P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v8i2.477 Vol. 8, No. 2. Maret 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

TOPIC MODELING OF PUBLIC DISCOURSE ON TWITTER ABOUT THE ASSET CONFISCATION BILL USING LATENT DIRICHLET ALLOCATION (LDA)

Azka Bima Aditya[1], Ahmad Abdul Chamid[2], Rizkysari MeiMaharani[3]

[123]Department of Informatics Engineering, Faculty of Engineering
Muria Kudus University, Kudus 59327, Indonesia
202251131@std.umk.ac.id, abdul.chamid@umk.ac.id, rizky.sari@umk.ac.id

Abstract

This study examines the structure of public discourse on Twitter regarding the Indonesian Asset Confiscation Bill, a policy initiative aimed at strengthening anti corruption enforcement and ensuring legal certainty. Moving beyond conventional sentiment classification, this research identifies how substantive public concerns are thematically organized within digital debate. A total of 14,319 cleaned and deduplicated tweets collected between January and September 2025 were analyzed using Latent Dirichlet Allocation with the optimal model configuration of nine topics selected based on coherence evaluation to ensure semantic interpretability. The findings reveal nine dominant thematic clusters, with law enforcement and regulatory enactment emerging as the primary focus, followed by legislative process dynamics, protest mobilization, party politics, and institutional accountability. These results indicate that online discourse is structured around normative concerns, particularly procedural clarity, fairness, and institutional legitimacy, rather than driven solely by emotional polarity. Scientifically, this study contributes by shifting the analytical emphasis from sentiment polarity toward systematic thematic mapping of digital political discourse using an optimized LDA framework tailored to Indonesian Twitter data characteristics. Practically, the findings provide policymakers with an evidence based monitoring instrument to identify priority public concerns, strengthen legislative communication strategies, and reduce interpretive ambiguity in sensitive regulatory deliberations.

Keywords: Asset Confiscation Bill; Twitter; topic modeling; LDA; coherence

Abstrak

Penelitian ini menganalisis struktur diskursus publik di Twitter mengenai RUU Perampasan Aset sebagai inisiatif kebijakan yang bertujuan memperkuat penegakan hukum antikorupsi dan menjamin kepastian hukum. Berbeda dari pendekatan analisis sentimen konvensional, penelitian ini berfokus pada pemetaan bagaimana kekhawatiran substantif publik terorganisasi secara tematik dalam perdebatan digital. Sebanyak 14.319 cuitan yang telah dibersihkan dan dihapus duplikasinya pada periode Januari hingga September 2025 dianalisis menggunakan Latent Dirichlet Allocation, dengan konfigurasi optimal sembilan topik yang dipilih berdasarkan evaluasi koherensi untuk memastikan keterpahaman semantik. Hasil penelitian menunjukkan sembilan klaster tema utama, dengan isu penegakan hukum dan pengesahan regulasi sebagai fokus dominan, diikuti dinamika proses legislasi, mobilisasi aksi protes, politik kepartaian, dan akuntabilitas institusional. Temuan ini mengindikasikan bahwa diskursus digital tidak semata didorong oleh reaksi emosional, tetapi terstruktur pada kekhawatiran normatif, terutama terkait kejelasan prosedural, keadilan, dan legitimasi kelembagaan. Secara ilmiah, penelitian ini berkontribusi dengan menggeser fokus analisis dari klasifikasi polaritas sentimen menuju pemetaan tematik sistematis diskursus politik digital melalui kerangka LDA yang dioptimalkan dan disesuaikan dengan karakteristik Twitter Indonesia. Secara praktis, hasil penelitian ini menyediakan instrumen berbasis data bagi pembuat kebijakan untuk mengidentifikasi isu prioritas publik, memperkuat strategi komunikasi legislasi, dan mengurangi ambiguitas interpretasi dalam perumusan regulasi yang sensitif.

Kata kunci: RUU Perampasan Aset; Twitter; pemodelan topik; LDA; koherensi

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i2.477
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

INTRODUCTION

Social media has now shifted far from its original role as a communication platform and has turned into a dynamic public space (Fatimah, 2025)(Fazri & Voutama, 2025). Among the various platforms available, Twitter (now known as X) plays a very important role as a space for the public to express opinions, share views, and debate social issues and government policies (Sander, 2025)(Gearhart et al., 2024)(Muhammad & Tanggahma, 2024). Conversations take place quickly, massively, and openly, reflecting the ever-changing dynamics of public opinion (Aditya et al., 2025). Therefore, data that appears on social media can be considered a true portrait of the sentiments, concerns, and aspirations of the public (Chamid, Nindyasari, Azizah, et al., 2025)(Chamid et al., 2023a). The ability to read and analyze these conversations is key for policymakers, especially the government, in understanding the direction of the developing public discourse.

Among various national policy issues, the Asset Confiscation Bill has become one of the topics that has attracted the most public attention. This regulation is viewed as a crucial step in strengthening law enforcement against corruption, a longstanding issue in Indonesia. Through the mechanism of confiscating assets derived from crime, this bill is expected to curb the practice of corruption while recovering state losses (Kaban & Kholiq, 2025)(Susilo et al., 2023). The discussion process has been dynamic because it involves various parties with different interests and perspectives. As a result, this dynamic often surfaces in the public sphere, especially on Twitter, which is then filled with various opinions, debates, and interpretations from the public that are diverse and often contradictory.

These diverse and conflicting opinions generate a huge amount of text data, making it difficult to process effectively using conventional text analysis methods (Chamid, Nindyasari, & Ghozali, 2025)(Cano-Marin et al., 2023). Efforts to understand this data face two main challenges. First, in terms of volume, the intensity of public conversation on Twitter generates data that far exceeds the capacity of manual analysis approaches. Second, in terms of language characteristics, text comments on Twitter are essentially unstructured and filled with various forms *of noise,* such as the use of informal language (*slang)*, abbreviations, and *typos*, which ultimately hinder the process of accurate interpretation (Chamid et al., 2023b).

Several previous studies have examined the issue of the Asset Confiscation Bill on social media, particularly Twitter, using a sentiment analysis approach. (Nugroho & Hasan, 2023) analyzed public sentiment towards this bill using the *Naïve Bayes* method and found that the majority of opinions were positive, although the model's accuracy was relatively low at 55%. Meanwhile, (Rofiqi & Akbar, 2024) used *Support Vector Machine (*SVM) and showed different results, namely a dominance of negative sentiment with an accuracy of 79.8% towards the bill. A similar study by (Sholekhah & Muntahanah, 2025) also compared the performance *of Naïve Bayes* and SVM on the issue of confiscating the assets of corruptors on Twitter, and found significant differences in public perception of this policy. These findings show that previous studies have generally focused on sentiment classification (positive–negative) and have not touched on the thematic aspects of the public discourse that has been formed.

On the other hand, research on topic modeling in social media has demonstrated the effectiveness of *Latent Dirichlet Allocation* (LDA) in revealing the complex structure of public discourse. (Kannitha et al., 2022) used LDA to map customer complaints against internet service providers on Twitter and successfully identified dominant topics with a high level of interpretability. (Ramadhan et al., 2025) applied a similar method to analyze political issues in the 2024 elections on Twitter and found the main topics that shaped the image of candidates in the digital space. Another study by (Nurhaliza et al., 2024) optimized LDA with a *bigram* approach and *coherence score* metrics, proving that parameter variation in the model can improve the accuracy of public theme identification.

Although LDA has proven effective in social media contexts, prior studies remain limited by conventional applications lacking systematic hyperparameter optimization, Indonesian Twitter-specific preprocessing, bigram modeling, single-metric evaluation, and combined quantitative-qualitative validation (Ramamoorthy et al., 2024)(Jelodar et al., 2019). This study selects LDA over alternatives like BERTopic or Dynamic Topic

Modeling for its superior interpretability via probabilistic topic-word distributions essential for nuanced public policy analysis, computational efficiency on big datasets without GPU requirements, flexibility in preprocessing Indonesian Twitter variations (unlike deep learning reliant on limited pretrained models), and focus on static thematic mapping—with a replicable framework integrating a six-stage preprocessing pipeline, bigrams, C_v+perplexity optimization, and interactive visualizations plus expert validation (Asghari et al., 2020).

This study proposes four methodological contributions based on LDA: (1) a six-stage preprocessing pipeline with a custom dictionary for Indonesian Twitter noise (Choirinnisa et al., 2025); (2) bigram modeling with optimal thresholds for policy collocations (Ramamoorthy et al., 2024)(Ye et al., 2023); (3) dual-metric optimization of the k number of topics via C_v coherence and perplexity (Chen & Komachi, 2023)(Kulkarni et al., 2023); and (4) interactive visualizations (word clouds, topic charts, intertopic maps) with expert validation for substantive topics in the Asset Confiscation Bill discourse (Lum & Chang, 2023).

Thus, the primary issue addressed in this research pertains to the application of comprehensive text preprocessing stages and the Latent Dirichlet Allocation (LDA) algorithm to effectively model public discourse, alongside determining the optimal number of topics that exhibit the highest mathematical coherence. The objective of this study is to identify and map the principal themes concerning the Asset Forfeiture Bill, thereby offering practical contributions to the government through an objective methodology for mapping societal aspirations. Through this analysis, the government can undertake in-depth monitoring of public opinion to discern shifts in focal attention, ranging from the predominance of law enforcement issues to the dynamics of student protest movements. These findings also serve as a foundation for responsive public communication strategies by aligning governmental messaging with key terminologies prevalent in digital spaces, such as 'reverse burden of proof' and 'impoverishment'. Strategically, this research proposes a more adaptive decision-making framework by directly integrating the structure of public discourse into the formulation of national regulations.

RESEARCH METHODS

This study uses a quantitative approach, which focuses on empirical proof of the effectiveness of the *Latent Dirichlet Allocation* (LDA) algorithm in mapping public discourse on Twitter. A quantitative approach was chosen because this study is oriented towards the processing of numerical data resulting from text transformation and mathematical model evaluation using the *Coherence Score* (C_v) and *Perplexity* metrics.

In the system development process, the *Data Science Pipeline* framework was used as the main methodology. This approach involves sequential stages from raw data collection to the formation *of machine-based* analysis models. Each stage is interconnected to ensure valid, interpretive, and representative topic modeling results for the analyzed public discourse structure.

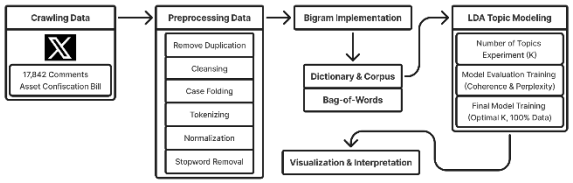


Figure 1. Research Methods

Several research stages are shown in Figure 1 as follows.

Data Crawling

The initial stage of this research was data collection (*data crawling*) from Twitter (X) using the keyword "UU Perampasan Aset" (Asset Confiscation Law). *Data crawling* is the process of automatically retrieving data from a digital source, with the aim of extracting large amounts of information without manual intervention (Lathifah et al., 2021)(Purwandari et al., 2023). In the context of this research, *crawling* is used to download data from the Twitter server in the form of users, tweet content, and other supporting attributes as a basis for analysis. Data collection is carried out by running a *Node.jsbased* script through the *tweet-harvest* package, which functions to automatically extract tweets and save them in *CSV* format. Before the extraction process was run, all the necessary dependencies were installed, and authentication was performed using a *Twitter Auth Token* to gain access to public data. The data collection period was set from January 1, 2025, to September 30, 2025, in

order to capture the dynamics of public conversation during the bill discussion process. Each crawling session generated around 600 to 800 tweets, and this process was repeated until a total of 17,842 public comments were collected as the main corpus for the study. After the entire crawling process was complete, the data was loaded into *a DataFrame* using *pandas.read_csv()* function was used, and the number of rows was verified to ensure the quality and completeness of the data obtained.

All data utilized in this study comprised publicly available Twitter (X) posts, with no access to private messages or personal information. User privacy was rigorously protected by excluding usernames and personal identifiers from analysis and publication. The study conducted aggregate-level discourse pattern identification rather than individual user evaluation, adhering to established social media research ethics principles that emphasize responsible use of public data while safeguarding user privacy.

Data Preprocessing

Data preprocessing is an important stage to clean and prepare the text for computational processing, with the aim of normalizing and reorganizing the data without changing the validity of *the dataset's* content (Putra et al., 2024). In the context of *Natural Language Processing,* this stage is necessary because Twitter data is generally filled with *noise* such as non-standard language, abbreviations, emojis, symbols, and spelling errors that can interfere with analysis (Rifaldi et al., 2023)(Jazuli et al., 2025). The *preprocessing* process is carried out systematically through several steps, starting with *cleansing* to remove non-text elements such as URLs, mentions, hashtags, and symbols (Kartika Sari et al., 2024), then *case folding* to standardize letters to lowercase to avoid word duplication (Muhammad Zuama Al Amin et al., 2025). After that, *tokenizing* is applied to break the text into word units (Arifin & Mahdiana, 2024), followed *by normalization* to convert non-standard words to standard forms using a conversion dictionary (Kusuma Wardana et al., 2025), and ending with *stopword removal* to remove common words that do not contribute significantly to the analysis (Sehabudin et al., 2025). In this study, the *stopword* list was also adjusted to the characteristics of social media, including informal expressions such as "wkwk" or swear words, so that the final corpus produced was cleaner and more relevant for the topic modeling process.

Bigram Implementation

After the *preprocessing* stage was completed, this study applied *bigrams* to identify pairs of words that frequently appeared together and had contextual meaning in public conversations on Twitter. This approach was important because the discourse on the Asset Confiscation Bill often involved two-word terms that could not be represented well if processed as single words. The bigram extraction results show the dominant occurrence of word pairs such as *"hukuman_mati", "pembuktian_terbalik", "anggota_dewan", "ketok_palu", dan "penegak_hukum",* confirming that issues of law enforcement, the legislative process, and the accountability of state officials are the focus of public attention in the analyzed dataset. Methodologically, the use of *n-grams* in text analysis serves to capture specific word sequences that reveal co-occurrence patterns that are not apparent in unigrams, and *bigrams* were chosen because they provide richer semantic information without excessively increasing feature dimensions (Ayu Anjani & Fauzan, 2021). Bigram formation was conducted using Gensim's Phrases model with parameters min_count=10 and threshold=50. The min_count parameter was set to 10 to ensure only word pairs appearing significantly across the corpus (17,842 documents) were considered, thereby mitigating noise from infrequent combinations (Sadeghzadehyazdi et al., 2021). Meanwhile, a threshold of 50 controlled inter-word association strength, retaining only semantically robust collocations as single tokens (Asghari et al., 2020)(Luo & Cui, 2024). These parameters were selected to balance representational completeness with model stability, enhancing LDA topic accuracy and interpretability.

Dictionary and Corpus Formation

The dictionary and *corpus* formation stage is carried out to convert text that has undergone a cleaning process into a numerical representation using the *Bag-of-Words* (BoW) approach. In this approach, each document is represented as *a vector* based on the frequency of word occurrence without regard to order or grammatical structure, allowing computers to assess the importance of a word in the context of the entire corpus in a simple but effective manner (Artanto, 2025). At this stage, *the dictionary* is formed as a set of unique words complete with their indices, while *the corpus* represents each document in pairs (word_id, frequency) that describe the distribution of word occurrences

(Roziewski & Kozłowski, 2021). The BoW representation was chosen because it is consistent with the characteristics of the *Latent Dirichlet Allocation* (LDA) algorithm, which models topics as a probabilistic distribution of words in *the corpus* and requires numerical data to identify word occurrence patterns mathematically. Through this process, the model obtains a structural foundation for recognizing word relationships and forming latent topics more accurately in the next modeling stage.

LDA Topic Modeling

The core stage of this research is the application of the *Latent Dirichlet Allocation* (LDA) algorithm as a *topic modeling* method to map the latent topic structure in public conversations on Twitter regarding the Asset Confiscation Bill. *Topic modeling* is a statistical approach in NLP that works through unsupervised learning by identifying groups of words that frequently appear together and interpreting them as a topic, while LDA is the most widely used method because it models each document as a mixture of several topics and each topic as a specific word distribution based on the probabilistic principle of (Setijohatmo et al., 2020).

The LDA model was implemented using Gensim's LdaModel with parameters optimized for stability, interpretability, and computational efficiency. The num_topics parameter was determined through coherence evaluation to identify the most representative topic count, while alpha='auto' enabled flexible, realistic topic distribution estimation per document. Eta=0.1 was selected to produce focused word distributions yielding specific, interpretable topic keywords. Training employed passes=10 for stable convergence through sufficient iterations, with chunksize=1000 and update_every=1 ensuring efficient incremental model updates. Reproducibility was guaranteed via random_state=42, and per_word_topics=True facilitated granular per-word topic distribution analysis.

Topic Coherence and Perplexity

The implementation of LDA in this study includes experiments on the number of topics, model evaluation, and final model training. The evaluation of the Latent Dirichlet Allocation (LDA) model is conducted utilizing two primary metrics: Topic Coherence (C_v) and Perplexity. Topic Coherence is employed to quantify the degree of semantic relatedness among words within a single topic. A higher coherence value indicates that the words in the topic are more conceptually relevant and more comprehensible to human interpreters. Conversely, Perplexity is utilized to measure the model's predictive capability on new data. A lower perplexity value signifies superior statistical generalization performance. Nonetheless, given the emphasis of this study on the interpretation of public discourse, the coherence metric is prioritized as the primary indicator for ascertaining the optimal number of topics, while perplexity serves as a supplementary measure to ensure model stability.

Visualization and Interpretation of Topics

Visualization and interpretation of topic results is the final stage, intending to present the modeling results visually so that they are easier to analyze and understand. LDA results can also be displayed interactively through *pyLDAvis* to observe the relationships between topics and the distribution of words within them. Based on these visualizations, manual interpretation is performed to label each topic and draw conclusions about the patterns of public discourse related to the Asset Confiscation Bill on Twitter.

Methodological Limitations

Although LDA is effective in identifying latent thematic structures, it has inherent limitations. The model relies on a Bag-of-Words representation that ignores syntactic structure and contextual nuances, making it unable to detect irony, sarcasm, or implicit rhetorical meanings that often appear in social media discourse. In addition, LDA treats the corpus as a static collection of documents and does not account for temporal dynamics, meaning shifts in discourse over time are not explicitly modeled. Consequently, the results reflect structural co-occurrence patterns rather than deeper pragmatic interpretation. Furthermore, the Bag-of-Words assumption disregards word order and grammatical relationships, which may limit semantic depth. While LDA is suitable for uncovering dominant thematic clusters, it does not fully capture the complexity of political communication in digital environments. These limitations suggest that future research could incorporate dynamic topic modeling or contextual embedding-based approaches to enhance temporal and semantic sensitivity.

RESULTS AND DISCUSSION

Data Crawling

The first stage of the research was data collection using the data *crawling* method from the Twitter (X) platform. The *crawling* process was carried out using a *Node.js* script through the *tweet-harvest* package with the keyword "Asset Confiscation Law" in the range of January 1 to September 30, 2025. This process yielded a total of 17,842 tweets, each accompanied by metadata such as publication time, number of interactions, language, and links to the original tweets. The structure of the crawled dataset is shown in Figure 2. The structure contains 15 columns with *float, integer,* and *object* data types**.** Some columns, such as *image_url* and *in_reply_to_screen_name*, have varying non-null values, while the *location* and *username* columns are empty. This information provides an initial overview of the characteristics of the dataset and the level of completeness of the available attributes.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17842 entries, 0 to 17841
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   conversation_id_str     17842 non-null  float64
 1   created_at              17842 non-null  object
 2   favorite_count          17842 non-null  int64
 3   full_text               17842 non-null  object
 4   id_str                  17842 non-null  float64
 5   image_url               1385 non-null   object
 6   in_reply_to_screen_name 12817 non-null  object
 7   lang                    17842 non-null  object
 8   location                0 non-null      float64
 9   quote_count             17842 non-null  int64
 10  reply_count             17842 non-null  int64
 11  retweet_count           17842 non-null  int64
 12  tweet_url               17842 non-null  object
 13  user_id_str             17842 non-null  float64
 14  username                0 non-null      float64
dtypes: float64(5), int64(4), object(6)
memory usage: 2.0+ MB
```

Figure 2. Twitter Crawling Data Structure

Data Preprocessing

After the data has been collected through the *crawling* process, the next step is *preprocessing*, which aims to clean and prepare the text so that it can be processed computationally. The first step is to remove duplicate data that appears during the repeated *crawling* process. Of the total 17,842 raw tweets collected, 14,319 entries were declared unique and suitable for analysis. Duplicate removal is an important part because the same conversation can be extracted more than once when tweets with high interaction often reappear in Twitter search results.

After the dataset has been cleaned of duplicates, a series of preprocessing steps is applied systematically. These steps include *cleansing* to remove non-text elements such as URLs, mentions, symbols, and emojis, followed by *case folding* to standardize all text to lowercase so that variations in capitalization do not cause duplication of meaning. *Normalization* is performed to convert non-standard words into standard forms, for example, "nggak" to "tidak" or "yg" to "yang". The next stage is *tokenizing* to break the text into words, followed by *stopword removal* to remove common words that do not contribute significantly to meaning. In this study, the *stopword* list was expanded to include expressions typical of social media, such as "wkwk," "anj," and other informal expressions. In addition, domain terms such as "uu," "ruu," "perampasan," and "aset" were also removed because all data focused on the same topic, so these words had the potential to cause bias and hinder the LDA model's ability to extract more informative derivative issues.

The results of the entire series of steps are combined in the *pipeline_text* column, which contains the final text representation ready for use in topic modeling, as shown in Table 1.

Table 1. Data Preprocessing

full text	situ presidennya tinggal balikin kpk ke mode independen trus sahkan uu perampasan aset buat hukuman pidana seberat2nya buat para koruptor nggak ada remisi buat para koruptor cabut hak politiknya. kalau bisa sekalian siapin peti mati bisa nggak?
Cleansing	situ presidennya tinggal balikin kpk ke mode independen trus sahkan uu perampasan aset buat hukuman pidana seberatnya buat para koruptor nggak ada remisi buat para koruptor cabut hak politiknya kalau bisa sekalian siapin peti mati bisa nggak
case folding	situ presidennya tinggal balikin kpk ke mode independen trus sahkan uu perampasan aset buat hukuman pidana seberatnya buat para koruptor nggak ada remisi buat para koruptor cabut hak politiknya kalau bisa sekalian siapin peti mati bisa nggak
normalization	situ presidennya tinggal balikin kpk ke mode independen terus sahkan uu perampasan aset buat hukuman pidana seberatnya buat para koruptor tidak ada remisi buat para koruptor cabut hak politiknya kalau bisa sekalian siapin peti mati bisa tidak
Tokenizing	['situ', 'presidennya', 'tinggal', 'balikin', 'kpk', 'ke', 'mode', 'independen', 'terus', 'sahkan', 'uu', 'perampasan', 'aset', 'buat', 'hukuman', 'pidana', 'seberatnya', 'buat', 'para', 'koruptor', 'tidak', 'ada', 'remisi', 'buat', 'para', 'koruptor', 'cabut', 'hak', 'politiknya', 'kalau', 'bisa', 'sekalian', 'siapin', 'peti', 'mati', 'bisa', 'tidak']

stopword removal	['situ', 'presidennya', 'tinggal', 'balikin', 'kpk', 'mode', 'independen', 'sahkan', 'hukuman', 'pidana', 'seberatnya', 'koruptor', 'remisi', 'koruptor', 'cabut', 'hak', 'politiknya', 'siapin', 'peti', 'mati']
pipeline text	situ presidennya tinggal balikin kpk mode independen sahkan hukuman pidana seberatnya koruptor remisi koruptor cabut hak politiknya siapin peti mati

Bigram Implementation

Bigram implementation was carried out to capture word pairs that appear together and have important contextual meaning in public discourse regarding the Asset Confiscation Bill on Twitter. Many key issues on this platform are not adequately represented when treated as single words, such as "hukuman_mati", "pembuktian_terbalik", or "ketok_palu". For this reason, *bigrams* were formed using *Gensim Phrases* with parameters *min_count = 10* and *threshold = 50,* so that only word pairs that truly have a strong association were combined into a single token. This process produced a more semantically rich representation of the text through the integration of *bigrams* into each document and became the basis for the visualization in Figure 3.

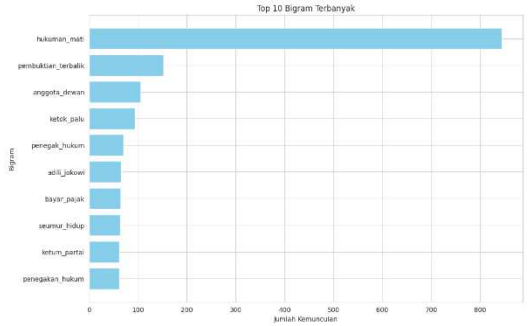


Figure 3. Bigram Visualization

The *bigram* extraction results show the dominance of word pairs such as "hukuman_mati", "pembuktian_terbalik", "anggota_dewan", and "ketok_palu" as *the bigrams* with the highest frequency. This pattern indicates that public attention is not only focused on the urgency of passing the Asset Confiscation Bill, but also on broader issues related to law enforcement, legislative mechanisms, and the accountability of public officials. The visualization in Figure 3 clearly shows how these *bigrams* appear much more frequently than other word pairs, confirming that *bigrams* are able to capture co-occurrence structures that are not visible in unigrams and provide a stronger semantic foundation for the topic modeling stage using LDA.

Dictionary and Corpus Formation

The dictionary and *corpus* were created by converting the *preprocessed* text and *bigrams* into numerical representations using the *Bag-of-Words* (BoW) approach. At this stage, *the dictionary* was built to contain all unique words and their numerical indices, while *the corpus* represented each document in pairs of word IDs and their frequency of occurrence. This numerical representation forms the basis for the *Latent Dirichlet Allocation* (LDA) algorithm to recognize word distribution patterns mathematically without considering the order of words in a sentence. Using this *dictionary* and *corpus* structure, the LDA model was trained to generate ten topics, which are then summarized in Table 2, containing the dominant words that best represent each topic in the public discourse on the Asset Confiscation Bill.

Table 2. LDA Topic Extraction Results Based on Dictionary and Corpus

Topic	Dominant Words
0	0.039*"menunggu" + 0.017*"pemiskinan" + 0.014*"semoga" + 0.013*"jaman" + 0.012*"mudah"
1	0.067*"dpr" + 0.064*"rakyat" + 0.021*"mahasiswa" + 0.021*"pengesahan" + 0.019*"anggota"
2	0.142*"rakyat" + 0.023*"langsung" + 0.019*"pajak" + 0.013*"sita" + 0.008*"selesai"
3	0.138*"demo" + 0.030*"dukung" + 0.026*"buzzer" + 0.025*"kemarin" + 0.013*"ppn"
4	0.043*"kim" + 0.039*"koalisi" + 0.033*"partai" + 0.025*"pdip" + 0.018*"psi"
5	0.102*"koruptor" + 0.047*"sahkan" + 0.038*"korupsi" + 0.024*"disahkan" + 0.021*"negara"
6	0.048*"jokowi" + 0.030*"partai" + 0.024*"dpr" + 0.022*"ketua" + 0.019*"pdip"
7	0.027*"anti" + 0.011*"pertamina" + 0.011*"kilat" + 0.010*"teriak" + 0.010*"mandek"
8	0.020*"mulyono" + 0.012*"parlemen" + 0.011*"kim_plus" + 0.010*"adili" + 0.009*"penguasa"
9	0.015*"media" + 0.013*"mustahil" + 0.010*"pemimpin" + 0.009*"kuat" + 0.009*"dukungan"

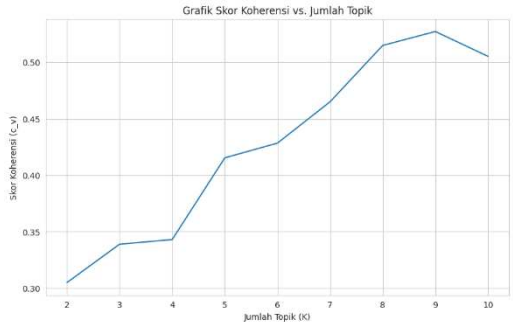
LDA Topic Modeling

a. Number of Topics Experiment (K)

The number of topics experiment was

conducted to determine the most representative K value in mapping the structure of public discourse related to the Asset Confiscation Bill. This process involved training the LDA model with a variety of topic numbers ranging from K = 2 to K = 10, where each model was then measured for its coherence value using the *Coherence Score* (C_v) metric. This metric was chosen because it is able to evaluate the semantic connection between words in a topic, so that a higher coherence score indicates that the topic is easier to understand and more semantically consistent.

The test results show that the coherence score increases gradually as the number of topics increases, with a more significant increase after K = 5. The highest value was achieved at K = 9, with a coherence score of 0.5273. This trend indicates that selecting K = 9 provides the best balance between topic granularity and semantic coherence, so this value was chosen as the optimal configuration to use in the final model training stage, as visualized in Figure 4.



Gambar 4. Grafik Skor Koherensi terhadap Variasi Jumlah Topik (K)

b. Model Evaluation Training (Coherence & Perplexity)

The evaluation model was trained using the K=9 value that previously produced the highest coherence score. Training was conducted by separating the data into 11,455 *training* documents and 2,864 *testing* documents. The evaluation results are shown in Figure 5, which shows that the model achieved a coherence score of 0.5415 on the training data, indicating that the relationship between words in a topic is at a strong semantic level. When tested on the *testing* data, *the log-perplexity* was -10.5410 with *a perplexity* of 1489.9016. This low *perplexity* value reflects that the model has good predictive ability in recognizing the distribution of words in new documents.

```
Total dokumen: 14319
Dokumen training: 11455
Dokumen testing: 2864

Mulai melatih model EVALUASI (K=9) pada data training...
Model EVALUASI selesai dilatih.

--- HASIL KUALITAS MODEL ---
Jumlah Topik: 9
Skor Koherensi (c_v) [pada data train]: 0.5415
Log Perplexity [pada data test]: -10.5410
Perplexity [pada data test]: 1489.9016
```

Gambar 5. Hasil Model Evaluasi (Coherence & Perplexity)

The combination of high coherence and low *perplexity* shows that the nine-topic configuration produces the most stable and interpretable performance compared to other K values. These evaluation results provide a strong basis for continuing the final model training using all documents so that the resulting topic structure is more comprehensive and representative of the public discourse on the Asset Confiscation Bill.

c. Final Model Training (Optimal K, 100% Data)

The final model was built using K=9, which had previously been proven to provide the best performance in the evaluation stage. All 14,319 documents were then used as full training data to ensure that all variations in language, opinion, and word co-occurrence patterns in public discourse were fully accommodated. Training using 100% of the data resulted in a more stable topic distribution and was able to capture the structure of public discourse more comprehensively than the evaluation model, which only utilized training data.

The results of the final model training show nine main topics that shape public discourse on the Asset Confiscation Bill. Each topic consists of words with the highest probability weight, such as the dominance of the words corruptor, legalize, and DPR (House of Representatives) in the topic of law enforcement; demo and student in the topic of protest actions; and party, Jokowi, and kpk (Corruption Eradication Commission) in the topic of national political dynamics. A complete list of dominant words from each topic is shown in Table 3, which serves as the basis for the labeling and interpretive analysis processes in the next

stage of this study.

Table 3. Final LDA Model Topic Extraction Results (K = 9)

Topic	Dominant Words
0	'0.035*"menunggu" + 0.025*"perpu" + 0.019*"parpol" + 0.016*"kecuali" + 0.016*"pemiskinan"'
1	'0.032*"bahas" + 0.027*"kepentingan" + 0.026*"dibahas" + 0.022*"suara" + 0.015*"kim_plus"'
2	'0.143*"rakyat" + 0.020*"pengesahan" + 0.016*"pakai" + 0.015*"langsung" + 0.014*"tuntutan"'
3	'0.143*"demo" + 0.073*"mahasiswa" + 0.029*"buzzer" + 0.025*"kemarin" + 0.011*"sok"'
4	'0.069*"tni" + 0.035*"cepat" + 0.033*"koalisi" + 0.024*"revisi" + 0.013*"polri"'
5	'0.106*"koruptor" + 0.048*"sahkan" + 0.043*"dpr" + 0.040*"korupsi" + 0.025*"disahkan"'
6	'0.012*"apbn" + 0.009*"mustahil" + 0.008*"kursi" + 0.008*"peraturan" + 0.007*"investor"'
7	'0.033*"tinggal" + 0.025*"jalan" + 0.024*"anti" + 0.013*"pro" + 0.013*"dikebut"'
8	'0.050*"partai" + 0.050*"jokowi" + 0.038*"kpk" + 0.036*"pdip" + 0.017*"mulyono"'

d. Topic Labeling and Interpretive Refinement Process

Topic labeling employed a qualitative-guided interpretive approach integrating LDA's probabilistic word distributions with contextual analysis of representative documents. Initially, the top five to ten highest-probability words per topic were identified as primary statistical indicators. Subsequently, documents exhibiting peak topic probabilities were examined to discern consistent narrative patterns, opinion orientations, and argumentative contexts. This methodology ensured topic labels derived from emergent semantic structures rather than isolated dominant terms.

For instance, Topic 5 dominated by "koruptor" (corruptor), "sahkan" (enact), "DPR" (House of Representatives), and "korupsi" (corruption)—exhibited consistent discourse demanding regulatory enactment acceleration and legal enforcement against corruption perpetrators, yielding the label "Law Enforcement and Regulatory Enactment." Topic 3, characterized by "demo" (demonstration), "mahasiswa" (students), and "buzzer," reflected mobilization dynamics and digital social responses, labeled "Protest Actions and Public Mobilization." Initially abstract Topic 0, featuring "menunggu" (waiting) and "kecuali" (except), revealed through contextual analysis with "perpu" (government regulation in lieu of law) and "parpol" (political parties) patterns of legislative uncertainty and inter-party political negotiations, designated "Legislative Uncertainty and Party Political Dynamics." This approach established labels through semantic consistency and narrative coherence rather than word frequency alone.

Visualization and Interpretation

The visualization of the modeling results was carried out using *pyLDAvis* to map the structure of nine topics formed from public conversations about the Asset Confiscation Bill. The map of the distance between topics in Figure 6 shows the distribution of topics in a two-dimensional field based on multidimensional *scaling***.** The size of the circles shows the proportion of each topic's appearance in the corpus, while the distance between circles indicates the semantic differences between topics. This visualization reveals the existence of one dominant topic with a circle size that is much larger than the other topics, indicating that the issues of law enforcement, corruption, and demands for regulatory approval are the main focus of public discourse. Meanwhile, smaller topics tend to cluster together, indicating contextual similarities in the social criticism surrounding the debate on the bill.

JURNAL RISET INFORMATIKA
Vol. 8, No. 2. Maret 2026

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i2.477

Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

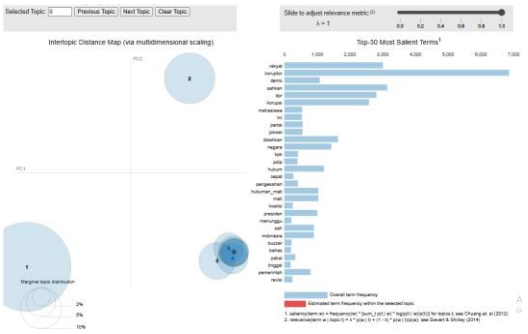


Figure 6. Map of Topic Distribution and Semantic Distance from LDA Results

To clarify the content structure of each topic, the *Top-30 Most Relevant Terms* graph is displayed for the two most representative topics, namely law enforcement and the legislative process. The visualization for the law enforcement topic in Figure 7 shows the dominance of words such as corruptor, legalize, DPR, and corruption, which reveals strong public sentiment regarding the need for firm legal action and the acceleration of law ratification. Meanwhile, the visualization for the legislative topic in Figure 8 shows terms such as "people," "ratification," "direct," "demand," and "use," reflecting public attention to the bill discussion process and demands for legislative decisions to be in the interests of the wider community. When viewed together with the previously formed topic table, this overall visualization shows that LDA modeling not only produces statistically separate topic structures but also reflects the complexity of public discourse, ranging from legal demands and national political dynamics to protest responses that emerge in the digital space.

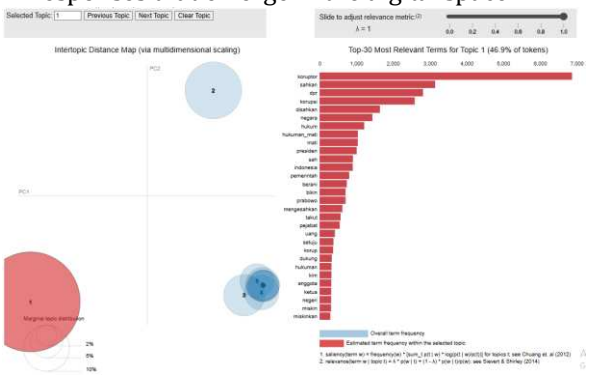


Figure 7. Distribution of Topics and Dominant Words on the Topic of Law Enforcement

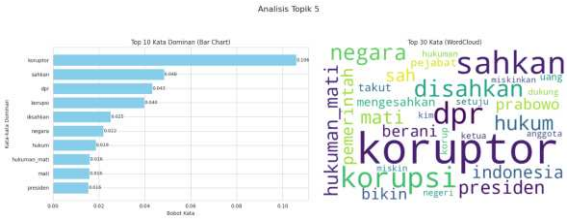


Figure 8. Bar Chart and Word Clouds on the Topic of Law Enforcement

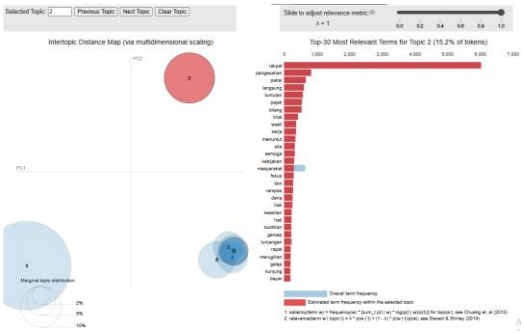


Figure 9. Distribution of Topics and Dominant Words in the Legislation Topic

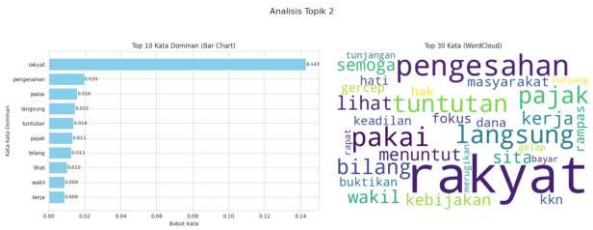


Figure 10. Bar Chart and Word Clouss in the Legislation Topic

Legal certainty emerges as a dominant theme because it functions as the operational foundation of a rule-of-law state while simultaneously acting as a mechanism for reducing uncertainty in coercive public policies. From the perspective of rule of law theory and institutional economics, regulations that are clear, consistent, and non-ambiguous reduce social costs and minimize the risk of discretionary abuse. When a policy involves sensitive matters such as state intervention in property rights or asset confiscation, the public rationally evaluates the precision of procedures, limits of authority, and oversight mechanisms before assessing the policy's normative objectives. Topic modeling results that highlight terms such as "rights," "courts," "procedures," and "protection" indicate that public discourse centers not merely on policy goals, but on the assurance of normative clarity.

Legal certainty is also structurally linked to public policy legitimacy. Within the framework of legal-rational authority, state power is accepted not because of coercive capacity, but because rules are impersonal, predictable, and consistently applied. Procedural justice theory demonstrates that perceived fairness of procedures exerts a stronger influence on policy acceptance than the threat of sanctions. When legal provisions are perceived as ambiguous or overly discretionary, the public tends to interpret them as vulnerable to abuse of power. At that point, legitimacy weakens, institutional trust declines, and implementation costs increase as voluntary compliance shifts toward skepticism or resistance.

Social media discourse, particularly on X (formerly Twitter), reinforces this dynamic. Digital platforms amplify concerns regarding normative ambiguity through framing effects and algorithmic visibility. Narratives questioning legal definitions, evidentiary standards, or potential criminalization risks often circulate more rapidly than formal governmental explanations. Public conversations on X reveal that procedural uncertainty is quickly reframed as a collective threat to civil rights, shaping a shared perception of institutional risk. Therefore, the dominance of legal certainty in topic modeling results is not merely a linguistic pattern, but a reflection of legitimacy negotiations in the digital era, where legal interpretation and institutional trust are continuously contested in online public spheres.

When compared with previous studies that applied sentiment analysis to public policy debates on social media, the findings of this study both reinforce and extend earlier conclusions. Prior research consistently shows that controversial legislative issues tend to generate polarized and predominantly negative sentiment, often reflecting distrust and skepticism. The present LDA-based analysis does not contradict this pattern; rather, it reveals the structural foundation beneath those sentiments. The dominance of themes related to legal certainty, procedural clarity, and institutional accountability indicates that negative sentiment is not merely emotional reaction, but is anchored in substantive normative concerns about rule-of-law guarantees and fairness in legislative processes. Moreover, while earlier studies typically categorize discourse into broad polarity classes (positive, negative, neutral), this research demonstrates that public opinion is organized into distinct thematic clusters such as law enforcement urgency, protest mobilization, party politics, and legislative uncertainty. This suggests that digital political discourse is not purely affect-driven but issue-structured. By moving beyond emotional classification toward thematic mapping, this study complements prior sentiment research and provides a more systematic explanation of how legitimacy debates around the Asset Confiscation Bill are constructed in online public spheres.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study demonstrates that an optimized LDA framework integrating structured preprocessing, bigram modeling, and dual-metric evaluation using coherence and perplexity is effective in mapping public discourse on the Asset Confiscation Bill using Indonesian Twitter data. The optimal configuration with K equal to 9 reveals that public discussion is organized into distinct thematic clusters, with law enforcement and regulatory enactment emerging as the dominant axis, followed by legislative dynamics, protest mobilization, and political contestation. These findings indicate that digital discourse is structured around substantive issues rather than driven purely by emotional polarity, particularly highlighting concerns related to legal certainty and institutional legitimacy. At the theoretical level, this research extends prior sentiment-based studies by shifting the analytical focus from polarity classification to thematic structure analysis. Instead of identifying opinions as simply positive or negative, the model uncovers the substantive foundations of public concern, especially regarding procedural clarity and accountability. However, the study is limited by its reliance on Twitter data within a specific time frame and by methodological constraints inherent in LDA, which uses a Bag-of-Words representation that does not capture sarcasm, contextual nuance, or temporal discourse evolution. Consequently, the results should be interpreted as structural

discourse mapping rather than a definitive measurement of overall public consensus. From a governance perspective, the dominance of topics related to law enforcement and legislative processes provides concrete and actionable insights. Topic modeling can serve as a data-driven monitoring instrument to detect shifts in public attention and identify specific legal mechanisms that require clearer explanation. By aligning communication strategies with dominant thematic clusters instead of reacting solely to fluctuating sentiment, government institutions can reduce interpretive ambiguity, strengthen procedural legitimacy, and improve institutional trust during the policymaking process.

Suggestion

Future research should integrate sentiment analysis, stance detection, or dynamic topic modeling to capture emotional orientation and temporal shifts in discourse more comprehensively. Comparative analysis across multiple social media platforms would enhance representativeness and robustness of findings. From a policy standpoint, it is recommended that government institutions incorporate structured discourse analytics into regulatory communication frameworks by developing topic-based monitoring systems that support adaptive messaging, transparent legislative explanation, and evidence-based public engagement during policy deliberation.

REFERENCES

Aditya, A. B., Samsudin, S., Rizki, W. P., Mahendra, M., & Setiawan, A. (2025). Analisis Sentimen Ulasan Aplikasi Gojek Menggunakan Support Vector Machine Dan Random Forest. *Jurnal Informatika Terpadu*, *11*(2), 134–143. https://doi.org/10.54914/jit.v11i2.1884

Arifin, M., & Mahdiana, D. (2024). IMPLEMENTATION OF DEEP LEARNING MODELS IN HATE SPEECH DETECTION ON TWITTER USING AN NATURAL LANGUAGE PROCESSING APPROACH. *Jurnal Teknik Informatika (Jutif)*, *5*(5), 1257–1266. https://doi.org/10.52436/1.jutif.2024.5.5.2043

Artanto, F. A. (2025). Comparison of Classification Algorithms with Bag of Words Feature in Sentiment Analysis. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *14*(3), 422–428. https://doi.org/10.32736/sisfokom.v14i3.2426

Asghari, M., Sierra-Sosa, D., & Elmaghraby, A. S. (2020). A topic modeling framework for spatio-temporal information management. *Information Processing & Management*, *57*(6), 102340. https://doi.org/10.1016/j.ipm.2020.102340

Ayu Anjani, S., & Fauzan, A. (2021). Implementasi n-Gram dalam Analisis Sentimen Masyarakat DIY terhadap PSBB Jawa-Bali Jilid II Menggunakan Naive Bayes Classifier. *STATISTIKA Journal of Theoretical Statistics and Its Applications*, *21*(2), 73–83. https://doi.org/10.29313/statistika.v21i2.294

Cano-Marin, E., Mora-Cantallops, M., & Sánchez-Alonso, S. (2023). Twitter as a predictive system: A systematic literature review. *Journal of Business Research*, *157*, 113561. https://doi.org/https://doi.org/10.1016/j.jbusres.2022.113561

Chamid, A. A., Nindyasari, R., Azizah, N., & Hariyadi, A. (2025). Analysis of Public Opinion on The Governor Candidate Debate Using LDA and IndoBERT. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*. https://doi.org/10.22219/kinetik.v10i3.2221

Chamid, A. A., Nindyasari, R., & Ghozali, M. I. (2025). Comparative Analysis of Machine Learning Algorithms for Predicting Patient Admission in Emergency Departments Using EHR Data. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *9*(2), 185–194. https://doi.org/10.29207/resti.v9i2.6188

Chamid, A. A., Widowati, & Kusumaningrum, R. (2023a). Graph-Based Semi-Supervised Deep Learning for Indonesian Aspect-Based Sentiment Analysis. *Big Data and Cognitive Computing*, *7*(1). https://doi.org/10.3390/bdcc7010005

Chamid, A. A., Widowati, & Kusumaningrum, R. (2023b). Multi-Label Text Classification on Indonesian User Reviews Using Semi-Supervised Graph Neural Networks. *ICIC Express Letters*, *17*(10), 1075–1084. https://doi.org/10.24507/icicel.17.10.1075

Chen, Z., & Komachi, M. (2023). Discontinuous Combinatory Constituency Parsing. *Transactions of the Association for Computational Linguistics*, *11*, 267–283. https://doi.org/10.1162/tacl_a_00546

Choirinnisa, D., Alzami, F., Indrayani, H., Rohmani, A., Nugraini, S., Zulfiningrumi, R., & Susanti, F. (2025). LDA Topic Modeling: Twitter-Based

Public Opinion on Indonesian Ministry of Finance. *Sinkron*, *9*, 849–863. https://doi.org/10.33395/sinkron.v9i2.14719

Fatimah, S. (2025). Transformasi Ruang Publik Digital: Tantangan Sosial dan Konstitusional dalam Demokrasi Era Media Baru. *Cakrawala*, *19*(1), 67–86. https://doi.org/10.32781/cakrawala.v19i1.785

Fazri, M., & Voutama, A. (2025). ANALISIS SENTIMEN PUBLIK TERHADAP DANANTARA DI MEDIA SOSIAL X MENGGUNAKAN NLP DAN PEMBELAJARAN MESIN. *JOISIE (Journal Of Information Systems And Informatics Engineering)*, *9*(1), 197. https://doi.org/10.35145/joisie.v9i1.4924

Gearhart, S., Zhang, B., & Adegbola, O. (2024). Tweeting, talking, or doing politics? Testing the influence of communication on democratic engagement. *Telematics and Informatics Reports*, *16*, 100167. https://doi.org/https://doi.org/10.1016/j.teler.2024.100167

Jazuli, A., Widowati, Chamid, A. A., & Kusumaningrum, R. (2025). Transformer-based semantic indexing for aspect-based sentiment analysis using an enhanced index generation algorithm with BERT. *International Journal of Advanced Technology and Engineering Exploration*, *12*(127), 907–926. https://doi.org/10.19101/IJATEE.2024.11110211 4

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Kaban, K. S., & Kholiq, A. (2025). Optimalisasi Regulasi Pidana Terkait Perampasan Aset Tindak Pidana Kejahatan Ekonomi Berlandaskan Perspektif Hukum Progresif Berkeadilan. *Jurnal Locus Penelitian Dan Pengabdian*, *4*(5), 1811–1823. https://doi.org/10.58344/locus.v4i5.4119

Kannitha, D. Z. T., Mustafid, M., & Kartikasari, P. (2022). PEMODELAN TOPIK PADA KELUHAN PELANGGAN MENGGUNAKAN ALGORITMA LATENT DIRICHLET ALLOCATION DALAM MEDIA SOSIAL TWITTER. *Jurnal Gaussian*, *11*(2), 266–277. https://doi.org/10.14710/j.gauss.v11i2.354

Kartika Sari, A., Akhmad Irsyad, Dinda Nur Aini, Islamiyah, & Stephanie Elfriede Ginting. (2024). Analisis Sentimen Twitter Menggunakan Machine Learning untuk Identifikasi Konten Negatif. *Adopsi Teknologi Dan Sistem Informasi (ATASI)*, *3*(1), 64–73. https://doi.org/10.30872/atasi.v3i1.1373

Kulkarni, M., Zhong, A., Guenon des mesnards, N., Movaghati, S., Sridhar, M., Xie, H., & Lu, J. (2023). An Empirical Analysis of Leveraging Knowledge for Low-Resource Task-Oriented Semantic Parsing. *Findings of the Association for Computational Linguistics: ACL 2023*, 1955–1969. https://doi.org/10.18653/v1/2023.findings-acl.123

Kusuma Wardana, R., Cahyono, N., Uyock Anggoro Saputro, Arifiyanto Hadi Negoro, & Nur Aini. (2025). IMPLEMENTASI ALGORITMA SVM DAN OPTIMASI MENGGUNAKAN KOMPARASI N-GRAM DAN GLOVE PADA SENTIMEN PENGGUNA APLIKASI ACCESS BY KAI. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(2), 3216–3223. https://doi.org/10.36040/jati.v9i2.13284

Lathifah, L., Handoyo, E., & Adi Soetrisno, Y. A. (2021). SISTEM CRAWLING DATA INSTRUMEN AKREDITASI BERBASIS SELENIUM DAN PANDAS. *Transient: Jurnal Ilmiah Teknik Elektro*, *10*(1), 84–91. https://doi.org/10.14710/transient.v10i1.84-91

Lum, Y., & Chang, C.-W. (2023). Modeling User Participation in Facebook Live by Applying the Mediating Role of Social Presence. *Information*, *15*(1), 23. https://doi.org/10.3390/info15010023

Luo, Y., & Cui, B. (2024). *Rev Gadget: A Java Deserialization Gadget Chains Discover Tool Based on Reverse Semantics and Taint Analysis* (pp. 229–240). https://doi.org/10.1007/978-3-031-53555-0_22

Muhammad, R. N., & Tanggahma, B. (2024). Pengaruh media sosial pada persepsi publik terhadap sistem peradilan: analisis sentimen di Twitter. *UNES Law Review*, *7*(1), 507–516.

Muhammad Zuama Al Amin, Muhammad Ariful Furqon, & Dwi Wijonarko. (2025). Deteksi Berita Hoaks Berbahasa Indonesia Menggunakan One-Dimensional Convolutional Neural Network. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *14*(2), 161–169. https://doi.org/10.22146/jnteti.v14i2.19050

Nugroho, K., & Hasan, F. N. (2023). Analisis Sentimen Masyarakat Mengenai RUU Perampasan Aset Di Twitter Menggunakan Metode Naïve Bayes. *SMATIKA JURNAL*, *13*(02), 273–283. https://doi.org/10.32664/smatika.v13i02.899

Nurhaliza, A. F., Setiawan, B. D., & Perdana, R. S. (2024). Penerapan Pemodelan Topik menggunakan Metode Latent Dirichlet Allocation terhadap Pembehasan Pemilu Indonesia tahun 2024 di Twitter. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *8*(7).

Purwandari, K., Perdana, R. B., Sigalingging, J. W. C., Rahutomo, R., & Pardamean, B. (2023). Automatic Smart Crawling on Twitter for Weather Information in Indonesia. *Procedia Computer Science*, *227*, 795–804. https://doi.org/10.1016/j.procs.2023.10.585

Putra, F., Tahiyat, H. F., Ihsan, R. M., Rahmaddeni, R., & Efrizoni, L. (2024). Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(1), 273–281. https://doi.org/10.57152/malcom.v4i1.1085

Ramadhan, S., Siswanto, T., & Sari, S. (2025). Sentiment Analysis And Topic Modelling Of Candidate News In The 2024 General Election On Twitter Social Media Using Latent Dirichlet Allocation (LDA) Method. *Intelmatics*, *5*, 33–41. https://doi.org/10.25105/v5i1.21058

Ramamoorthy, T., Kulothungan, V., & Mappillairaju, B. (2024). Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1329185

Rifaldi, D., Abdul Fadlil, & Herman. (2023). Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet "Mental Health." *Decode: Jurnal Pendidikan Teknologi Informasi*, *3*(2), 161–171. https://doi.org/10.51454/decode.v3i2.131

Rofiqi, L., & Akbar, M. (2024). Analisis Sentimen Terkait RUU Perampasan Aset dengan Support Vector Machine. *JEKIN - Jurnal Teknik Informatika*, *4*(3), 529–538. https://doi.org/10.58794/jekin.v4i3.824

Roziewski, S., & Kozłowski, M. (2021). LanguageCrawl: a generic tool for building language models upon common Crawl. *Language Resources and Evaluation*, *55*(4), 1047–1075. https://doi.org/10.1007/s10579-021-09551-7

Sadeghzadehyazdi, N., Batabyal, T., & Acton, S. T. (2021). Modeling spatiotemporal patterns of gait anomaly with a CNN-LSTM deep neural network. *Expert Systems with Applications*, *185*, 115582. https://doi.org/10.1016/j.eswa.2021.115582

Sander, A. (2025). Social Media, Democracy, and the Popular Public Sphere. *Constellations*, *32*(2), 330–342. https://doi.org/10.1111/1467-8675.12773

Sehabudin, S., Irma Purnamasari, A., Bahtiar, A., & Wahyudin, E. (2025). PENINGKATAN MODEL KLASIFIKASI SENTIMEN PUBLIK DI YOUTUBE CNBC INDONESIA TERHADAP MOBIL LISTRIK MENGGUNAKAN ALGORITMA SUPPORT VEKTOR MACHINE. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(4), 6888–6896. https://doi.org/10.36040/jati.v9i4.14184

Setijohatmo, U. T., Rachmat, S., Susilawati, T., & Rahman, Y. (2020). Analisis metoda Latent Dirichlet Allocation untuk klasifikasi dokumen laporan tugas akhir berdasarkan pemodelan topik. *Prosiding Industrial Research Workshop and National Seminar*, *11*(1), 402–408.

Sholekhah, A., & Muntahanah, M. (2025). Perbandingan Naïve Bayes dan Support Vector Machine Dalam Analisa Sentimen Tentang Penyitaan Aset Koruptor di Twitter. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(3), 981–989. https://doi.org/10.57152/malcom.v5i3.2068

Susilo, J. S., Danil, E., & Mulyati, N. (2023). Pemiskinan koruptor sebagai alternatif pidana tambahan dalam pemberantasan tindak pidana korupsi di Indonesia dikaitkan dengan rancangan undang-undangan perampasan aset. *Unes Law Review*, *6*(1), 3718–3730.

Ye, X., Fu, Y.-K., Wang, H., & Zhou, J. (2023). Information asymmetry evaluation in hotel E-commerce market: Dynamics and pricing strategy under pandemic. *Information Processing & Management*, *60*(1), 103117. https://doi.org/10.1016/j.ipm.2022.103117