siddhartha B S & N. M. Niveditha, 2021). Topic modeling is a technique that can identify and categorize latent topics from a set of text data, making it easier for researchers to understand the main themes discussed by Twitter users.

In the scope of government, especially the Ministry of Finance, no one has discussed topic modeling of public sentiment related to the Ministry of Finance's policies. However, NLP has been used in several other fields to create topic modeling, for example, in research on identifying hidden patterns of COVID-19 fake news. This research discusses how to solve the problem of spreading misinformation about COVID-19 by developing an analysis approach that combines sentiment analysis and topic modeling. Among the models used, namely Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), the highest coherence score in LDA is 0.66 for 20 negative sentiment topics and 0.573 for 18 positive fake news topics. This research provides a valuable method for detecting and analyzing misinformation and emphasizes the importance of understanding the patterns of fake news related to COVID-19 (Ahammad, 2024).

In a different field, topic modeling is also used to identify clusters of specific crime problems from free-text unstructured modus operandi. The study analyzed free-text narrative descriptions of house burglaries over two years in major metropolitan areas in the UK and found that the topic modeling algorithm could cluster burglary problems substantively and be applied to support operational decisions (Birks et al., 2020). LDA is often used in research to create topic modeling; apart from the scope of health and crime topic modeling using LDA also exists in software engineering research to see how topics are generated from the model (Silva et al., 2021), subevents detection research by creating scalable and modular topic modeling algorithms to identify sub-events and create labels so that the representation is more accurate by evaluating the approach using two large-scale Twitter corpus, namely Brazilian political protests and the Zika Virus epidemic in the world (Nolasco & Oliveira, 2019). As well as in research that discusses public discussions about the impact of AI and offers contextual topic modeling to identify main scientific topics, sub-themes, and cross-topic themes about AI in the energy field by combining LDA, BERT, and clustering, which results in 8 topics with the methods used (Saheb et al., 2022).

Based on some of the previous research above, in which topic modeling did not use the LSA method much, the author realized this analysis gap, so they decided to use this method in topic modeling research from public sentiment. Also there has been no research related to topic modeling using the LSA model in government agencies, especially the Ministry of Finance. Therefore, this analysis is expected to provide deeper insights into public perceptions of policies taken by the Ministry of Finance and assist policymakers in understanding issues of concern to the public. Latent Semantic Analysis (LSA) is one of the popular methods used in topic modeling. LSA is a topic modeling method that extracts underlying associations between words in textual data by reducing the dimensionality of a term-document matrix (Silva et al., 2021).

This research aims to apply LSA in topic modeling on tweets containing the keyword ‘Kemenkeu’ with the TruncatedSVD approach. TruncatedSVD works more efficiently with sparse matrices as it does not center the data before applying SVD (Egorova et al., 2022). With this approach, the research is expected to identify the main topics that are the focus of public conversations related to the Ministry of Finance on the Twitter platform. The research consists of several stages, starting with data crawling from Twitter using the keyword 'Kemenkeu', then the preprocessing stage of text data to clean and normalize the data, and end with the application of Latent Semantic Analysis (LSA) based on TruncatedSVD for topic modeling. The analysis results are also presented in the form of visualizations to facilitate understanding of the interpretation in this study.

LITERATURE REVIEW

Latent Semantic Analysis (LSA) is a powerful technique in the field of natural language processing (NLP) and information retrieval that leverages the statistical properties of word co-occurrences to uncover latent structures in textual data. LSA operates on the premise that words that are used in similar contexts tend to have similar meanings, which allows it to capture semantic relationships between terms and documents effectively.

One of the foundational aspects of LSA is its ability to reduce dimensionality in text data, which is crucial for improving the efficiency of various NLP tasks. By applying singular value decomposition (SVD) to a term-document matrix, LSA can identify patterns and relationships that are not immediately apparent in the raw data. This process not only enhances the retrieval of relevant documents but also aids in topic modeling, where LSA can reveal underlying themes within a corpus of texts (Silva et al., 2021). For instance, Silva et al. discuss the role of probabilistic topic models, including LSA, in software engineering research, highlighting their utility in discovering topics through statistical analysis of word frequencies and co-occurrences (Silva et al., 2021). Similarly, the combination of LSA with fuzzy clustering methods has shown promise in improving topic detection accuracy, as demonstrated by Murfi et al.(Murfi et al., 2022).

Moreover, LSA's application extends beyond traditional text analysis into more specialized domains. For example, Huyut and Meram propose a method for creating regulation-relatedness maps using LSA, which provides an objective and statistically grounded approach to understanding relationships between regulatory concepts

(Huyut et al., 2022). This contrasts with conventional methods that often rely on subjective expert judgment, thereby introducing inconsistencies. The ability of LSA to generate such maps underscores its versatility and effectiveness in various contexts.

LSA has been employed to enhance the understanding of public sentiment on social media platforms. For instance, Wang et al. utilized NLP techniques, including LSA, to analyze public sentiment regarding governmental COVID-19 measures, demonstrating how LSA can be integrated into broader sentiment analysis frameworks (Wang et al., 2020). This application illustrates LSA's capability to process and interpret large volumes of unstructured data, providing insights into public opinion dynamics.

Furthermore, LSA has been instrumental in addressing challenges associated with unstructured data, such as non-standard words and text normalization. As noted by Finansyah et al., the normalization of text data is essential for effective information retrieval, and LSA can play a significant role in this preprocessing stage by enhancing the semantic coherence of the data (Finansyah et al., 2022). This preprocessing is vital for ensuring that subsequent analyses yield accurate and meaningful results.

METHOD

This research aims to perform topic modeling using Latent Semantic Analysis (LSA) based on Truncated Singular Value Decomposition (TruncatedSVD) on Twitter data with the keyword 'Kemenkeu'. The stages of this research are divided into several main steps, consisting of data collection, data preprocessing by applying several different cases, LSA application, and comparative analysis results from different cases, as shown in Fig. 1.

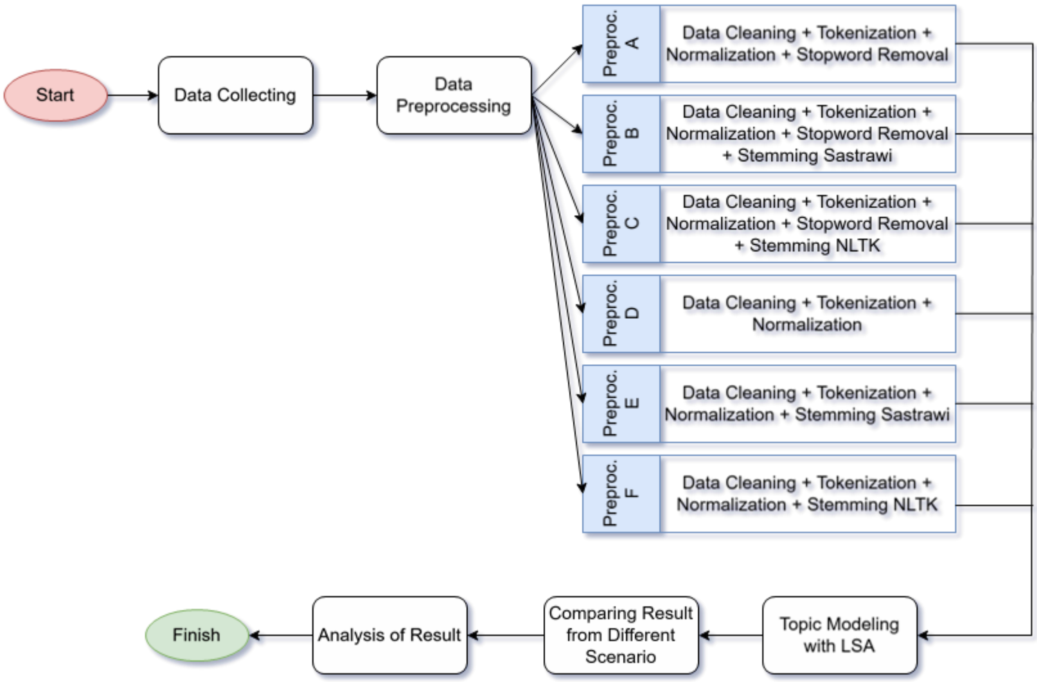

Start
Data Collecting
Data Preprocessing
Preproc. A
Data Cleaning + Tokenization + Normalization + Stopword Removal
Preproc. B
Data Cleaning + Tokenization + Normalization + Stopword Removal + Stemming Sastrawi
Preproc. C
Data Cleaning + Tokenization + Normalization + Stopword Removal + Stemming NLTK
Preproc. D
Data Cleaning + Tokenization + Normalization
Preproc. E
Data Cleaning + Tokenization + Normalization + Stemming Sastrawi
Preproc. F
Data Cleaning + Tokenization + Normalization + Stemming NLTK
Topic Modeling with LSA
Comparing Result from Different Scenario
Analysis of Result
Finish

Fig 1. Research Flow Diagram

Data Collection

Data collection is a methodical process of obtaining information for research objectives from various sources such as focus groups, interviews, records, and technological devices. The data in this research is taken from the Twitter (X) platform through a data crawling process using the Tweet Harvest tool to obtain tweets and relevant metadata. The data collected from January 1, 2024, to July 10, 2024, with a total of 10,099 tweets, consisting of 14 columns (conversation_id_str, created_at, favorite_count, full_text, id_str, in_reply_to_screen_name, location, quote_count, reply_count, retweet_count, tweet_url, user_id_str, username, F1) which are then used as the basis for analysis in this study.

Data Preprocessing

After the data is obtained, the next step is preprocessing to clean and prepare it for analysis. Data preprocessing in text mining involves cleaning and translating raw text data into a structured format, including

*name of corresponding author

stages such as Tokenization, normalization, and eliminating extraneous information to improve analysis accuracy (Viani et al., 2021). The following is a further explanation of the preprocessing stage in this research.

Data Cleaning

This cleaning process removes non-alphabetic characters, such as numbers, punctuation marks (except underscore), irrelevant symbols, emojis, and mentions, and deletes retweet words and account usernames in the data. In text mining, data cleaning must be done because this process affects the quality of the dataset that will be used in the research. In addition, cleaning also facilitates the execution process and provides optimal results so that the research objectives can be adequately achieved. Data cleaning has a vital role in data management because quality decisions also come from quality data (Peng et al., 2024).

Tokenization

Tokenization in the NLTK library separates each text in the data into single-word units (tokens) (Morozovskii & Ramanna, 2023). Tokenization helps understand and analyze text by separating information into more manageable parts. With Tokenization, data analysis can be done in more depth and provide better insight into the data used in the research.

Normalization

The text normalization stage on the data is technically done by converting the text into lowercase letters and correcting inappropriate or non-standard spelling. Text normalization can remove punctuation marks, spell correction, or rearrange words (Finansyah et al., 2022). The words that become the benchmark for word repair consist of a word repair dictionary dataset, which is then used to repair words that have been tokenized in this study.

Stopword Removal

The stopword removal stage is the part to remove common words that have no special meaning in topic analysis, such as “dan” (and), “atau” (or), “itu” (that). As for this research, the author determines the stopword words that do not want to be removed consisting of “tidak” (no), “belum” (yet), “bukan” (not), “namun” (however), “daripada” (than), “sebelum” (before), “sedangkan” (while), “kecuali” (except), “selain” (besides). Some of the deleted words can change the meaning and information of the sentence because some have an important role when combined with other sentences (Huwaidah et al., 2021). Therefore, this customization is done to adjust the preprocessing to suit the purpose of the research analysis.

Stemming

Stemming is the process of converting a word into its base form. It can also be interpreted as reducing inflection in words to their primary form; sometimes, some words are invalid in the stemming base language used (Qorib et al., 2023). In the scope of Indonesian, the available stemming methods are proven to provide high-accuracy results, but stemming for non-formal Indonesian text processing is not much (Rianto et al., 2021). The stemming used in the scenario is stemming from NLTK and Sastrawi. Natural Language Toolkit (NLTK) is a Python package commonly used for research and educational purposes. It includes text preprocessing, Tokenization, and other accessible features (Nair & Thushara, 2024). While Sastrawi is a stemming library that can perform tasks for text processing in Indonesian (Hubert et al., 2021). Stemming aims to simplify text data by changing words to their basic form, so that the data is more consistent, easier to analyze, and ready to be used in modeling.

Topic Modeling with Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a method that locates relevant data from a large data document and allows for detecting similarities between data (Huyut et al., 2022). LSA works by finding semantic relationships between words in documents using a Singular Value Decomposition (SVD) based approach. SVD is a matrix algebra technique that reorients and moves dimensions in vector space (Huyut et al., 2022). In this study, researchers used the Truncated Singular Value Decomposition (TruncatedSVD) method to handle large data sets and improve data processing efficiency. The following is a more detailed explanation of this stage.

Vectorization

Vectorization can also be referred to as Term-Document Matrix (TDM). This stage constructs a term-document matrix X. This matrix has a size of $m \times n$, where m is the number of unique words (terms) in the corpus and n is the number of documents in the corpus. The identified documents are then placed into the x-value of the matrix, while the unique words across all documents are placed on the y-axis. These terms appear in more than one document (Sagum et al., 2023).

Each element x_{ij} in the X matrix indicates the weight of word i in document j. At this stage, each document is represented as a vector in word space. The following is the equation form for the vectorization stage or term-document matrix in number 1.

$$\mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \qquad (1)$$

Dimensionality Reduction (TruncatedSVD)

Truncated Singular Value Decomposition (TruncatedSVD) is performed after the vectorization process to reduce the dimensionality of the numerical representation generated by TF-IDF. TruncatedSVD produces a lower dimensional representation with orthogonal dimensions or features (Murfi et al., 2022) and is applied to retain the most relevant important information while reducing computational complexity. Reducing the dimension of the original feature space is an important step in text processing (Li et al., 2024). LSA uses this reduced dimensionality to find semantic relationships between words in the data. TruncatedSVD only retains most of the decomposition result's top components (dimensions). Where k is the number of retained components (reduced dimensions), which is smaller than the number of original dimensions and is written as in number 2.

$$X_k = U_k S_k V_k^T \qquad (2)$$

Approximate Reconstruction of the Term-Document Matrix (TDM)

After dimensionality reduction is performed, an approximation representation of the original TDM is obtained, X_k. This representation preserves the structure of the most important topics by discarding the less significant components, written as in number 3.

$$X_k = U_k S_k V_k^T \qquad (3)$$

This X_k matrix represents the relationship between documents and words in a new topic space with lower dimensions. The relationship between words and documents can be identified at this stage because redundancy and noise in the initial data have been minimized.

Topic Modeling and Similarity

The last stage of this method is to analyze the results of dimension reduction. Topics in documents and words can be identified based on the values in the matrix. In addition, the similarity between documents and words is calculated by performing operations such as cosine similarity on the result vectors of LSA. The following is the cosine similarity equation between two documents, d_i and d_j, as shown in number 4.

$$\mathit{Cosine\ Similarity}\ (d_i, d_j) = \frac{v_i^T v_j}{||v_i||\,||v_j||} \qquad (4)$$

V_i and V_j are document vectors that have been reduced in dimension in the V_i^T matrix. With this semantic relationship between words and documents, it can be easier to analyze the topic context. The LSA model then identifies hidden topics in the tweet data based on the distribution of words and documents. The authors then determine the number of topics and their topic theme labels based on the experimental results by considering the interpretation and relevance of the resulting topics.

Comparing Results from Different Scenarios

After the whole series of preprocessing and topic modeling, the next step is to compare the results of all scenarios applied in the research to find the best results. This research compares the results of several scenarios in preprocessing consisting of (a) Stopword removal without stemming, (b) Stopword removal with Sastrawi stemming, (c) Stopword removal with NLTK stemming, (d) Without stopword removal and stemming, (e) Without stopword removal with Sastrawi stemming, (f) Without stopword removal with NLTK stemming. This analysis is based on the results of the words (keywords) that appear in the topic modeling process using LSA and based on personal analysis by considering the keywords produced. The author makes an assessment limitation, namely avoiding the number of keywords that appear that are not informative and not by the Ministry of Finance. The frequency of occurrence of keywords related to the context of the Ministry of Finance is prioritized in this research

RESULT

All processes carried out in this study are by those described in the research methodology section, which includes data collecting, data preprocessing, topic modeling using LSA, and comparing the results of different scenarios. The following is the result of each process, which is also the research output.

Data Collection

Data collection for research was carried out using the Tweet Harvest tool, with a total acquisition of 10,099 records for the keyword search 'Kemenkeu'. As explained in the method section, using the tools used, 14 attributes (columns) contain tweet details. Table 1 shows the raw crawling results and examples of some of the records obtained in the data collection process.

Table 1. Raw Data Crawling Results

indeks	conversation_id_str	created_at	full_text	eng
0	1741959210336764321	Mon Jan 01 23:06:54 +0000 2024	Kementerian Keuangan (Kemenkeu) secara resmi telah menaikkan tarif cukai hasil tembakau (CHT) rata-rata 10% pada awal 2024. Harga rokok jadi makin mahal. https://t.co/014NtplWgL	The Ministry of Finance (Kemenkeu) has officially increased the excise tax on tobacco products (CHT) by an average of 10% in early 2024. The price of cigarettes has become more expensive. https://t.co/014NtplWgL
1	1741945163239592203	Mon Jan 01 22:11:05 +0000 2024	Kemenkeu Pastikan Gaji PNS Naik 8 Persen per 1 Januari Namun Dirapel https://t.co/JgeCQd9b3c #aktualcom #Aktualofficial	Ministry of Finance Ensures Civil Servant Salaries Increase by 8 Percent as of January 1 But in Arrears https://t.co/JgeCQd9b3c #aktualcom #Aktualofficial
2	1741613826691440780	Mon Jan 01 17:04:12 +0000 2024	@IrmansyahAdhy Food Estatenya Gagal dan Lahan Terbengkalai (Data Walhi dan Greenpeace).!! Kalau Hilirisasi.. kenapa tidak ada Pemasukan yg Signifikan untuk RI (Data di Kemenkeu).Jilat boleh.. Tolol jangan..!!	@IrmansyahAdhy Food Estate Failed and Land Abandoned (Data from Walhi and Greenpeace).!! If Downstream.. why is there no Significant Income for RI (Data from the Ministry of Finance). You can lick it.. Don't be stupid..!!

Data Preprocessing

A data preprocessing stage is required to facilitate text data processing and optimize topic modeling results using LSA. This stage includes cleaning, Tokenization, normalization, stopword removal, and stemming. During this process, several different scenarios are tested to obtain the most optimal results. The details of the process and results will be explained thoroughly in the following stages.

Data Cleaning

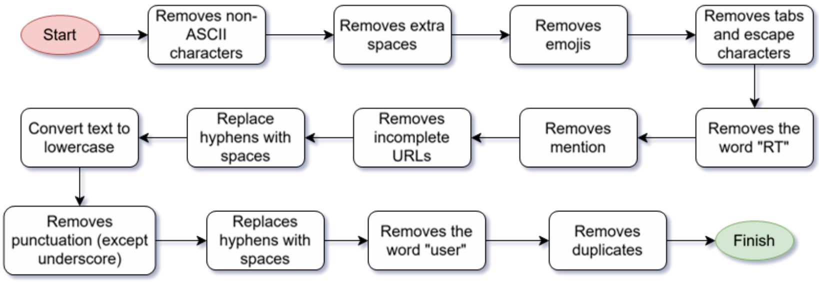

Start
Removes non-ASCII characters
Removes extra spaces
Removes emojis
Removes tabs and escape characters
Removes the word "RT"
Removes mention
Removes incomplete URLs
Replace hyphens with spaces
Convert text to lowercase
Removes punctuation (except underscore)
Replaces hyphens with spaces
Removes the word "user"
Removes duplicates
Finish

Fig 2. Data Cleaning Flow Diagram

*name of corresponding author

After cleaning the data on the entire dataset, as shown in Fig. 2, clean text is produced without capital letters, hashtags, numbers, and punctuation marks. This data cleaning maximizes other preprocessing sequences that will be carried out after data cleaning to produce maximum results. Table 2 shows a comparison of raw text data obtained after crawling and text that has passed the cleaning process.

Table 2. Data Cleaning Result

full_text	clean_text	eng
Kementerian Keuangan (Kemenkeu) secara resmi telah menaikkan tarif cukai hasil tembakau (CHT) rata-rata 10% pada awal 2024. Harga rokok jadi makin mahal. https://t.co/014NtplWgL	kementerian keuangan kemenkeu secara resmi telah menaikkan tarif cukai hasil tembakau cht rata rata percent pada awal number harga rokok jadi makin mahal	The Ministry of Finance has officially increased the excise tariff on tobacco products by an average of 100 percent at the start of the year, making cigarette prices even more expensive.
Kemenkeu Pastikan Gaji PNS Naik 8 Persen per 1 Januari Namun Dirapel https://t.co/JgeCQd9b3c #aktualcom #Aktualofficial	kemenkeu pastikan gaji pns naik number persen per number januari namun dirapel aktual com aktualofficial	Ministry of Finance ensures civil servant salaries increase by 10 percent per January, but are paid in arrears aktual com aktualofficial
@IrmansyahAdhy Food Estatenya Gagal dan Lahan Terbengkalai (Data Walhi dan Greenpeace).!! Kalau Hilirisasi.. kenapa tidak ada Pemasukan yg Signifikan untuk RI (Data di Kemenkeu).Jilat boleh.. Tolol jangan..!!	food estatenya gagal dan lahan terbengkalai data walhi dan greenpeace kalau hilirisasi kenapa tidak ada pemasukan yg signifikan untuk ri data di kemenkeu jilat boleh tolol jangan	the food estate failed and the land was abandoned, data from Walhi and Greenpeace, if it was downstream, why wasn't there any significant income for RI? Data from the Ministry of Finance, licking it is allowed, idiots, don't do it

Tokenization

After data cleaning is done, the next step is tokenizing each word in the text in the document. This process separates each word and gives it a code/token for each word. The results of this tokenization process can be seen in Table 3.

Table 3. Tokenization Result

clean_text	tokens	eng
kementerian keuangan kemenkeu secara resmi telah menaikkan tarif cukai hasil tembakau cht rata rata percent pada awal number harga rokok jadi makin mahal	['kementerian', 'keuangan', 'kemenkeu', 'secara', 'resmi', 'telah', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'rata', 'rata', 'percent', 'pada', 'awal', 'number', 'harga', 'rokok', 'jadi', 'makin', 'mahal']	['ministry', 'finance', 'kemenkeu', 'officially', 'has', 'raised', 'tariff', 'excise', 'results', 'tobacco', 'cht', 'average', 'average', 'percent', 'at', 'beginning', 'number', 'price', 'cigarettes', 'so', 'more', 'expensive']
kemenkeu pastikan gaji pns naik number persen per number januari namun dirapel aktual com aktualofficial	['kemenkeu', 'pastikan', 'gaji', 'pns', 'naik', 'number', 'persen', 'per', 'number', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'make sure', 'salary', 'civil servants', 'increase', 'number', 'percent', 'per', 'number', 'january',

*name of corresponding author

		'however', 'raised', 'actual', 'com', 'aktualofficial']
food estatenya gagal dan lahan terbengkalai data walhi dan greenpeace kalau hilirisasi kenapa tidak ada pemasukan yg signifikan untuk ri data di kemenkeu jilat boleh tolol jangan	['food', 'estatenya', 'gagal', 'dan', 'lahan', 'terbengkalai', 'data', 'walhi', 'dan', 'greenpeace', 'kalau', 'hilirisasi', 'kenapa', 'tidak', 'ada', 'pemasukan', 'yg', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estate', 'failed', 'and', 'land', 'abandoned', 'data', 'walhi', 'and', 'greenpeace', 'if', 'downstream', 'why', 'not', 'there is', 'revenue', 'that', 'significant', 'for', 'ri', 'data', 'in', 'kemenkeu', 'lick', 'allowed', 'stupid', 'don't']

Normalization

The next stage in the preprocessing process is the normalization stage. At this stage, each word that has been separated and already has a token is converted into its formal form. Changing informal words to formal is done to approach the language standards used and improve the quality of modeling. In Table 4 below, it can be seen that several words are not standardized, for example, "mending", which is then changed to "lebih baik" (better) at the normalization stage.

Table 4. Normalization Result

tokens	normalize	eng
['food', 'estatenya', 'gagal', 'dan', 'lahan', 'terbengkalai', 'data', 'walhi', 'dan', 'greenpeace', 'kalau', 'hilirisasi', 'kenapa', 'tidak', 'ada', 'pemasukan', 'yg', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estatenya', 'gagal', 'dan', 'lahan', 'terbengkalai', 'data', 'walhi', 'dan', 'greenpeace', 'kalau', 'hilirisasi', 'kenapa', 'tidak', 'ada', 'pemasukan', 'yang', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estate', 'failed', 'and', 'land', 'abandoned', 'data', 'walhi', 'and', 'greenpeace', 'if', 'downstream', 'why', 'not', 'there is', 'revenue', 'which', 'significant', 'for', 'ri', 'data', 'in', 'kemenkeu', 'lick', 'allowed', 'stupid', 'don't']
['kemenkeu', 'pastikan', 'gaji', 'pns', 'naik', 'number', 'persen', 'per', 'number', 'januari', 'tapi', 'dirapel']	['kemenkeu', 'pastikan', 'gaji', 'pns', 'naik', 'nomor', 'persen', 'per', 'nomor', 'januari', 'tapi', 'dirapel']	['kemenkeu', 'make sure', 'salary', 'civil servants', 'increase', 'number', 'percent', 'per', 'number', 'January', 'but', 'increase']
['kalau', 'pintar', 'apalagi', 's2', 'luar', 'negeri', 'mending', 'cari', 'kerja', 'di', 'singapore', 'atau', 'hong', 'kong', 'aja', 'pns', 'non', 'kemenkeu', 'hidupnya', 'melarat']	['kalau', 'pintar', 'apalagi', 's2', 'luar', 'negeri', 'lebih baik', 'cari', 'kerja', 'di', 'singapore', 'atau', 'hong', 'kong', 'saja', 'pns', 'non', 'kemenkeu', 'hidupnya', 'melarat']	['if', 'smart', 'especially', 's2', 'abroad', 'country', 'better', 'look for', 'work', 'in', 'singapore', 'or', 'hong', 'kong', 'only', 'civil servant', 'non', 'kemenkeu', 'life', 'poor']

Stopword Removal

Stopword removal is a stage performed to remove common words without meaning for the analysis process. Examples of stopword words in Indonesian are "dan" (and), "yang" (which), "di" (at), "ke" (to), "dengan" (with), "itu" (that), "ini" (these), "adalah" (is), "dalam" (in). These words have no critical information value for modeling, so they are removed at the preprocessing stage. However, in this study, the author conducted custom stopwords because he saw that some common words that often appear (stopwords) in documents when combined with other essential words have different meanings when compared to those not accompanied by stopword words. Therefore, custom stopwords are used in modeling. The words that are not wanted in this stage include "tidak" (no), "belum" (yet), "bukan" (not), "namun" (however), "daripada" (than), "sebelum" (before), "sedangkan" (while), "kecuali" (except), "selain" (besides). Can be seen below in Table 5 the difference in words before and after stopword removal.

Table 5. Stopword Removal Result

normalize	stopword_removal	eng
['kementerian', 'keuangan', 'kemenkeu', 'secara', 'resmi', 'telah', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'rata', 'rata', 'persen', 'pada', 'awal', 'nomor', 'harga', 'rokok', 'jadi', 'makin', 'mahal']	['kementerian', 'keuangan', 'kemenkeu', 'resmi', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'persen', 'nomor', 'harga', 'rokok', 'mahal']	['ministry', 'finance', 'kemenkeu', 'official', 'increase', 'tariff', 'excise', 'results', 'tobacco', 'cht', 'percent', 'number', 'price', 'cigarettes', 'expensive']
['kemenkeu', 'pastikan', 'gaji', 'pns', 'naik', 'nomor', 'persen', 'per', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'pastikan', 'gaji', 'pns', 'nomor', 'persen', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'make sure', 'salary', 'civil servants', 'number', 'percent', 'number', 'january', 'however', 'received', 'actual', 'com', 'actualofficial']
['food', 'estatenya', 'gagal', 'dan', 'lahan', 'terbengkalai', 'data', 'walhi', 'dan', 'greenpeace', 'kalau', 'hilirisasi', 'kenapa', 'tidak', 'ada', 'pemasukan', 'yang', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estatenya', 'gagal', 'lahan', 'terbengkalai', 'data', 'walhi', 'greenpeace', 'hilirisasi', 'tidak', 'pemasukan', 'signifikan', 'ri', 'data', 'kemenkeu', 'jilat', 'tolol']	['food', 'estate', 'fail', 'land', 'abandoned', 'data', 'walhi', 'greenpeace', 'downstream', 'not', 'revenue', 'significant', 'ri', 'data', 'kemenkeu', 'lick', 'stupid']

Stemming

Stemming is a process of converting words in documents into root form (basic words); in this case, stemming removes certain affixes and prefixes and then returns them to basic words. In this research, the author compares two libraries supporting the stemming process, namely, NLTK and Sastrawi. In its implementation, it is also combined with texts that use stopword removal and are not used for some stemming scenarios. The results of each scenario will be explained below using either NLTK or Sastrawi.

a. NLTK stemming

NLTK is basically focuses on processing data in English and does not provide special support for Indonesian. Therefore, in Table 6, it can be seen that the stemming results using NLTK are not significant and almost do not even change the entire initial word to the root form. In some words, there is also a deletion of letters, which causes the word to have no meaning because the letters are incomplete.

Table 6. Result of Stemming Using NLTK Without Stopword Removal

normalize	stemming	eng
['kementerian', 'keuangan', 'kemenkeu', 'secara', 'resmi', 'telah', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'rata', 'rata', 'persen', 'pada', 'awal', 'nomor', 'harga', 'rokok', 'jadi', 'makin', 'mahal']	['kementerian', 'keuangan', 'kemenkeu', 'secara', 'resmi', 'telah', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'rata', 'rata', 'persen', 'pada', 'awal', 'nomor', 'harga', 'rokok', 'jadi', 'makin', 'mahal']	['ministry', 'finance', 'kemenkeu', 'officially', 'has', 'increased', 'tariff', 'excise', 'results', 'tobacco', 'cht', 'average', 'average', 'percent', 'at', 'beginning', 'number', 'price', 'cigarettes', 'so', 'more', 'expensive']
['kemenkeu', 'pastikan', 'gaji', 'pns', 'naik', 'nomor', 'persen', 'per', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'pastikan', 'gaji', 'pn', 'naik', 'nomor', 'persen', 'per', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualoffici']	['kemenkeu', 'make sure', 'salary', 'pn', 'increase', 'number', 'percent', 'per', 'number', 'january', 'however', 'received', 'actual', 'com', 'aktualoffici']

*name of corresponding author

['food', 'estatenya', 'gagal', 'dan', 'lahan', 'terbengkalai', 'data', 'walhi', 'dan', 'greenpeace', 'kalau', 'hilirisasi', 'kenapa', 'tidak', 'ada', 'pemasukan', 'yang', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estatenya', 'gagal', 'dan', 'lahan', 'terbengkalai', 'data', 'walhi', 'dan', 'greenpeac', 'kalau', 'hilirisasi', 'kenapa', 'tidak', 'ada', 'pemasukan', 'yang', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estate', 'failed', 'and', 'land', 'abandoned', 'data', 'walhi', 'and', 'greenpeac', 'if', 'downstream', 'why', 'not', 'there is', 'revenue', 'which', 'significant', 'for', 'ri', 'data', 'in', 'kemenkeu', 'lick', 'allowed', 'stupid', 'don't']

This condition also occurs in scenario experiments that use stopword removal. Table 7 shows that the results do not change after the stemming process is performed.

Table 7. Result of Stemming Using NLTK With Stopword Removal

stopword_removal	stemming	eng
['kementerian', 'keuangan', 'kemenkeu', 'resmi', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'persen', 'nomor', 'harga', 'rokok', 'mahal']	['kementerian', 'keuangan', 'kemenkeu', 'resmi', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'persen', 'nomor', 'harga', 'rokok', 'mahal']	['ministry', 'finance', 'kemenkeu', 'official', 'increase', 'tariff', 'excise', 'results', 'tobacco', 'cht', 'percent', 'number', 'price', 'cigarettes', 'expensive']
['kemenkeu', 'pastikan', 'gaji', 'pns', 'nomor', 'persen', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'pastikan', 'gaji', 'pn', 'nomor', 'persen', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualoffici']	['kemenkeu', 'make sure', 'salary', 'pn', 'number', 'percent', 'number', 'january', 'however', 'received', 'actual', 'com', 'aktualoffici']
['food', 'estatenya', 'gagal', 'lahan', 'terbengkalai', 'data', 'walhi', 'greenpeace', 'hilirisasi', 'tidak', 'pemasukan', 'signifikan', 'ri', 'data', 'kemenkeu', 'jilat', 'tolol']	['food', 'estatenya', 'gagal', 'lahan', 'terbengkalai', 'data', 'walhi', 'greenpeac', 'hilirisasi', 'tidak', 'pemasukan', 'signifikan', 'ri', 'data', 'kemenkeu', 'jilat', 'tolol']	['food', 'estate', 'fail', 'land', 'abandoned', 'data', 'walhi', 'greenpeac', 'downstream', 'not', 'revenue', 'significant', 'ri', 'data', 'kemenkeu', 'lick', 'stupid']

b. Sastrawi stemming

The steaming process using Sastrawi produces better results, and there is a difference in the change of affixed words into their basic form only. This can be seen in Table 8, which compares before and after the use of stemming using Sastrawi in the scenario without stopword removal.

Table 8. Result of Stemming Using Sastrawi Without Stopword Removal

normalize	stemming	eng
['kementerian', 'keuangan', 'kemenkeu', 'secara', 'resmi', 'telah', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'rata', 'rata', 'persen', 'pada', 'awal', 'nomor', 'harga', 'rokok', 'jadi', 'makin', 'mahal']	['menteri', 'uang', 'kemenkeu', 'cara', 'resmi', 'telah', 'naik', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'rata', 'rata', 'persen', 'pada', 'awal', 'nomor', 'harga', 'rokok', 'jadi', 'makin', 'mahal']	['minister', 'money', 'kemenkeu', 'method', 'official', 'has', 'increase', 'tariff', 'excise', 'results', 'tobacco', 'cht', 'average', 'average', 'percent', 'at', 'beginning', 'number', 'price', 'cigarettes', 'so', 'more', 'expensive']

['kemenkeu', 'pastikan', 'gaji', 'pns', 'naik', 'nomor', 'persen', 'per', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'pasti', 'gaji', 'pns', 'naik', 'nomor', 'persen', 'per', 'nomor', 'januari', 'namun', 'rapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'pasti', 'salary', 'civil servants', 'increase', 'number', 'percent', 'per', 'number', 'january', 'however', 'rapel', 'actual', 'com', 'aktualofficial']
['food', 'estatenya', 'gagal', 'dan', 'lahan', 'terbengkalai', 'data', 'walhi', 'dan', 'greenpeace', 'kalau', 'hilirisasi', 'kenapa', 'tidak', 'ada', 'pemasukan', 'yang', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estatenya', 'gagal', 'dan', 'lahan', 'bengkalai', 'data', 'walhi', 'dan', 'greenpeace', 'kalau', 'hilir', 'kenapa', 'tidak', 'ada', 'pasu', 'yang', 'signifikan', 'untuk', 'ri', 'data', 'di', 'kemenkeu', 'jilat', 'boleh', 'tolol', 'jangan']	['food', 'estate', 'fail', 'and', 'land', 'abandoned', 'data', 'walhi', 'and', 'greenpeace', 'if', 'downstream', 'why', 'not', 'there is', 'vasu', 'which', 'significant', 'for', 'ri', 'data', 'in', 'kemenkeu', 'lick', 'allowed', 'stupid', 'don't']

While Table 9 is the result of stemming using Sastrawi in a scenario that uses stopword removal. The stemming results for this scenario are also good and show clear word changes, words that originally had affixes turned into basic words. This shows that Sastrawi is able to produce more consistent and accurate basic word forms for Indonesian, thus increasing the readability of the data by the model and reducing the risk of irrelevant features. In contrast, NLTK stemming results tend to be less suitable for Indonesian, which can affect the effectiveness of the modeling process.

Table 9. Result of Stemming Using Sastrawi With Stopword Removal

stopword_removal	stemming	eng
['kementerian', 'keuangan', 'kemenkeu', 'resmi', 'menaikkan', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'persen', 'nomor', 'harga', 'rokok', 'mahal']	['menteri', 'uang', 'kemenkeu', 'resmi', 'naik', 'tarif', 'cukai', 'hasil', 'tembakau', 'cht', 'persen', 'nomor', 'harga', 'rokok', 'mahal']	['minister', 'money', 'minister of finance', 'official', 'increase', 'tariff', 'excise', 'results', 'tobacco', 'cht', 'percent', 'number', 'price', 'cigarettes', 'expensive']
['kemenkeu', 'pastikan', 'gaji', 'pns', 'nomor', 'persen', 'nomor', 'januari', 'namun', 'dirapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'pasti', 'gaji', 'pns', 'nomor', 'persen', 'nomor', 'januari', 'namun', 'rapel', 'aktual', 'com', 'aktualofficial']	['kemenkeu', 'pasti', 'salary', 'civil servant', 'number', 'percent', 'number', 'january', 'however', 'rapel', 'actual', 'com', 'aktualofficial']
['food', 'estatenya', 'gagal', 'lahan', 'terbengkalai', 'data', 'walhi', 'greenpeace', 'hilirisasi', 'tidak', 'pemasukan', 'signifikan', 'ri', 'data', 'kemenkeu', 'jilat', 'tolol']	['food', 'estatenya', 'gagal', 'lahan', 'bengkalai', 'data', 'walhi', 'greenpeace', 'hilir', 'tidak', 'pasu', 'signifikan', 'ri', 'data', 'kemenkeu', 'jilat', 'tolol']	['food', 'estate', 'fail', 'land', 'abandoned', 'data', 'walhi', 'greenpeace', 'downstream', 'not', 'vase', 'significant', 'ri', 'data', 'kemenkeu', 'lick', 'stupid']

Topic Modeling Using LSA

The stages of creating topic modeling using LSA in this study consist of several processes, starting with forming the model by importing the corpus to be used and then the vectorization and TruncatedSVD stages. Furthermore, the optimal topic for each scenario is obtained from the coherence results, which then become guidelines in the distribution of topic modeling. Below, the results of each process will be explained in more detail, and a visualization of the results will be provided to make it easier to read the data and see the distribution of data in this study.

a. coherence score analysis

Before clustering the topics in this research, a coherence score analysis is conducted for each preprocessing scenario to determine the optimal number of topics. This optimal number of topics will be the total number of topics used in topic modeling. Table 10 shows the results of the coherence score and optimal number of topics in each preprocessing scenario.

Table 10. Result of Coherence Score and Optimal Number of Topics for Each Scenario

Preprocessing Scenario	Coherence Score	Optimal Number of Topic
A (Stopword Removal + No Stemming)	0.4245	3
B (Stopword Removal + Stemming Sastrawi)	0.5289	3
C (Stopword Removal + Stemming NLTK)	0.4219	3
D (No Stopword Removal + No Stemming)	0.5632	2
E (No Stopword Removal + Stemming Sastrawi)	0.5198	3
F (No Stopword Removal + Stemming NLTK)	0.5634	2

Based on Table 10, it is known that the variation of the optimal number of topics is dominated by the acquisition of 3 topics that occur in 4 scenarios. In comparison, the remaining two scenarios have the acquisition of 2 topics.

b. vectorization

Vectorization is the first stage in topic modeling, which converts text data into numerical representations that algorithms can process. This stage converts text into numeric with the TF-IDF method. The following are the results of the vectorization stage, which can be seen in Table 11.

c. TruncatedSVD

The next step after the data is converted into vector form is to perform dimensionality reduction to see the latent relationship between words and documents. SVD will reduce the dimension of TF-IDF matrix so that only important features are retained. The results of the topics generated after passing this process are as in Table 11 which displays the results of topic modeling for each scenario which contains words that appear and have a relationship in each document so as to define certain topics.

d. topic modeling

Table 11. Topic Modeling Result

Preprocessing Scenario	Topic Modeling Result
A (Stopword Removal + No Stemming)	Topik 1: ('0.706*"nomor" + 0.377*"kemenkeu" + 0.164*"triliun" + 0.157*"pajak" + 0.147*"tidak" + 0.123*"rp" + 0.116*"gaji" + 0.115*"persen" + 0.112*"anggaran" + 0.098*"pns"') Topik 2: ('0.525*"kemenkeu" + 0.398*"tidak" + 0.207*"ya" + 0.157*"cukai" + 0.153*"bea" + 0.109*"pajak" + 0.098*"nya" + 0.096*"orang" + 0.088*"bukan" + 0.072*"magang"') Topik 3: ('0.468*"cukai" + 0.457*"bea" + 0.328*"pajak" + 0.089*"mi" + 0.087*"bp" + 0.081*"dirjen" + 0.078*"ditjen" + 0.068*"kantor" + 0.066*"barang" + 0.065*"keuangan"')

*name of corresponding author

B (Stopword Removal + Stemming Sastrawi)	Topik 1: ('0.666*"nomor" + 0.371*"kemenkeu" + 0.158*"pajak" + 0.155*"tidak" + 0.155*"triliun" + 0.139*"uang" + 0.127*"gaji" + 0.116*"anggar" + 0.115*"menteri" + 0.114*"rp"') Topik 2: ('0.536*"nomor" + 0.162*"triliun" + 0.127*"rp" + 0.054*"persen" + 0.051*"capai" + 0.05*"maret" + 0.041*"cair" + 0.035*"hutang" + 0.035*"catat" + 0.032*"thr"') Topik 3: ('0.475*"gaji" + 0.464*"pns" + 0.201*"cair" + 0.179*"rapel" + 0.169*"naik" + 0.161*"asn" + 0.159*"pensiun" + 0.138*"kemenkeu" + 0.111*"polri" + 0.111*"tni"')
C (Stopword Removal + Stemming NLTK)	Topik 1: ('0.705*"nomor" + 0.377*"kemenkeu" + 0.164*"triliun" + 0.158*"pajak" + 0.148*"tidak" + 0.122*"rp" + 0.116*"gaji" + 0.115*"persen" + 0.112*"anggaran" + 0.097*"pn"') Topik 2: ('0.523*"kemenkeu" + 0.397*"tidak" + 0.207*"ya" + 0.157*"cukai" + 0.154*"bea" + 0.109*"pajak" + 0.098*"nya" + 0.096*"orang" + 0.088*"bukan" + 0.072*"magang"') Topik 3: ('0.47*"cukai" + 0.46*"bea" + 0.326*"pajak" + 0.09*"mi" + 0.087*"bp" + 0.081*"dirjen" + 0.077*"ditjen" + 0.069*"kantor" + 0.066*"barang" + 0.064*"keuangan"')
D (No Stopword Removal + No Stemming)	Topik 1: ('0.448*"nomor" + 0.303*"kemenkeu" + 0.231*"yang" + 0.222*"di" + 0.196*"dan" + 0.168*"tidak" + 0.14*"ini" + 0.136*"dari" + 0.128*"pajak" + 0.127*"ada"') Topik 2: ('0.675*"nomor" + 0.2*"triliun" + 0.157*"rp" + 0.087*"persen" + 0.069*"maret" + 0.067*"gaji" + 0.061*"bisnis" + 0.059*"tahun" + 0.057*"hingga" + 0.055*"hutang"')
E (No Stopword Removal + Stemming Sastrawi)	Topik 1: ('0.41*"nomor" + 0.288*"kemenkeu" + 0.226*"yang" + 0.212*"di" + 0.193*"dan" + 0.162*"tidak" + 0.138*"ini" + 0.131*"ada" + 0.131*"dari" + 0.125*"pajak"') Topik 2: ('0.689*"nomor" + 0.201*"triliun" + 0.159*"rp" + 0.091*"persen" + 0.075*"gaji" + 0.075*"maret" + 0.068*"cair" + 0.067*"capai" + 0.061*"bisnis" + 0.061*"tahun"') Topik 3: ('0.363*"gaji" + 0.34*"pns" + 0.163*"cair" + 0.157*"naik" + 0.141*"rapel" + 0.135*"asn" + 0.123*"tidak" + 0.118*"pensiun" + 0.093*"kalau" + 0.086*"tni"')
F (No Stopword Removal + Stemming NLTK)	Topik 1: ('0.447*"nomor" + 0.302*"kemenkeu" + 0.231*"yang" + 0.222*"di" + 0.196*"dan" + 0.168*"tidak" + 0.14*"ini" + 0.135*"dari" + 0.128*"pajak" + 0.127*"ada"') Topik 2: ('0.675*"nomor" + 0.2*"triliun" + 0.156*"rp" + 0.087*"persen" + 0.069*"maret" + 0.067*"gaji" + 0.061*"bisni" + 0.059*"tahun" + 0.057*"hingga" + 0.055*"sebesar"')

e. comparing results from different scenarios

Based on a series conducted by the author by applying different scenarios, it can be concluded that most scenarios have different results. The following is a detailed explanation of the topic modeling results for each scenario.

1. Stopword Removal + No Stemming (Scenario A)
   The results of scenario A show that the keywords that appear in the topic modeling include “kemenkeu”, “nomor” (number), “pajak” (tax), “rp”, “anggaran” (budget), “ditjen”, “pns”, “gaji” (salary), “bea” (duty) and “cukai” (excise). The words that appear are quite relevant to the research topic, but some meaningless words appear such as “tidak” (no), “bukan” (not), “mi”, “dan” (and) as well as “nya” (its).
2. Stopword Removal + Stemming Sastrawi (Scenario B)
   Literary stemming affects the topic words produced, and the keyword results that appear in scenario B are excellent compared to other scenarios. There are no meaningless words that appear as in other scenarios. The keywords that appear are very relevant to the topic of the Ministry of Finance, such as “pajak” (tax), “uang” (money), “gaji” (salary), “menteri” (minister), “anggar” (budget), “rp”, “hutang” (debt), and so on.
3. Stopword Removal + Stemming NLTK (Scenario C)
   Scenario C with a different stemming application, namely NLTK, produces many keywords that are not meaningful and irrelevant to the Ministry of Finance. In addition, NLTK, which is actually used for English text mining, causes some words that appear not to have complete letters so that they become uninformative, for example, “pn” (which should be “pns” but one letter is removed during the stemming process), “mi”, “bp”, “nya”, “tidak” (no), “ya” (yes), and “not” (bukan).
4. No Stopword Removal + No Stemming (Scenario D)
   The analysis shows that scenario D, excluding stopword removal and stemming processes, produced lower-quality keywords than other scenarios. In the results of this scenario there are also many irrelevant and meaningless words such as “yang” (which), “di” (in), “dan” (and), “tidak” (not), “ini” (this), “dari” (from), “ada” (there is), and “hingga” (until), which these words should disappear if using stopword removal during preprocessing.
5. No Stopword Removal + Stemming Sastrawi (Scenario E)
   Although the use of Sastrawi stemming produces quite good keywords, as in scenario B, in scenario E, the resulting keywords are not good enough because they are not accompanied by stopword removal during preprocessing. The words that are not relevant and informative related to the Ministry of Finance include the words “di” (in), “dan” (and), “yang” (which), “tidak” (not), “ini” (this), and so on (the results of this scenario are not much different from the results of scenario D).
6. No Stopword Removal + Stemming NLTK (Scenario F)
   The results of scenario F using NLTK stemming produce several words that are not meaningful and do not have complete letters. This is not much different from the results of scenario C and also scenario D where many uninformative keywords appear such as “di” (in), “dan” (and), “dari” (from), “yang” (which), and “hingga” (until).

Based on the results of topic modeling by comparing several scenarios to choose the maximum results in this study, the author analyzes that the best result for topic modeling using the LSA method with the ‘Kemenkeu’ keyword is the result of scenario B, namely with the application of stopword removal and Sastrawi stemming. This scenario has a coherence score of 0.5289, which is not the highest score compared to other scenarios but also not too low. It can be identified that the use of stopword removal helps remove meaningless and irrelevant words to the Ministry of Finance topic. In addition, stemming using Sastrawi, a very supportive library for Indonesian text mining, helps generate keywords for topic modeling that are appropriate and in accordance with the desired context. Keywords that appear such as “nomor” (number), “kemenkeu”, “pajak” (tax), “triliun” (trillion), “uang” (money), “gaji” (salary), “anggar” (budget), “menteri” (minister), “rp” have good quality because they are by the context of the Ministry of Finance to facilitate the creation of topic labels in the following process. After calculating the coherence score, scenario B has an optimal topic of 3. The following graph shows the optimal number of topics for this scenario can be seen in Fig. 3.

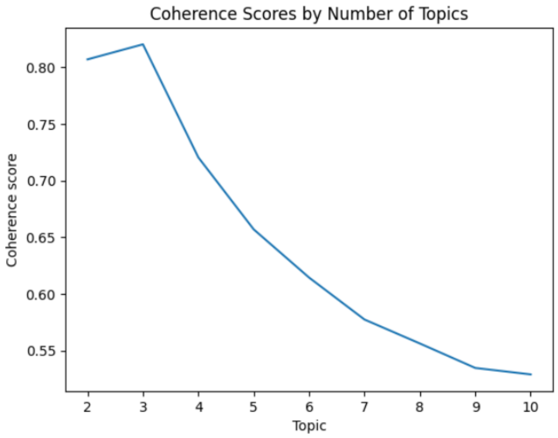
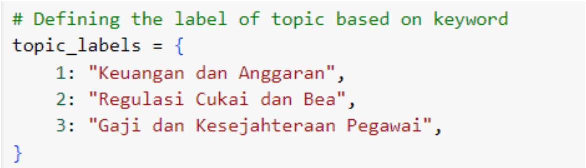


Fig 3. Coherence Score by Number of Topic of Scenario B Line Graph

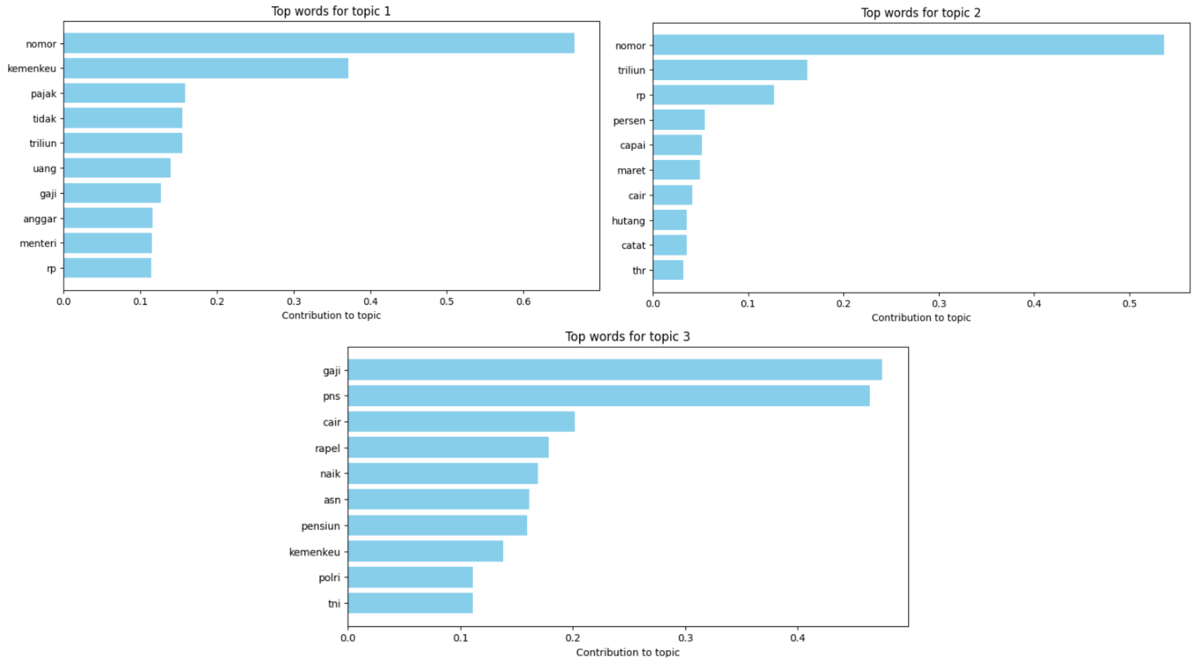
After obtaining the coherence score and the optimal number of topics for scenario B, the author performs labeling to determine the topic theme based on the optimal number of topics. This labeling is done based on the results of the words (keywords) that appear in the topic modeling process, then combines them and interprets them into a new topic theme unit. The result of this topic theme labeling can be seen in Fig. 4.



```
# Defining the label of topic based on keyword
topic_labels = {
    1: "Keuangan dan Anggaran",
    2: "Regulasi Cukai dan Bea",
    3: "Gaji dan Kesejahteraan Pegawai",
}
```

Fig 4. Labeling the Topics Based on Keyword

Based on the combination of keywords in each topic generated by scenario B, there are three main topics consisting of Keuangan dan Anggaran (Finance and Budget), Regulasi Cukai dan Bea (Excise and Customs Regulations), and Gaji dan Kesejahteraan Pegawai (Salaries and Employee Welfare) as shown in Fig. 4. Furthermore, to support the results of the analysis and assist readers in understanding the results of topic modeling, visualization of the results using a combination of matplotlib and Seabron libraries is carried out, which shows a histogram of the frequency of contribution of each keyword in the topic as in Fig. 5 to enable data representation in the form of bars and a more interactive and neat label arrangement.



*name of corresponding author

Fig 5. Keyword Frequency Visualization using Seaborn and Matplotlib

Fig. 5 shows the frequency values of the keywords that appear for each topic, sorted from largest to smallest frequency. For example, the keyword “kemenkeu” has a frequency of 0.37 on topic 1, namely the topic of Keuangan dan Anggaran (Finance and Budget). In contrast, with the same keyword on topic 3 Gaji dan Kesejahteraan Pegawai (Salaries and Employee Welfare), “kemenkeu” obtained a frequency result of 0.13. This proves that not every same keyword has the same semantic frequency for different topics. In addition to using the histogram form to visualize the frequency of keywords that appear in the topic, to facilitate understanding, visualization using word cloud is also carried out, which shows quickly what text content (keywords) appears in the topic and the dominant words according to the size of the word displayed. Here, in Fig. 6, visualization is done using Word Cloud.

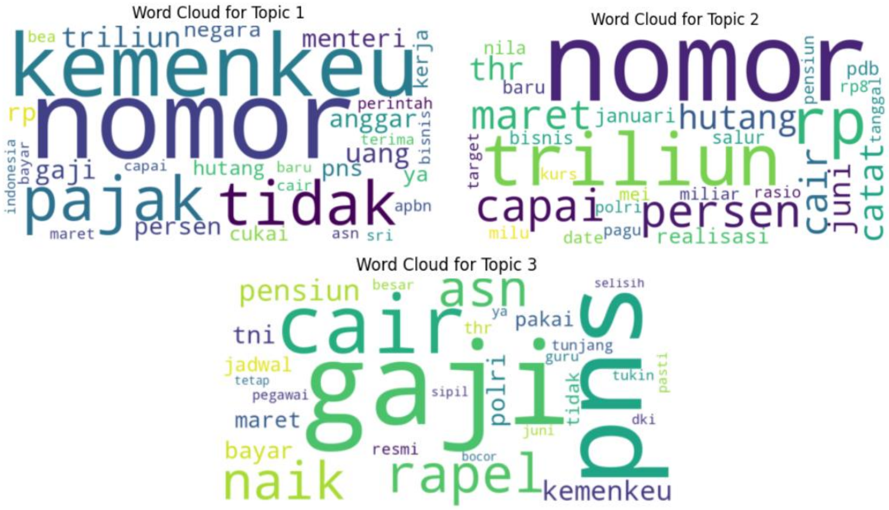


Fig 6. Keyword Frequency Visualization using World Cloud

In addition to the visualization showing the frequency of keywords, the author also uses t-SNE to visualize the distribution of topics in scenario B using different colors; this visualization is shown in Fig. 7.

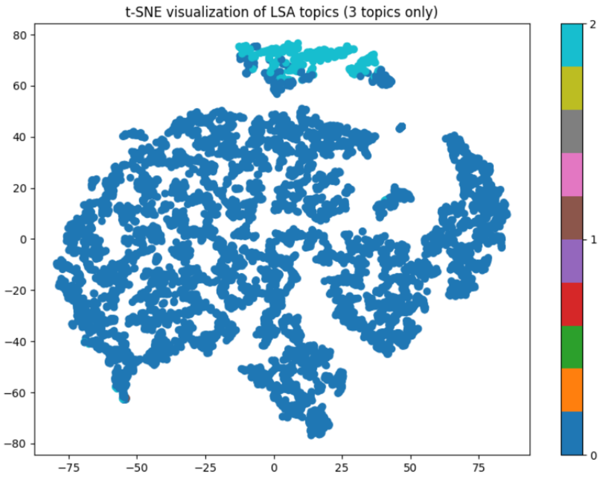


Fig 7. Topics Distribution Visualization

Fig. 7 shows that the distribution of topics in the document appears in 2 prominent colors, namely blue and cyan. This distribution should have three colors because scenario B has three optimal topics that appear. However, the non-appearance of 1 color is because the number of topic distributions in the document is fairly imbalanced compared to the other 2 topics. This is also shown in Fig. 8 below.

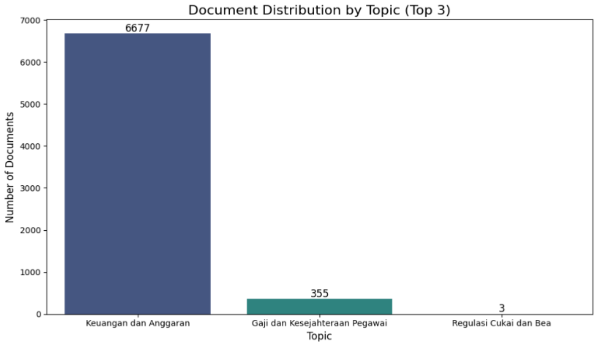


Fig 8. Document Distribution by Topic

Fig. 8 above shows that the distribution of documents for the three topics (sorted by the largest value) that exist in topic modeling using LSA has a considerable difference in number. Topic 1, Keuangan dan Anggaran (Finance and Budget), has a total of 6,677 documents with this topic, and Topic 3, Gaji dan Kesejahteraan Pegawai (Salaries and Employee Welfare), has 355 documents. In contrast, Topic 2, Regulasi Cukai dan Bea (Excise and Customs Regulations), has only three documents. This high imbalance based on the number between Topic 2 and Topics 1 and 3 causes the color of the visualization using t-SNE in Fig. 7 to not appear and is dominated by the color of Topic 1 and Topic 3.

Below in Table 12 is a sample of research results that display raw text tweets (full_text), text tweets after passing the cleaning and stemming process, dominant topics, keyword topics, and topic labels based on keywords. This research identifies what topics are public discussions based on the ‘Kemenkeu’ tweet data obtained. LSA filters each word and makes it a keyword to determine the topic being discussed. The results of this analysis help the Ministry of Finance identify topics that are widely discussed by the public and support internal decision-making related to policies issued. With topic labels, decisions can be taken more quickly and efficiently, while considering public needs in a more targeted manner.

Table 12. Sample Result of Research

doc_ num	full_text	stemming	dom_ topic	keyword	topic_label
19	@user Aturan sma pedomannya kan kalian kemenkeu yg bikin..jkw tnggal teken stlh perintahkn spy buat aturan majakin lbh banyak lgi rakyat...soalny mnjam ke LN udh ga boleh lgi..kbnyakn ngutang ..SDA udh abis dijaminkn sma diambil Cina..	['atur', 'sma', 'pedoman', 'kemenkeu', 'bikin', 'jkw', 'tnggal', 'teken', 'stlh', 'perintahkn', 'spy', 'atur', 'majakin', 'rakyat', 'soalny', 'mnjam', 'ln', 'tidak', 'kbnyakn', 'ngutang', 'sda', 'abis', 'dijaminkn', 'sma', 'ambil', 'cina']	1	rp, menteri, anggar, gaji, uang, triliun, tidak, pajak, kemenkeu, nomor	Keuangan dan Anggaran
20	@user Jadi pns kemenkeu terus penempatan madiun bisa wkwkwkww	['pns', 'kemenkeu', 'tempat', 'madiun', 'wkwkwkww']	3	tni, polri, kemenkeu, pensiun, asn, naik, rapel, cair, pns, gaji	Gaji dan Kesejahteraan Pegawai
21	@user @user @user Anies di pecat karena menemukan korupsi di kemendikbud. Dilaporkan ke kemenkeu bahwa nama orang² yg di transfer ini sudah tiada. Prabowow di pecat karena ngilangin nyawa anak manusia. Loe milih mana laler warteg?	['anies', 'pecat', 'temu', 'korupsi', 'kemendikbud', 'lapor', 'kemenkeu', 'nama', 'orang', 'transfer', 'tiada', 'prabowow', 'pecat', 'ngilangin', 'nyawa', 'anak', 'manusia', 'loe', 'pilih', 'laler', 'warteg']	1	rp, menteri, anggar, gaji, uang, triliun, tidak, pajak, kemenkeu, nomor	Keuangan dan Anggaran

DISCUSSIONS

Performance of Different Preprocessing Scenarios

The comparison of six preprocessing scenarios revealed significant variations in topic modeling effectiveness. Scenario B (Stopword Removal + Sastrawi Stemming) demonstrated superior performance with a coherence score of 0.5289, despite not having the highest absolute score. This finding aligns with previous research by Rianto et al. (Rianto et al., 2021), which emphasized the importance of appropriate stemming methods for Indonesian text processing. The superiority of Sastrawi stemming over NLTK can be attributed to its specialized design for Indonesian language morphology, particularly in handling affixes and root word identification.

Topic Distribution Analysis

The highly uneven distribution of documents across topics (6,677 for Finance and Budget, 355 for Salaries and Employee Welfare, and 3 for Excise and Customs Regulations) reveals several interesting patterns:

1. **Dominant Public Interest**: The overwhelming focus on finance and budget topics (66.1% of documents) suggests that public discourse primarily centers on fiscal policy and financial management aspects of the Ministry of Finance's operations.
2. **Policy Impact**: The significant presence of salary and employee welfare-related discussions (3.5% of documents) reflects public interest in government employment policies, particularly concerning civil servant compensation.
3. **Specialized Topics**: The limited discussion of excise and customs regulations (0.03% of documents) indicates that these topics, while important, generate less public engagement on social media platforms.

Semantic Relationship Patterns

The LSA implementation revealed distinct semantic patterns in public discourse:

1. **Keyword Co-occurrence**: High-frequency keywords within the Finance and Budget topic ("kemenkeu", "pajak", "anggaran") demonstrate strong semantic relationships, indicating cohesive public discussion around fiscal policy.
2. **Cross-topic Relationships**: The appearance of certain keywords across multiple topics (e.g., "kemenkeu" in both Finance and Budget and Salaries and Employee Welfare) suggests interconnected policy discussions.
3. **Temporal Context**: The frequency distribution of keywords indicates temporal patterns in public discourse, potentially reflecting responses to specific policy announcements or implementations.

Methodological Implications

The research findings have several methodological implications:

1. **Preprocessing Impact**: The significant variation in results across different preprocessing scenarios highlights the critical importance of appropriate text preprocessing for Indonesian social media data.
2. **LSA Effectiveness**: The successful identification of coherent topics demonstrates LSA's capability in handling Indonesian social media content, despite the informal nature of tweet data.
3. **Dimensionality Reduction**: The effectiveness of TruncatedSVD in maintaining semantic relationships while reducing computational complexity supports its use in large-scale social media analysis.

Policy and Practical Implications

The research findings have several practical implications for policy makers and communication strategists:

1. **Public Communication Strategy**: The identified topic distribution can inform the Ministry of Finance's communication strategy, helping to address areas of high public interest more effectively.
2. **Policy Focus Areas**: The prevalence of certain topics in public discourse can help prioritize policy explanations and public engagement efforts.
3. **Stakeholder Engagement**: Understanding the semantic relationships between topics can improve stakeholder engagement by addressing interconnected policy concerns comprehensively.

Limitations and Future Research Directions

Several limitations of the current study suggest directions for future research:

1. **Data Temporality**: The analysis covers a specific time period, and future studies could benefit from longitudinal analysis to capture temporal changes in public discourse.
2. **Language Complexity**: While Sastrawi stemming improved results, the handling of informal Indonesian language and social media slang could be further refined.
3. **Topic Granularity**: The highly uneven topic distribution suggests that alternative topic modeling approaches might reveal more nuanced sub-topics within the dominant categories.

4. **Sentiment Integration**: Future research could benefit from integrating sentiment analysis with topic modeling to understand not just what people discuss, but also their attitudes toward different policy areas.
5. **Comparative Analysis**: Future studies could compare LSA results with other topic modeling approaches such as LDA or BERTopic for Indonesian social media content.

The findings of this study contribute to the growing body of research on social media analysis in public policy contexts, while highlighting the importance of appropriate methodological choices for processing Indonesian language content. The results provide valuable insights for both practitioners in public communication and researchers in text mining and policy analysis.

CONCLUSION

It can be concluded that this topic modeling research with scenarios using stopword removal and literary stemming resulted in 3 main topics related to the Ministry of Finance, namely Keuangan dan Anggaran (Finance and Budget), Gaji dan Kesejahteraan Pegawai (Salaries and Employee Welfare), and Regulasi Cukai dan Bea (Excise and Customs Regulations). The LSA method effectively extracted dominant topics from the tweet data. LSA can identify semantic relationships between texts in tweet data to produce keywords and topics being discussed by the public. In addition, the use of combinations in research to generate many different scenarios is also needed to support the results of the analysis and compare the quality of execution results with different conditions—for example, the use of stopword removal and stemming, which significantly affects the results of topic modeling. The language supported by the library, such as stemming, also affects the results based on the language of the tweet data used in the study. Stemming libraries with supporting languages (Indonesian) produce better results than libraries that only support English-based text mining processing. The results of this study provide in-depth information for the Ministry of Finance to understand public responses and perceptions related to policies issued by the Ministry of Finance. This research has limitations regarding the sample size of the tweets used and the possibility of bias in data selection, which only includes tweets with specific keywords.

Further research could consider using a more sophisticated model and expanding the data to cover a more extended period to be more representative. Overall, this study shows that topic modeling using LSA can provide meaningful insights into public discussions on issues related to the Ministry of Finance on social media. However, further development is needed to improve the model. Further research is also recommended to explore other models in text mining, such as Latent Dirichlet Allocation (LDA), which is more often used for topic modeling. In addition, deep learning-based approaches, such as BERT or IndoBERT, can also be used to label modeled tweet data, so that the analysis results can be more diverse and in-depth.

ACKNOWLEDGMENT

This research has been conducted in collaboration with the Bureau of Communications and Information Services of the Ministry of Finance.

REFERENCES

Ahammad, T. (2024). Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach. *Natural Language Processing Journal*, *6*, 100053. https://doi.org/10.1016/j.nlp.2024.100053

Birks, D., Coleman, A., & Jackson, D. (2020). Unsupervised identification of crime problems from police free-text data. *Crime Science*, *9*(1), 18. https://doi.org/10.1186/s40163-020-00127-4

Chen, Y., He, S., Yang, Y., & Liang, F. (2023). Learning Topic Models: Identifiability and Finite-Sample Analysis. *Journal of the American Statistical Association*, *118*(544), 2860–2875. https://doi.org/10.1080/01621459.2022.2089574

Egorova, E., Glukhov, G., & Shikov, E. (2022). Customer transactional behaviour analysis through embedding interpretation. *Procedia Computer Science*, *212*, 284–294. https://doi.org/10.1016/j.procs.2022.11.012

Finansyah, A. Y. W., Afiahayati, F., & Sutanto, V. M. (2022). Performance Comparison of Similarity Measure Algorithm as Data Preprocessing Stage: Text Normalization in Bahasa. *Scientific Journal of Informatics*, *9*(1), 1–7. https://doi.org/10.15294/sji.v9i1.30052

Hepworth, N. (2024). *Public Financial* Management *and Internal Control: The Importance of Managerial Capability for Successful Reform in Developing and Transition Economies*. Springer International Publishing. https://doi.org/10.1007/978-3-031-35066-5

Hubert, Phoenix, P., Sudaryono, R., & Suhartono, D. (2021). Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier. *Procedia Computer Science*, *179*, 498–506. https://doi.org/10.1016/j.procs.2021.01.033

Huwaidah, A., Adiwijaya, & Faraby, S. A. (2021). Argument Identification in Indonesian Tweets on the Issue of Moving the Indonesian Capital. *Procedia Computer Science*, *179*, 407–415. https://doi.org/10.1016/j.procs.2021.01.023

Huyut, M. M., Kocaoğlu, B., & Meram, Ü. (2022). Regulation Relatedness Map Creation Method with Latent Semantic Analysis. *Computers, Materials and Continua*, *72*(1), 2093–2107. https://doi.org/10.32604/cmc.2022.024190

Li, Q., Zhao, S., He, T., & Wen, J. (2024). A simple and efficient filter feature selection method via document-term matrix unitization. *Pattern Recognition Letters*, *181*, 23–29. https://doi.org/10.1016/j.patrec.2024.02.025

Morozovskii, D., & Ramanna, S. (2023). Rare words in text summarization. *Natural Language Processing Journal*, *3*, 100014. https://doi.org/10.1016/j.nlp.2023.100014

Murfi, H., Rosaline, N., & Hariadi, N. (2022). Deep autoencoder-based fuzzy c-means for topic detection. *Array*, *13*, 100124. https://doi.org/10.1016/j.array.2021.100124

Nair, R. P., & Thushara, M. G. (2024). Investigating Natural Language Techniques for Accurate Noun and Verb Extraction. *Procedia Computer Science*, *235*, 2876–2885. https://doi.org/10.1016/j.procs.2024.04.272

Nolasco, D., & Oliveira, J. (2019). Subevents detection through topic modeling in social media posts. *Future Generation Computer Systems*, *93*, 290–303. https://doi.org/10.1016/j.future.2018.09.008

Parveen, N., Chakrabarti, P., Hung, B. T., & Shaik, A. (2023). Twitter sentiment analysis using hybrid gated attention recurrent network. *Journal of Big Data*, *10*(1), 50. https://doi.org/10.1186/s40537-023-00726-3

Peng, J., Shen, D., Nie, T., & Kou, Y. (2024). RLclean: An unsupervised integrated data cleaning framework based on deep reinforcement learning. *Information Sciences*, *682*, 121281. https://doi.org/10.1016/j.ins.2024.121281

Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, *212*, 118715. https://doi.org/10.1016/j.eswa.2022.118715

Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, *8*(1), 26. https://doi.org/10.1186/s40537-021-00413-1

Sagum, R. A., Clacio, P. A. C., Cayetano, R. E. R., & Lobrio, A. D. F. (2023). Philippine Court Case Summarizer using Latent Semantic Analysis. *8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023)*, *227*, 474–481. https://doi.org/10.1016/j.procs.2023.10.548

Saheb, T., Dehghani, M., & Saheb, T. (2022). Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis. Sustainable *Computing: Informatics and Systems*, *35*, 100699. https://doi.org/10.1016/j.suscom.2022.100699

Siddhartha B S, & N. M. Niveditha, (second). (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Journal of University of Shanghai for Science and Technology*, 1–9.

Silva, C. C., Galster, M., & Gilson, F. (2021). Topic modeling in software engineering research. *Empirical Software Engineering*, *26*(6), 120. https://doi.org/10.1007/s10664-021-10026-0

Stevany, R. (2024, July 28). Indonesia Pengguna X atau Twitter Terbanyak Keempat di Dunia. *Radio Republik* Indonesia. https://rri.co.id/lain-lain/859350/indonesia-pengguna-x-atau-twitter-terbanyak-keempat-di-dunia

Viani, N., Botelle, R., Kerwin, J., Yin, L., Patel, R., Stewart, R., & Velupillai, S. (2021). A natural language processing approach for identifying temporal disease onset information from mental healthcare text. *Scientific Reports*, *11*(1), 757. https://doi.org/10.1038/s41598-020-80457-0

Wang, S., Schraagen, M., Tjong Kim Sang, E., & Dastani, M. (2020). Public Sentiment on Governmental COVID-19 Measures in Dutch Social Media. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.17