Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

LDA Topic Modeling: Twitter-Based Public Opinion on Indonesian Ministry of Finance

Dina Choirinnisa[1)], Farrikh Alzami[2)*], Heni Indrayani[3)], Asih Rohmani[4)], Siti Hadiati Nugraini[5)] Rahmawati Zulfiningrumi[6)], Fitri Susanti[7)]
[1,2,3,4,5,6)] Universitas Dian Nuswantoro, [7)] Kementerian Keuangan Indonesia
[1)] dinach.nisa@gmail.com, [2)]alzami@dsn.dinus.ac.id, [3)]heni.indrayani@dsn.dinus.ac.id, [4)]aseharsoyo@dsn.dinus.ac.id, [5)]shnugraini@dsn.dinus.ac.id, [6)] rahmawati.zulfiningrum@dsn.dinus.ac.id, [7)] fitris@kemenkeu.go.id

Submitted :April 23, 2025 | **Accepted** : May 4, 2025 | **Published** : May 7, 2025

Abstract: People in the modern era use social media daily to exchange opinions regarding government policies, such as discussions related to the Indonesian Ministry of Finance (Kemenkeu). This study aims to analyze the topics of discussion about the Ministry of Finance on the Twitter platform, now known as 'X', and to determine the results of more effective preprocessing. The data in this study was taken from Twitter using the Tweet Harvest Tool with the keyword 'Ministry of Finance' from January 2024 to July 2024. The data is then processed through cleaning, preprocessing, calculation of coherence values, LDA modeling, and visualization. The preprocessing process includes several scenarios to compare the best results that are easy for the reader to understand. The highest coherence value obtained is 0.572250 by using stemming from NLTK library. The most effective preprocessing results are normalization, tokenization, stopwords, and stemming using Sastrawi. Modeling is done to find latent topics through LDA topic modeling techniques. Visualizing the intertopic distance map provides information on the distance between each topic. The results show that the distance between one topic and another has a variety of distance variations. This study shows that social media platforms can serve as a source of evaluation for the Indonesian government. The findings of these topics are helpful as insights for readers and the Kemenkeu. Finally, the analysis identified several key topics in public discussion, including fiscal policy, budget transparency, and the Ministry of Finance's performance in addressing current economic issues.

Keywords: ***kemenkeu, topic modelling, LDA, twitter, preprocessing***

INTRODUCTION

Latent Dirichlet Allocation (LDA) has emerged as a powerful probabilistic model for discovering latent topics in large text collections (Blei et al., 2002). However, applying LDA to social media data, particularly Twitter, presents unique challenges due to the informal, short, and noisy nature of tweets. These challenges are further compounded when dealing with non-English languages like Indonesian, where standard preprocessing techniques may not be directly applicable. One of the critical challenges in LDA implementation is the preprocessing pipeline, which significantly impacts topic quality and coherence. Previous studies have shown that preprocessing steps such as tokenization, normalization, and stemming can dramatically affect the performance of topic models (Fan et al., 2021). In the context of Indonesian social media text, the choice between different stemming libraries (such as NLTK and Sastrawi) and the combination of preprocessing steps remain underexplored areas that directly influence model performance.

While LDA has been successfully applied in various domains, including crowdfunding campaigns (Muzumdar et al., 2024), e-commerce sentiment analysis (Parveen et al., 2021), and COVID-19 public opinion analysis (Nurmalasari et al., 2023), most studies focus on the application results rather than systematically evaluating the impact of different preprocessing scenarios on LDA performance. This creates a significant research gap in understanding which preprocessing combinations yield the most coherent and interpretable topics for Indonesian social media data. Furthermore, existing research often implements LDA with a single preprocessing pipeline without comparing alternative approaches. Studies by Kang et al. (Kang et al., 2019) on biomedical research trends and Lee et al. (Lee et al., 2024) on virtual education strategies demonstrate the effectiveness of LDA but do not address the preprocessing challenges specific to informal social media language. The selection of optimal

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

preprocessing steps remains largely empirical, lacking systematic evaluation using coherence metrics and other quantitative measures.

This study addresses these methodological gaps by systematically evaluating six different preprocessing scenarios for LDA topic modeling on Indonesian Twitter data. We compare the effects of various combinations of tokenization, normalization, stopword removal, and stemming (using both NLTK and Sastrawi libraries) on model coherence and topic interpretability. Our approach provides empirical evidence for selecting appropriate preprocessing pipelines when applying LDA to Indonesian social media text.

To demonstrate the practical applicability of our findings, we apply our methodology to analyze public opinion regarding the Indonesian Ministry of Finance (Kemenkeu) on Twitter. Social media platforms have become valuable sources for understanding public sentiment toward government institutions (Boulianne et al., 2024)(Studies, 2024), making this case study relevant for both methodological validation and practical policy insights.

The objectives of this study are: 1) To systematically evaluate different preprocessing scenarios for LDA topic modeling on Indonesian Twitter data; 2) To identify the most effective preprocessing pipeline based on coherence scores and topic interpretability; 3) To demonstrate the application of the optimized LDA model in analyzing public discourse about government institutions. This research contributes to the methodological understanding of LDA implementation for non-English social media analysis while providing practical insights for government institutions seeking to leverage social media analytics.

METHOD

In this study, the flow of research stages is visualized as a flow, as shown in Fig. 1.

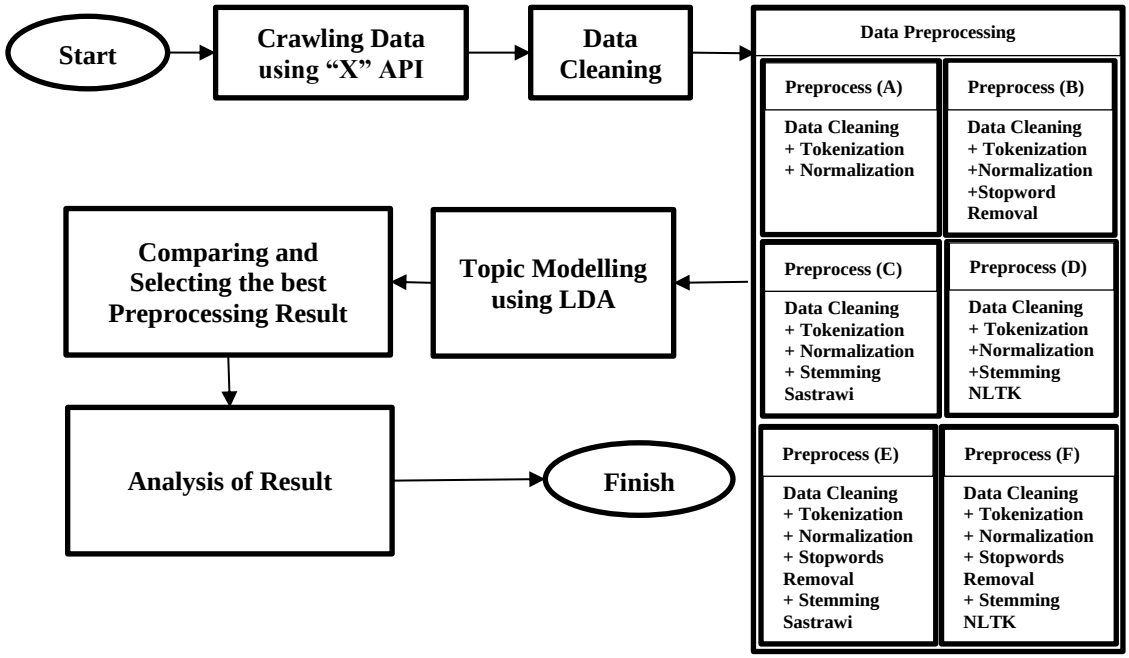

Fig. 1 Research Flow

The detailed description of Fig. 1 can be seen as follows:

Crawling Data Using the" X " API

The Data used in this study came from crawling through the platform API "X" using Tweet Harvest, designed by a Youtuber with the account name Helmi Satria. The Data was collected from January 2024 to July 2024 using the keyword "Kemenkeu." Known data generated a total of 10,099 data. The resulting data is in the form of CSV and ready to be processed. The data from crawling is processed using Google Colab with Python version 3.10.12.

Data Cleaning

To ensure unique entries and accurate topic modeling, the data is cleaned by removing duplicates, expanding abbreviations, correcting spelling, and eliminating excess spaces, non-standard ASCII characters, mentions (@username), incomplete URLs, irrelevant single characters, and punctuation (except underscores). Additionally, the words "user" and emojis are removed to reduce noise. These steps produce 7,016 clean entries, refining and

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

standardizing the dataset for optimal LDA performance. The following data-cleaning process flow can be seen in Fig. 2.

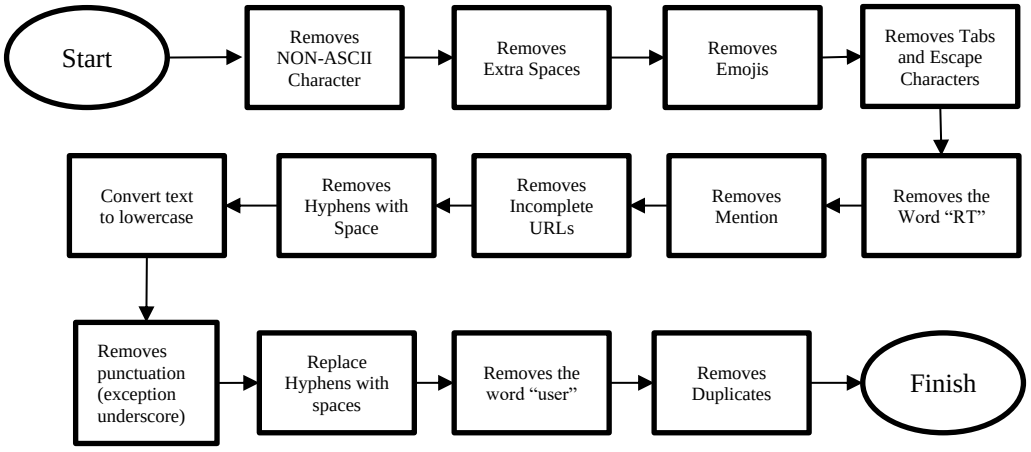

Fig. 2 Data Cleaning Flow

Preprocessing

Data preprocessing is essential to facilitate analysis, improve model accuracy, minimize noise interference, and provide relevant (Finansyah et al., 2022)data. Data Preprocessing includes the following stages:

Normalization

Normalization is turning unusual words into normal ones(Lubis et al., 2022).

Tokenization

Tokenization is a process that breaks down raw text into smaller parts called tokens (words or sentences), aiding in context understanding and building models for natural language processing(Novira, 2019).

Stopwords

Stopwords are too familiar and frequently used and do not provide significant meaning for learning(Rinandyaswara et al., 2022).

Stemming

Stemming Stemming is a natural language processing (NLP) process that reduces words to their primary or root form; this stage is widely used to improve the efficiency of Information Retrieval Systems by reducing word variations to their primary form (Abidin et al., 2024). In this process, we compare two stemming libraries, NLTK and Sastrawi. NLTK (Natural Language Toolkit) is the most popular Python library for natural language processing; it can process text quickly and easily and has many built-in Corpus covering different types of text (Fan et al., 2021). The definition of Pustaka Sastrawi is a simple Python library that can reduce inflected words in Indonesian to their primary form(Rabbani, n.d.).

The definition of Pustaka Sastrawi is a simple Python library that can reduce inflected words in Indonesian to their primary form

Topic Modelling using LDA

This study uses topic modeling LDA (Latent Dirichlet Allocation) with Gensim library. LDA is a generative probabilistic model that assumes each document is a mixture of a few topics, and each word's presence is attributable to one of the document's topics. It uses a Bayesian inference process to discover the underlying topic structure in a collection of documents. LDA is well-suited for analyzing social media content like tweets because it can uncover latent themes in large volumes of unstructured text data, allowing us to identify the main topics of discussion about Kemenkeu without predefining categories. We implemented LDA using the Gensim library in Python, which provides efficient and scalable topic modeling capabilities. We used the coherence score metric to determine the optimal number of topics. We iteratively tested different numbers of topics (from 2 to 20) and selected the number that yielded the highest coherence score, balancing between model complexity and interpretability. We set the following LDA parameters: number of topics = [optimal number], alpha = 'auto', beta

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

= 'auto', and number of iterations = 1000. These settings were chosen based on preliminary experiments and literature recommendations. We evaluated our LDA model using the coherence score (C_v measure), which assesses the semantic similarity between high-scoring words in each topic. A higher coherence score indicates more human-interpretable topics. We set a random seed (seed = 42) before running the LDA model to ensure reproducibility.
Moreover, the representation of LDA modeling can be seen in Fig. 3.

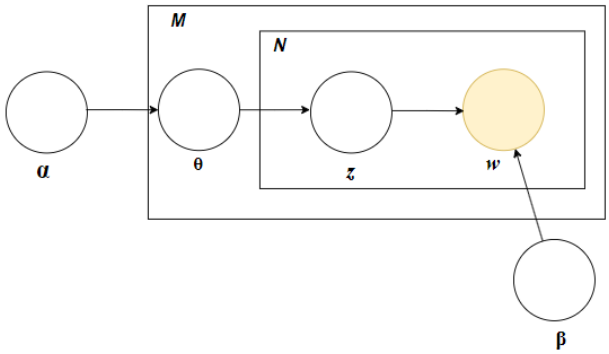

Fig. 3 Representation of LDA Modelling(Habibi, 2020)

Based on Fig. 3, it is shown that α controls the distribution of topics in the document if a high α value indicates that the document is likely to have many topics, θ indicates the distribution of topics in the document, β controls the distribution of words in each topic, and z is the latent variable that indicates the topic for each word. The working process of LDA starts with initializing the values of α and β and the distribution of topics (θ) and topics for each word (z). Then, the value of z is iteratively updated by Gibbs sampling until converging. Once inference is complete, the distribution of topics and related words can be analyzed to find the most(Blei et al., 2002) critical topics in the document.

Comparing and Selecting the best preprocessing result

Preprocessing scenarios are modeled using LDA modeling to compare the results of each scenario. The results of a quality LDA model topic have the right, meaningful, and easy-to-understand topic. Then, the data must go through an optimal preprocessing process, such as loading and cleaning the text by removing words that have no semantic value, converting all words to lowercase letters, and removing spaces and punctuation so that the resulting topic is relevant(Asmussen & Møller, 2019).

Analysis of Result

Result analysis involves systematically evaluating the results of the best LDA model to determine latent topics in the discussion of 'Kemenkeu'. This process includes analyzing the data generated from the best modeling results and creating visualizations to make it easier for readers. By examining patterns and trends in outcomes, researchers can draw meaningful conclusions about the strengths and weaknesses of each scenario. This analysis helps analyze the model for each scenario, make necessary adjustments, and ensure that the results are reliable and can serve as information, insight, or advice to the reader.

RESULT

In this study, the results and discussions will be presented in the form of sub-chapters as follows:

Data Cleaning

The data cleaning process aims to produce clean data that can be processed with several preprocessing scenarios.The data cleansing flow is shown in Fig. 1. Table 1 shows the difference between raw data and cleaned data.

Table 1. Data Cleaning Result

Raw Data	Data Clean
@R****W******19 @k****h*b151 @K****p******f Pesan diterima @k***nh**151 @k**n*** @p******g****n02	pesan diterima
@p*****n****ij**** @H**R****P*** @pr*** Kalau pintar apalagi S2 luar negeri mending cari kerja di Singapore atau	kalau pintar apalagi s2 luar negeri mending cari kerja di singapore atau hong kong aja pns non kemenkeu hidupnya melarat

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Hong Kong aja. PNS non kemenkeu hidupnya melarat	
Kemenkeu Pastikan Gaji PNS Naik 8 Persen per 1 Januari Tapi Dirapel https://t.co/EKS4Wk9LpP	kemenkeu pastikan gaji pns naik persen per januari tapi dirapel

Preprocessing Process

In the preprocessing stage, there are several scenarios; this is done to find out and compare the best results from preprocess (a), preprocess (B), preprocess (C), preprocess (D), preprocess (E), and preprocess (F). Preprocessing results from several scenarios as follows:

Preprocess (A)

In the preprocessing scenario (A), the steps are to process the tokenized and normalized cleaning data—tokenization using NLTK libraries and normalization of tokens based on dictionaries. The results of preprocessing (A) can be seen in Table 2.

Table 2. Result Preprocess (A)

Data Clean	Token	Normalization
pesan diterima	pesan,diterima	pesan,diterima
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,mending,cari, kerja,di,singapore,atau, hong,kong,aja,pns,non, kemenkeu,hidupnya, melarat	kalau,pintar,apalagi,s2,luar,negeri, lebihbaik,cari,kerja,di,singapore, atau,hong,kong,saja,pns,non, kemenkeu,hidupnya,melarat
kemenkeu pastikan gaji pns naik persen per januari tapi dirapel	kemenkeu,pastikan,gaji, pns,naik,persen,per,januari ,tapi,dirapel	kemenkeu,pastikan,gaji,pns,naik, persen,per,januari, tapi,dirapel

Preprocess (B)

The preprocess Scenario (B) is almost the same as preprocess (A), but the preprocess stage (B) is added to the stopword removal process. Table 1 shows the tokenization and normalization process. Table 3 shows the results of preprocess (B).

Table 3. Result of Preprocess (B)

Data Clean	Stopwords Removal
pesan diterima	pesan,diterima
kalau pintar apalagi s2 luar negeri mending cari kerja di singapore atau hong kong aja pns non kemenkeu hidupnya melarat	pintar,s2,negeri,lebih baik,cari,kerja, singapore,hong,kong,pns,non, kemenkeu,hidupnya,melarat
kemenkeu pastikan gaji pns naik persen per januari tapi dirapel	kemenkeu,pastikan,gaji,pns, persen,januari,dirapel

Preprocess (C)

In the preprocess Scenario (C), add the steaming process. At this stage, the steaming process is done using a literary library. The results of preprocess (C) before stemming, i.e. tokenization and normalization, can be seen in Table 1. The results of preprocess (C) can be seen in Table 4.

Table 4. Result of Preprocess (C)

Data Clean	Stemming Sastrawi
pesan diterima	pesan,terima
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,hidup, melarat
kemenkeu pastikan gaji pns naik persen per januari tapi dirapel	kemenkeu,pasti,gaji,pns,naik,persen,per,januari,tapi ,rapel

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Preprocess (D)

Preprocess (D) is similar to preprocess (C). Here, the stemming process uses the NLTK library. In preprocess (D), net data and NLTK stemming results are displayed, as can be seen in Table 5.

Table 5. Result of Preprocess (D)

Data Clean	Stemming NLTK
pesan diterima	pesan,diterima
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, singapor,atau,hong,kong,saja,pn,non,kemenkeu,hidupnya,melarat
kemenkeu pastikan gaji pns naik persen per januari tapi dirapel	kemenkeu,pastikan,gaji, pn,naik,persen,per, januari,tapi,dirapel

Preprocess (E)

The preprocess scenario (E) combines all the stages of preprocessing, starting with data clean, tokenization, normalization, stopword removal, and stemming. At this stage, the stemming process uses literary literature. The preprocessing results that including data clean, tokenization, and normalization can be seen in Table 1. The results of stopword removal and stemming using literature can be seen in Table 6.

Table 6. Result of Preprocess (E)

Stopwords	Stemming Sastrawi
pesan,diterima	pesan,terima
pintar,s2,negeri,lebihbaik,cari,kerja,singapore, hong,kong,pns,non,kemenkeu,hidupnya,melarat	pintar,s2,negeri,lebihbaik,cari,kerja,singapore,hong ,kong,pns,non,kemenkeu,hidup,melarat
kemenkeu,pastikan,gaji,pns,persen,januari,dirapel	kemenkeu,pasti,gaji,pns,persen,januari,rapel

Preprocess (F)

Preprocess (F) is similar to preprocess (E). However, in preprocess scenario (F), The stemming process uses the NLTK library. The preprocess (F) results only display the stemming results using the NLTK library, as seen in Table 7.

Table 7. Result of Preprocess (F)

Data Clean	Stemming NLTK
pesan diterima	pesan,diterima
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,singapor,atau,hong,kong,saja,pn, non,kemenkeu,hidupnya, melarat
kemenkeu pastikan gaji pns naik persen per januari tapi dirapel	kemenkeu,pastikan,gaji, pn,naik,persen,per,januari, tapi,dirapel

Comparing and Selecting the Best Data

Data To evaluate its effectiveness, the preprocessing process involves several scenarios that are compared to each other. The coherence value, which measures the interrelationships between words in the resulting topic, will be used to analyze each preprocessing stage. In addition, the results of topic modeling using LDA are assessed to determine how well the model identifies the underlying structure of the data. How clear and understandable the resulting topic is will determine the quality of the results, which provides valuable insights for further analysis

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

data. The resulting topic's clarity and understanding will determine the results' quality, which will provide valuable insights for further analysis.

Coherence Value

Preprocessing is done with various optimized scenarios to get the best results. The coherence value and the ideal number of topics of each tested scenario will be calculated. The coherence value indicates how well the words in a topic are interconnected and the number of topics produced. The results of this calculation will show how effective the LDA model is in capturing the underlying structure of the data. The coherence values and the optimal number of topics of each tested scenario are described in detail in Table 8.

Table 8. Coherence Score and Number of Topics

Preprocessing Scenario	Coherence Score	Topic
Preprocess (A) (Tokenization + Normalization)	0.553403	4
Preprocess (B) (Tokenization + Normalization +Stopwords Removal)	0.392844	13
Preprocess (C) (Tokenization +Normalization +Stemming Sastrawi)	0.555225	3
Preprocess (D) (Tokenization +Normalization +Stemming NLTK)	0.572250	2
Preprocess (E) (Tokenization +Normalization +Stopwords Removal+Stemming Sastrawi)	0.329886	13
Preprocess (F) (Tokenization +Normalization +Stopwords Removal + Stemming NLTK)	0.572250	2

From Table 8. Preprocessing scenarios D and F yielded the highest coherence score of 0.572250. This can be attributed to the effectiveness of the NLTK stemming process in these scenarios, which better preserved the semantic relationships between words while reducing vocabulary size. The consistency in results between these two scenarios suggests that the addition of stopword removal in scenario F did not significantly impact the coherence, indicating that the NLTK stemming process was robust in maintaining topic coherence. The varying optimal number of topics across preprocessing scenarios provides insights into the impact of different text-processing techniques on topic granularity. Scenarios B and E, which involved stopword removal, resulted in more optimal topics (13), suggesting that removing common words allowed for more nuanced topic differentiation. Conversely, scenarios D and F, which achieved the highest coherence, identified only 2 optimal topics. This could indicate that these scenarios resulted in more generalized, broader topics encapsulating multiple related sub-themes. The trade-off between coherence and topic granularity highlights the importance of carefully considering preprocessing steps about the desired level of topic specificity.

Examining the coherence scores across scenarios reveals that stemming (whether using NLTK or Sastrawi) generally improved coherence compared to scenarios without stemming. However, the addition of stopword removal (scenarios B and E) decreased coherence scores. This suggests that while stemming helps consolidate related words and improve topic coherence, aggressive word removal through stopword filtering may eliminate some contextually important terms, potentially reducing the overall coherence of the identified topics.

These findings emphasize the critical role of preprocessing in LDA topic modeling. While higher coherence scores generally indicate more interpretable topics, researchers must balance this with the desired level of topic granularity. For analyzing public opinion on Kemenkeu, the choice between a model with fewer, more coherent topics (as in scenarios D and F) versus one with more numerous, potentially more specific topics (as in scenarios B and E) would depend on the specific analytical goals and the level of detail required for policy insights.

Fig. 4 shows the following visualization of a grouped bar chart based on LDA modeling results from the preprocessing scenario

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

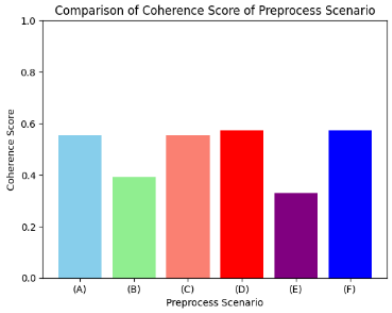

Fig. 4 Visualization of Coherence Score

Topic Modelling LDA

Result In this process, the LDA used is LDA with version 4.3.3. The preprocessing scenarios implemented are used to identify and analyze problems arising from the data. Each scenario will produce different results, so assessing those differences is essential. Table 9 will present the results of the LDA modeling in detail, including information on the topics found and the coherence values per word. LDA modeling results can be seen in Table 9 to provide a clear picture of the topic analysis results obtained from various preprocessing scenarios.

Table 9. Results of weighting each word of the LDA model

Preprocessing scenario	Result of Modelling LDA
Preprocess (A) (Tokenization + Normalization)	**Topic: 0 Word**: 0.057*"kemenkeu" + 0.041*"di" + 0.023*"ada" + 0.021*"kalau" + 0.018*"sudah" + 0.016*"saja" + 0.014*"ke" + 0.010*"sampai" + 0.009*"dan" + 0.009*"yang" **Topic: 1 Word**: 0.040*"dan" + 0.038*"kemenkeu" + 0.020*"untuk" + 0.012*"di" + 0.012*"pajak" + 0.011*"yang" + 0.011*"dalam" + 0.009*"oleh" + 0.008*"indonesia" + 0.008*"ekonomi" **Topic: 2 Word**: 0.025*"kemenkeu" + 0.023*"dengan" + 0.013*"atau" + 0.013*"masih" + 0.011*"akan" + 0.010*"pemerintah" + 0.009*"tahun" + 0.008*"soal" + 0.008*"gaji" + 0.007*"uang_kita" **Topic: 3 Word**: 0.047*"kemenkeu" + 0.043*"yang" + 0.025*"dari" + 0.023*"tidak" + 0.020*"itu" + 0.019*"ini" + 0.016*"ya" + 0.012*"juga" + 0.012*"dan" + 0.011*"tapi"
Preprocess (B) (Tokenization + Normalization + Stopwords Removal)	**Topic: 0 Word**: 0.043*"bea_masuk" + 0.032*"eselon" + 0.030*"bisnis_pembaruan_pembaruan_bisnis" + 0.024*"te" + 0.024*"perekonomian" + 0.021*"menteri_keuangan_sri_mulyani" + 0.020*"kinerja_apbn" + 0.019*"viral" + 0.018*"kemenkeu_membuka_suara" + 0.018*"china" **Topic: 1 Word**: 0.000*"bertanggung" + 0.000*"daftar_magang" + 0.000*"lkpp" + 0.000*"panas_bumi" + 0.000*"geo_dipa" + 0.000*"gencarkan_eksplorasi" + 0.000*"tingkat_kemiskinan" + 0.000*"bansos_ekonomi_domestik_turunkan" + 0.000*"rabu_kurs_pajak_terbaru" + 0.000*"tanggal_berlaku" **Topic: 2 Word**: 0.000*"bertanggung" + 0.000*"daftar_magang" + 0.000*"lkpp" + 0.000*"panas_bumi" + 0.000*"geo_dipa" + 0.000*"gencarkan_eksplorasi" + 0.000*"tingkat_kemiskinan" + 0.000*"bansos_ekonomi_domestik_turunkan" + 0.000*"rabu_kurs_pajak_terbaru" + 0.000*"tanggal_berlaku" **Topic: 3 Word**: 0.082*"pakai" + 0.030*"rp_triliun" + 0.030*"juni" + 0.028*"kemenkeu_pastikan" + 0.020*"aturan_berlaku" + 0.020*"iuran_tapera_dilaksanakan_sesuai" + 0.018*"pembiayaan" + 0.018*"saham" + 0.018*"tepercaya" + 0.018*"investor_daily_invest_as" **Topic: 4 Word**: 0.000*"bertanggung" + 0.000*"daftar_magang" + 0.000*"lkpp" + 0.000*"panas_bumi" + 0.000*"geo_dipa" + 0.000*"gencarkan_eksplorasi" + 0.000*"tingkat_kemiskinan" + 0.000*"bansos_ekonomi_domestik_turunkan" + 0.000*"rabu_kurs_pajak_terbaru" + 0.000*"tanggal_berlaku" **Topic: 5 Word**: 0.000*"bertanggung" + 0.000*"daftar_magang" + 0.000*"lkpp" + 0.000*"panas_bumi" + 0.000*"geo_dipa" + 0.000*"gencarkan_eksplorasi" + 0.000*"tingkat_kemiskinan" + 0.000*"bansos_ekonomi_domestik_turunkan" + 0.000*"rabu_kurs_pajak_terbaru" + 0.000*"tanggal_berlaku" **Topic: 6 Word**: 0.183*"kemenkeu" + 0.040*"tidak" + 0.024*"ya" + 0.014*"uang" + 0.011*"pajak" + 0.010*"negara" + 0.008*"orang" + 0.008*"nya" + 0.008*"pemerintah" + 0.007*"apbn" **Topic: 7 Word**: 0.045*"dana" + 0.043*"masuk" + 0.041*"tolong" + 0.040*"teman" + 0.036*"langsung" + 0.035*"bilang" + 0.024*"as" + 0.023*"rapat" + 0.021*"tim" + 0.020*"mendukung" **Topic: 8 Word:** 0.000*"bertanggung" + 0.000*"daftar_magang" + 0.000*"lkpp" + 0.000*"panas_bumi" + 0.000*"geo_dipa" + 0.000*"gencarkan_eksplorasi" + 0.000*"tingkat_kemiskinan" + 0.000*"bansos_ekonomi_domestik_turunkan" + 0.000*"rabu_kurs_pajak_terbaru" + 0.000*"tanggal_berlaku"

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Preprocessing scenario	Result of Modelling LDA
	Topic: 9 Word: 0.045*"terjaga" + 0.036*"kementerian_keuangan_kemenkeu" + 0.032*"yaa" + 0.027*"masyarakat" + 0.024*"simak" + 0.023*"juli" + 0.022*"miliar" + 0.022*"yuk" + 0.019*"umkm" + 0.018*"kppn" **Topic: 10 Word**: 0.069*"indonesia" + 0.049*"asn" + 0.029*"pertumbuhan_ekonomi" + 0.024*"menke" + 0.023*"jalan" + 0.021*"hasil" + 0.019*"belum" + 0.019*"kali" + 0.018*"mencapai" + 0.017*"RAPBN (Rencana Anggaran Pendapatan dan Belanja Negara)" **Topic: 11 Word**: 0.000*"bertanggung" + 0.000*"daftar_magang" + 0.000*"lkpp" + 0.000*"panas_bumi" + 0.000*"geo_dipa" + 0.000*"gencarkan_eksplorasi" + 0.000*"tingkat_kemiskinan" + 0.000*"bansos_ekonomi_domestik_turunkan" + 0.000*"rabu_kurs_pajak_terbaru" + 0.000*"tanggal_berlaku" **Topic: 12 Word**: 0.090*"negeri" + 0.064*"pns" + 0.062*"kerja" + 0.038*"kaya" + 0.037*"aturan" + 0.027*"juta" + 0.023*"pendidikan" + 0.018*"berdasarkan" + 0.017*"becuk" + 0.017*"bank"
Preprocess (C) (Tokenization + Normalization + Stemming Sastrawi)	**Topic: 0 Word**: 0.039*"kemenkeu" + 0.036*"yang" + 0.020*"tidak" + 0.018*"ada" + 0.015*"di" + 0.015*"dan" + 0.014*"itu" + 0.014*"ini" + 0.013*"dari" + 0.013*"ya" **Topic: 1 Word**: 0.039*"kemenkeu" + 0.027*"dan" + 0.019*"untuk" + 0.009*"yang" + 0.008*"dari" + 0.008*"dalam" + 0.008*"satu" + 0.007*"ekonomi" + 0.007*"indonesia" + 0.007*"menteri_uang" **Topic: 2 Word**: 0.064*"kemenkeu" + 0.020*"sama" + 0.020*"di" + 0.016*"ke" + 0.009*"kerja" + 0.008*"dulu" + 0.007*"gaji" + 0.006*"asn" + 0.006*"aku" + 0.006*"nih"
Preprocess (D) (Tokenization + Normalization + Stemming NLTK)	**Topic: 0 Word**: 0.046*"kemenkeu" + 0.027*"yang" + 0.015*"di" + 0.015*"tidak" + 0.012*"ada" + 0.011*"itu" + 0.010*"dan" + 0.010*"ini" + 0.009*"kalau" + 0.009*"ya" **Topic: 1 Word**: 0.036*"kemenkeu" + 0.021*"dan" + 0.013*"untuk" + 0.009*"di" + 0.008*"dari" + 0.007*"yang" + 0.006*"dalam" + 0.005*"dengan" + 0.004*"oleh" + 0.004*"indonesia"
Preprocess (E) (Tokenization + Normalization + Stopwords Removal+ Stemming Sastrawi)	**Topic: 0 Word:** 0.100*"kemenkeu" + 0.036*"anggar" + 0.035*"pajak" + 0.027*"orang" + 0.025*"nya" + 0.024*"menteri" + 0.023*"negara" + 0.023*"pakai" + 0.022*"lapor" + 0.022*"layan" **Topic: 1 Word**: 0.081*"data" + 0.072*"gaji" + 0.058*"pegawai" + 0.052*"anak" + 0.051*"ri" + 0.035*"triliun" + 0.028*"aman" + 0.026*"miliar" + 0.024*"tukin" + 0.019*"youtube" **Topic: 2 Word**: 0.058*"asn" + 0.049*"laksana" + 0.038*"bi" + 0.036*"atur" + 0.034*"tumbuh_ekonomi" + 0.033*"belum" + 0.029*"turun" + 0.028*"ikn" + 0.023*"proses" + 0.021*"eselon" **Topic: 3 Word**: 0.148*"kemenkeu" + 0.086*"tidak" + 0.052*"ya" + 0.021*"perintah" + 0.016*"ekonomi" + 0.012*"pas" + 0.011*"jabat" + 0.010*"bukan" + 0.010*"masuk" + 0.010*"tri" **Topic: 4 Word**: 0.000*"daftar_magang" + 0.000*"bansos_ekonomi_domestik_turun" + 0.000*"tingkat_miskin" + 0.000*"rupiah_anjlok" + 0.000*"subsidi_energi" + 0.000*"rp37" + 0.000*"antisipasi" + 0.000*"panas_bumi" + 0.000*"kemenperin_kemendag" + 0.000*"gencar_eksplorasi" **Topic: 5 Word:** 0.090*"menteri_uang" + 0.055*"nih" + 0.039*"hasil" + 0.038*"djp" + 0.031*"kait" + 0.026*"simak" + 0.023*"bijak" + 0.020*"resmi" + 0.019*"daftar" + 0.017*"it" **Topic: 6 Word**: 0.000*"daftar_magang" + 0.000*"bansos_ekonomi_domestik_turun" + 0.000*"tingkat_miskin" + 0.000*"rupiah_anjlok" + 0.000*"subsidi_energi" + 0.000*"rp37" + 0.000*"antisipasi" + 0.000*"panas_bumi" + 0.000*"kemenperin_kemendag" + 0.000*"gencar_eksplorasi" **Topic: 7 Word**: 0.000*"daftar_magang" + 0.000*"bansos_ekonomi_domestik_turun" + 0.000*"tingkat_miskin" + 0.000*"rupiah_anjlok" + 0.000*"subsidi_energi" + 0.000*"rp37" + 0.000*"antisipasi" + 0.000*"panas_bumi" + 0.000*"kemenperin_kemendag" + 0.000*"gencar_eksplorasi" **Topic: 8 Word**: 0.113*"lpdp" + 0.066*"terima" + 0.057*"kemenag" + 0.025*"korup" + 0.021*"s2" + 0.018*"engga" + 0.006*"asal" + 0.001*"agama" + 0.000*"barang_hibah" + 0.000*"link" **Topic: 9 Word**: 0.225*"kemenkeu" + 0.028*"kerja" + 0.018*"negeri" + 0.016*"biaya" + 0.014*"bumn" + 0.013*"semester" + 0.013*"jalan" + 0.013*"pns" + 0.013*"giat" + 0.011*"sosialisasi" **Topic: 10 Word**: 0.000*"daftar_magang" + 0.000*"bansos_ekonomi_domestik_turun" + 0.000*"tingkat_miskin" + 0.000*"rupiah_anjlok" + 0.000*"subsidi_energi" + 0.000*"rp37" + 0.000*"antisipasi" + 0.000*"panas_bumi" + 0.000*"kemenperin_kemendag" + 0.000*"gencar_eksplorasi" **Topic: 11 Word**: 0.000*"daftar_magang" + 0.000*"bansos_ekonomi_domestik_turun" + 0.000*"tingkat_miskin" + 0.000*"rupiah_anjlok" + 0.000*"subsidi_energi" + 0.000*"rp37" + 0.000*"antisipasi" + 0.000*"panas_bumi" + 0.000*"kemenperin_kemendag" + 0.000*"gencar_eksplorasi" **Topic: 12 Word**:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Preprocessing scenario	Result of Modelling LDA
	0.094*"uang" + 0.047*"indonesia" + 0.043*"apbn" + 0.038*"jaga" + 0.022*"lanjut" + 0.018*"yaa" + 0.017*"serta" + 0.017*"dukung" + 0.017*"hadir" + 0.016*"baca"
Preprocess (F) (Tokenization + Normalization + Stopwords Removal+ Stemming NLTK)	**Topic: 0 Word**: 0.036*"kemenkeu" + 0.021*"dan" + 0.013*"untuk" + 0.009*"di" + 0.008*"dari" + 0.007*"yang" + 0.006*"dalam" + 0.005*"dengan" + 0.004*"oleh" + 0.004*"indonesia" **Topic: 1 Word**: 0.046*"kemenkeu" + 0.027*"yang" + 0.015*"di" + 0.015*"tidak" + 0.012*"ada" + 0.011*"itu" + 0.010*"dan" + 0.010*"ini" + 0.009*"kalau" + 0.009*"ya"

At Table 9. The occurrence of topics with zero values, particularly in scenarios B and E, warrants careful interpretation. These zero-value topics could result from over-segmentation due to aggressive preprocessing, particularly stopword removal. For example, in scenario B, topics 1, 2, 4, 5, 8, and 11 show zero values for all words. This suggests that the preprocessing in this scenario removed too many words, leaving these topics without significant content. The implications of these zero-value topics are twofold: firstly, they indicate that the chosen number of topics might be too high for the given preprocessing scenario, and secondly, they suggest that the preprocessing steps may need to be adjusted to preserve more meaningful content.

DISCUSSIONS

Analysis Results

Based on the processing results in this study, it is useful to find latent topics in comments on the Twitter platform or 'X'. It is known that in each preprocessing scenario processed using LDA modeling produces different results. It can be seen in Table 8 that the coherence results in Scenario (D) and Scenario (F) are the same and highest among other scenarios. Table 9 presents the results of coherence values per topic of each scenario. To choose the best results, each scenario outcome must be analyzed. Analysis of the scenario results as follows:

Preprocess (A)
Four topics emerge from the preprocess scenario (A). Within these four topics, non-semantic words such as ’yang (which),' 'dari (of),' 'itu (that),' and ’juga (also)’ remain. These words can cause the modeling results to be inaccurate and obscure the topic's relevance.

Preprocess (B)
On this scenario appeared 13 topics. Each topic has brought up important topics such as ’kemenkeu’,’pajak(taxes)’,’kinerja_apbn’. However, in this scenario there are as many as 4 topics that have a value of 0, this indicates inaccuracies in extracting keywords.

Preprocess (C)
The result of modeling with preprocess Scenario (C) raises 3 topics. In the results, there are meaningless words, similar to those in the preprocess scenario (A). However, in this scenario has brought up relevant topics such as ’kemenkeu’, ’gaji (salary)’, ’asn’.

Preprocess (D)
The modeling results of this scenario are the same as the preprocess scenarios that do not use stopwords, giving rise to meaningless words.

Preprocess (E)
The results of this scenario raise 13 topics. This topic only brings up a few words that are not semantic. The topics in this scenario are pretty relevant because there are dominant topics such as ’kemenkeu’, ’pajak(taxes)’, ’gaji(salary)’. However, there is still a topic value of 0.

Preprocess (F)
The result of scenario (F) uses the stopword process, but the word does not semantically appear anyway. Dominant's topic on this scenario is 'kemenkeu’.

From the analysis of each scenario, the best results are the results of preprocess scenarios (E). The results of preprocess scenario (E) contain dominant topics that are the focus of this study; the resulting topics are minimal, raising the word is not semantic. Compared to scenarios that produce similar results in Scenario (B), scenario (E) produces more understandable results and focus on topics around the ‘kemenkeu’. To provide a clear picture

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

related to the most frequently occurring words and facilitate the reader, LDA modeling results will be presented using worldcloud visualization, which can be seen in Fig. 5.

Fig. 5 World Cloud of Top 10 Dominant World for Each Topic

The word cloud visualization makes it easy to read each dominant topic. It displays the 10 dominant words that emerge from each topic. The dominant word appears in a larger form than the non-dominant word.
Based on the weighting of values on each topic that has been done, as seen in Table 9, a map visualization of inter-topic distances is made to see how close or far the distance between topics is, which helps identify similar or overlapping topics. Overall, these visualizations help interpret the results and ensure that topic modeling generates separate and complementary topics. Figure 6 shows a visualization of distance maps between topics.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14719
e-ISSN : 2541-2019
p-ISSN : 2541-044X

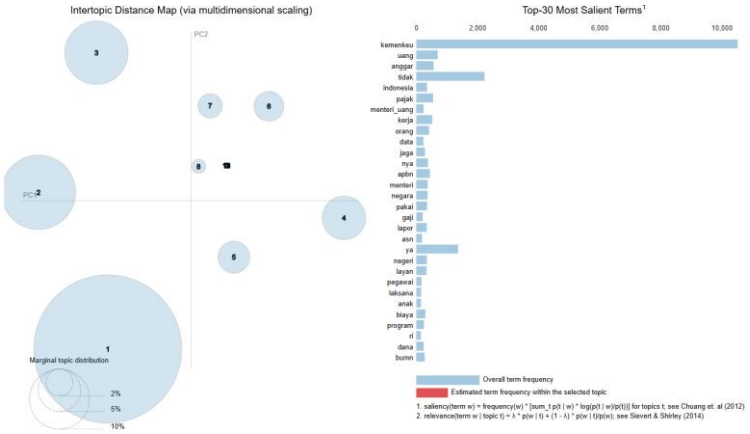

Fig. 6 Intertopic Distance Map Visualization

In the visualization of the distance map between topics, it can be seen that there are two panels, namely, the left panel and the right panel. On the left panel, a circle of topics represents the distance between topics. Topic 7 and topic 6 are closely spaced, which suggests that the two topics have something in common. In addition, overlapping topics also have similar words in them, such as topic 9 to topic 13. The resulting visualization illustrates that Topic 1 is a dominant topic compared to other topics because the circle size is more significant. It can be seen in Figure 6 that there are 30 dominant topics, such as 'kemenkeu', 'uang (money)', 'pajak (taxes)', and 'anggar (anggar)'. This dominant topic can be used as an in-depth view or primary information for the Kemenkeu. To make it easier to determine the discussion contained in each topic, it is done by manually labeling the topic by looking at the content of words on each topic, which can be seen in Figure 7.

```
# Based on the output above, we can assign topic labels as follows
topics_keywords = {
    1: "Keuangan Negara",     # Topic 1 label based on the words that appear
    2: "Gaji Pegawai",        # Topic 2 label based on the words that appear
    3: "Regulasi",            # Topic 3 label based on the words that appear
    4: "Ekonomi",             # Topic 4 label based on the words that appear
    5: "Bantuan Sosial",      # Topic 5 label based on the words that appear
    6: "Kebijakan Keuangan",  # Topic 6 label based on the words that appear
    7: "Bantuan Sosial",      # Topic 7 label based on the words that appear
    8: "Bantuan Sosial",      # Topic 8 label based on the words that appear
    9: "Isu Korupsi",         # Topic 9 label based on the words that appear
    10: "Kebijakan PNS",      # Topic 10 label based on the words that appear
    11: "Bantuan Sosial",     # Topic 11 label based on the words that appear
    12: "Bantuan Sosial",     # Topic 12 label based on the words that appear
    13: "APBN"                # Topic 13 label based on the words that appear
    }
```

Fig. 7 Manual Labelling

It can be seen in Figure 7; Topic 1 is labeled 'Keuangan Negara (state finance)' because each word describes the state's finances, while Topic 2 is labeled 'gaji pegawai (employee salaries)' because the dominant word is salary. In Topic 3 it is labeled 'regulasi (regulation)' and in topic 4 it is labeled 'Ekonomi (Economics)'. Topics 5, 7, 8, 11, and 12 are labeled 'bantuan sosial (social assistance)' because they have the same word value and do not have a dominant word but there are words about work, natural disasters, and the rupiah plummeting. Topic 6 is labeled 'Kebijakan Keuangan (financial policy)' because the word in the topic describes the word 'resmi (official)', 'djp'. Topic 9 is labeled 'isu korupsi (corruption issues)' because it contains the words 'corrupt (korupsi)' and 'barang_hibah'. On topic 10 is labeled 'Kebijakan PNS (civil servant Policy)' and topic 13 is labeled' APBN' because there are dominant words, namely 'uang (money)' and apbn.
Table 9 contains word weighting values using LDA for preprocess (E). The word weighting results will be explained per topic in Table 10. This is done to provide Kemenkeu with a clear understanding of the topic discussed.

Table 10. Analysis of Result

Index	Topic	Result
0	Keuangan Negara (State Finance)	Related to the performance of the Ministry of Finance in managing and allocating state finances
1	Gaji Pegawai (Office Salary)	There is a linkage of agency employee salaries to the Ministry of Finance in the form of employee salary distribution and salary policy

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

2	Regulasi (Regulation)	Regulations relating to state financial policy are the responsibility of the Ministry of Finance. Regulations are highlighted by the public, such as tax and customs regulations
3	Ekonomi (Economy)	The Ministry of Finance plays a vital role in the country's economy. This includes economic matters in Indonesia, such as state Budget Management, Debt Policy, and post-pandemic economic recovery.
4,6,7,10, 11	Bantuan Sosial (Social Assistance)	This social assistance discusses budgeting funds to the community, scholarships, and distribution to social institutions.
5	Kebijakan Keuangan (Financial Policy)	Twitter users' discussions of financial policy provide insight into the close relationship between MOF decisions and financial policy perceived by the public.
8	Isu Korupsi (Corruption Issues)	Discussion of corruption issues by Twitter users explained the picture of receiving corrupt funds and gratuities for officials
9	Kebijakan PNS (PNS Policy)	Civil servant policy towards the Ministry of Finance highlighted by Twitter users related to salary budgeting, civil servant welfare, and civil servant benefits
12	APBN	APBN Indonesia includes the preparation of the APBN, its implementation, and management. The Kemenkeu highlighted related to the management of the APBN, where every decision of the APBN is the responsibility of the Kemenkeu

Based on the analysis of the topics generated from the LDA model, it is known that the topics produced are very closely related to the responsibilities, functions, and morality of the Ministry of Finance towards Indonesian state finances. This study proves that the people of Indonesia are still concerned about the performance of the Ministry of Finance. This information and insight can be used to evaluate the Ministry of Finance's current performance. These topics prove that the Twitter platform is still relevant in serving as input, advice, and criticism for agencies in any sector.

Moreover, the inter-topic distance map (Figure 6) reveals important insights about topic relationships. Topics 7 and 6 show close proximity, suggesting thematic overlap in discussions about "Bantuan Sosial" (Social Assistance) and "Kebijakan Keuangan" (Financial Policy). In contrast, Topics 1 and 13 maintain significant distance despite both discussing financial matters, indicating distinct contextual discussions - state finance management versus APBN specifics. The varying circle sizes, with Topic 1 being notably larger, confirm its dominance in public discourse about Kemenkeu. This visualization validates that our LDA model successfully captures distinct yet related themes in public opinion.

Despite scenarios D and F achieving the highest coherence score (0.572250), we selected scenario E as the optimal preprocessing pipeline for several reasons: (1) Topic Interpretability: Scenario E produced 13 topics with clear thematic distinctions, while scenarios D and F only generated 2 topics, potentially oversimplifying the discourse. (2) Semantic Clarity: Scenario E minimized non-semantic words while preserving meaningful terms like 'kemenkeu', 'pajak', and 'gaji', essential for understanding public concerns. (3) Balanced Granularity: The 13 topics in scenario E provided sufficient granularity to capture diverse aspects of public opinion without excessive fragmentation. (4) Practical Relevance: The topics identified in scenario E (Table 10) directly aligned with Kemenkeu's key responsibilities and public concerns, making the results more actionable for policy insights. While some topics showed zero values, this trade-off was acceptable given the improved interpretability and practical applicability of the results.

We employed coherence score (C_v measure) as our primary validation metric due to its strong correlation with human judgment of topic quality. The coherence scores across our six preprocessing scenarios ranged from 0.329886 to 0.572250, demonstrating significant variation based on preprocessing choices. Our validation approach included: - Systematic comparison of coherence scores across all scenarios (Table 8) - Analysis of the relationship between coherence and topic number - Evaluation of word co-occurrence patterns within topics The high coherence score of 0.572250 achieved by scenarios D and F validates the effectiveness of NLTK stemming, while the moderate score of 0.329886 in scenario E was offset by superior topic interpretability and practical relevance.

This study makes several methodological contributions to LDA implementation for non-English social media analysis: (1) Systematic Preprocessing Evaluation: We provide empirical evidence comparing six different preprocessing pipelines, demonstrating that the choice of stemming library (NLTK vs. Sastrawi) significantly impacts model performance for Indonesian text. (2) Trade-off Analysis Framework: We establish a framework for balancing coherence scores with topic interpretability, showing that the highest coherence score does not always yield the most useful results for practical applications. (3) Optimal Pipeline for Indonesian Twitter Data: Our

*name of corresponding author

findings indicate that combining tokenization, normalization, stopword removal, and Sastrawi stemming (scenario E) provides the best balance for Indonesian social media text analysis. (4) Reproducible Methodology: We present a systematic approach for evaluating preprocessing scenarios that can be applied to other Indonesian language datasets or government institutions. These contributions address the gap in preprocessing optimization for LDA on Indonesian social media data, providing researchers with evidence-based guidelines for future studies.

CONCLUSION

Based on the research results, Twitter data with the keyword 'Kemenkeu' amounted to 10,099, taken using the scraping method using TweetHarvest. The process of data cleaning and preprocessing processes this Data. In the preprocessing stage, there are six scenarios; some involve two libraries, namely, in the stemming process using NLTK and Sastawi libraries. Based on the analysis, it is found that in the steaming process, the Sastrawi library produces better base words compared to NLTK; for example, in the word 'diterima (accepted)', NLTK still produces the word 'diterima (accepted)' but in the literary Library produces the word 'terima (accept)'. The following process is to find the coherence value in each preprocessing scenario; the highest coherence value is 0.572250 in Scenario (D) and Scenario (F). Where both scenarios use the NLTK library, preprocessing results are modeled using LDA modeling, the results that are easy to understand will be selected to be the best results. It is known that preprocess scenario (E) is the best result compared to other scenarios because it does not contain many semantic words, each word is easy to understand, and the dominant topic is related to the focus of this study, but there are topics with a coherence value of 0 and similar topic values as in topics 4, 6, 7, 10, and 11, this also happens in preprocess Scenario (B). The visualization of the inter-topic distance map illustrates that each topic has a considerable distance, meaning that the topic of discussion on the Kemenkeu is diverse. With the dominant topic is Topic 1 and the dominant word is 'kemenkeu'. Overall, this study not only provides insights related to the discussion topics around Kemenkeu but also compares the results from various preprocessing scenarios

ACKNOWLEDGMENT

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

REFERENCES

Abidin, Z., Junaidi, A., & Wamiliana. (2024). Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review. *Journal of Information Systems Engineering and Business Intelligence*, *10*(2), 217–231. https://doi.org/10.20473/jisebi.10.2.217-231

Adhikari, S. (2022). Social Media and its Impacts in Human Minds. *Unity Journal*, *3*(01), 317–330. https://doi.org/10.3126/unityj.v3i01.43335

Alif Nur Iman, S. (2024). PENGARUH TREND PLATFORM DIGITAL SEBAGAI EDUKASI POLITIK TERHADAP PENINGKATAN PARTISIPASI POLITIK MASYARAKAT DI KOTA SURABAYA TAHUN 2023. *Journal of Comprehensive Science*, *3*(1), 37–48.

Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0255-7

Blei, D. M., Ng, A. Y., & Jordan, M. T. (2002). Latent dirichlet allocation. *Advances in Neural Information Processing Systems*, *3*, 993–1022.

Boulianne, S., Hoffmann, C. P., & Bossetta, M. (2024). Social media platforms for politics: A comparison of Facebook, Instagram, Twitter, YouTube, Reddit, Snapchat, and WhatsApp. In *New Media and Society* (Nomor July). https://doi.org/10.1177/14614448241262415

Erlisya, V., Aulia, A., Tobing, N. B., Saputra, B., Raja, M., Haji, A., & Korespondensi, T. (2024). Analisis Penyalahgunaan Kekuasaan dari Pejabat Kemenkeu yang Dilakukan oleh Rafael Alun Trisambodo. *Analisis Penyalahgunaan Kekuasaan (Erlisya, dkk.) Madani: Jurnal Ilmiah Multidisiplin*, *2*(5), 298–302. https://doi.org/10.5281/zenodo.11422692

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*(March), 1–17. https://doi.org/10.3389/fenrg.2021.652801

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

Habibi, P. W. C. and M. (2020). Entity Profiling to Identify Actor Involvement in Topics of Social Media Content. *IJCCS Indones. J. Comput. Cybern. Syst.*, *14*. https://doi.org/10.22146/ijccs.59869

Kang, H. J., Kim, C., & Kang, K. (2019). Analysis of the trends in biochemical research using latent dirichlet allocation (LDA). *Processes*, *7*(6), 1–14. https://doi.org/10.3390/PR7060379

Keuangan, M. (n.d.). *Organisasi dan Tata Kerja Kementerian Keuangan*.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

https://jdih.kemenkeu.go.id/in/dokumen/peraturan/c1

Lee, J., Kim, H., & Kron, F. (2024). Virtual education strategies in the context of sustainable health care and medical education: A topic modelling analysis of four decades of research. *Medical Education*, *58*(1), 47–62. https://doi.org/10.1111/medu.15202

Lubis, A. R., Prayudani, S., Lubis, M., & Nugroho, O. (2022). Sentiment Analysis on Online Learning during the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method. *2022 1st International Conference on Information System and Information Technology, ICISIT 2022*, *July*, 106–111. https://doi.org/10.1109/ICISIT54091.2022.9872926

Muzumdar, P., Kurian, G., & Basyal, G. P. (2024). A Latent Dirichlet Allocation (LDA) Semantic Text Analytics Approach to Explore Topical Features in Charity Crowdfunding Campaigns. *Asian Journal of Economics, Business and Accounting*, *24*(1), 1–10. https://doi.org/10.9734/ajeba/2024/v24i11207

Novira, S. T. (2019). *Sistem Pendukung Keputusan Pemilihan Jurusan Sekolah Menengah Kejuruan Dengan Menggunakan Metode Analytical Hierarchy …*. *2015*, 186–188. http://repositori.unsil.ac.id/782/

Nurmalasari, S., Hidayanto, A. N., Huwaida, L. A., & Wulandari, H. (2023). Sentiment Analysis and Topic Modeling of Citizen Satisfaction with the Indonesian Government in Handling a Pandemic. *OPSearch: American Journal of Open Research*, *2*(7), 246–256. https://doi.org/10.58811/opsearch.v2i6.61

Parveen, N., Santhi, M. V. B. T., Ramani Burra, L., Pellakuri, V., & Pellakuri, H. (2021). WITHDRAWN: Women's e-commerce clothing sentiment analysis by probabilistic model LDA using R-SPARK. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2020.10.064

Rabbani, H. A. (n.d.). *Sastrawi 1.0.1*.

Rinandyaswara, R., Sari, Y. A., & Furqon, M. T. (2022). Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi). *Jurnal Teknologi Informasi dan Ilmu Komputer*, *9*(4), 717. https://doi.org/10.25126/jtiik.2022934707

Rivaldy, A., Fedria Wowor, H. A., Maisya, S. R., & Safitri, D. (2021). Penggunaan Twitter Dalam Meningkatkan Melek Politik Mahasiswa Ilmu Komunikasi Universitas Negeri Jakarta. *Perspektif Komunikasi: Jurnal Ilmu Komunikasi Politik dan Komunikasi Bisnis*, *5*(1), 41. https://doi.org/10.24853/pk.5.1.41-48

Sahria, Y., & Hatta Fudholi, D. (2017). Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA (Latent Dirichlet Allocation). *Masa Berlaku Mulai*, *1*(3), 336–344.

Sakti, R. E., & Nainggolan, B. (2023). Understanding the Role of Social Media Toward Satisfaction of Government in Indonesia. *Jurnal Komunikasi Indonesia*, *12*(1). https://doi.org/10.7454/jkmi.v12i1.1185

Sonk, M., & Tunger, D. (2024). Trend mining with Orange – using topic modeling in futures research with the example of urban mobility. *European Journal of Futures Research*, *12*(1), 1–7. https://doi.org/10.1186/s40309-024-00229-1

Studies, L. (2024). *The Effect Of Political Influencer On Online Political Participation In Twitter/X*. *6*(2).

Tan, X., Zhuang, M., Lu, X., & Mao, T. (2021). An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model. *IEEE Access*, *9*, 15860–15871. https://doi.org/10.1109/ACCESS.2021.3052566

*name of corresponding author