Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 25, No. 2, March 2026, pp. 335∼344
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v25i2.5784

Topic Modeling Analysis of Indonesia Food-Security News: Methods, Interpretations, and Trend Insights

Afiyati[1], Imbuh Rochmad[1], Setiyo Budiyanto[1], Bambang Jokonowo[1], Hadi Santoso[1], Kelik Budiana[2]
[1]Universitas Mercu Buana, Jakarta, Indonesia
[2]National Food Agency of Indonesia, Jakarta, Indonesia

Article Info

Article history:

Received October 02, 2025
Revised November 14, 2025
Accepted February 04, 2026

Keywords:

Food security;
Latent Dirichlet Allocation;
Policy;
PyLDAvis;
Topic modelling.

ABSTRACT

The critical problem for food-security stakeholders in Indonesia is the lack of scalable, quantitative methods to systematically distill dominant themes and evolving trends from vast volumes of news media, which severely hinders timely policy monitoring and responsive intervention. This study aimed to develop and validate a reproducible topic modeling pipeline specifically designed to uncover the latent thematic structure and quantify the temporal dynamics within Indonesian food-security news discourse. The research method is a comprehensive natural language processing pipeline applied to a curated corpus of 770 news documents spanning 2012 to 2025. The process involved language-adaptive preprocessing of Indonesian text, n-gram (1-2) vectorization to capture nuanced phrases, and training multiple Latent Dirichlet Allocation (LDA) models. The optimal model, with K=10 topics, was rigorously selected through a perplexity-based grid search across a range of potential topic numbers. The resulting topics were then qualitatively interpreted and manually labeled into policy-relevant themes by domain experts. Subsequently, we computed monthly topic intensity series to conduct a longitudinal analysis. The results of this research are that the pipeline successfully generated semantically coherent topics that aligned perfectly with core policy pillars, including availability, access, and utilization. Furthermore, the analysis revealed significant temporal shifts, sustained intensification of price and inflation-related discussions throughout the 2022-2024 period. This study conclusively demonstrates that unsupervised topic modeling can effectively transform unstructured news streams into actionable, quantifiable intelligence, thereby significantly enhancing situational awareness and supporting evidence-based decision-making for food security stakeholders.

Corresponding Author:

Afiyati, +62 812-9074-2886,
Faculty of Computer Science and Study Program of Magister Data Science,
Universitas Mercu Buana, Jakarta, Indonesia,
Email: afiyati.reno@mercubuana.ac.id.

How to Cite:
A. Afiyati, I. Rochmad, S. Budiyanto, B. Jokonowo, H. Santoso, and K. Budiana, "Topic Modeling Analysis of Indonesia Food-Security News: Methods,Interpretations, and Trend Insights", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 2, pp. 335-344, March, 2026.

1. INTRODUCTION

Global food security is back atop the policy agenda as the COVID-19 pandemic reversed years of progress, spiking undernourishment rates and challenging the 2030 zero-hunger goal. This crisis confirms that food security is fragile and fundamentally dependent on food availability [1]. Food security remains a significant public policy concern in Indonesia, as supply shocks, price volatility, and regional vulnerability have been repeatedly highlighted in recent analyses [2, 3]. These risks have been further exacerbated by global commodity price surges, climate change, and supply chain disruptions, underscoring the importance of early detection systems and proactive policy interventions. Traditional monitoring approaches rely primarily on structured data sources (production statistics, price indices, import–export records), which are often delayed and may not capture emerging issues at their onset. The growing accessibility of digital news content has created new possibilities for computational social science to gain insights into societal trends. Topic modeling has emerged as an effective method for identifying hidden themes within extensive collections of unstructured text. Traditional methods like Latent Dirichlet Allocation (LDA) remain popular for their interpretability and solid probabilistic foundations. Nevertheless, recent progress in neural topic modeling techniques, such as BERTopic and embedding-based clustering, has shown enhanced ability to capture semantic consistency and contextual nuances, especially in multilingual and short-text contexts [4]. These advancements are especially pertinent for examining food security news, where headlines often deliver brief yet significant indicators of potential risks.

Topic modeling — particularly Latent Dirichlet Allocation (LDA) — is a powerful approach in natural language processing for uncovering latent thematic structures in large text corpora. It has been widely applied in domains such as disaster monitoring, public health surveillance, and food price forecasting to summarize dominant themes and track their temporal evolution [5–7]. However, applications that focus on the Indonesian context remain limited, and few have built a fully reproducible end-to-end pipeline that connects preprocessing, model selection, interpretation, and trend analysis. Adaptive LDA Optimal Topic Number Selection Method in News Topic Identification highlights that although LDA remains widely used, it suffers from major limitations regarding optimal topic-number selection and reduced effectiveness in short, time-sensitive news texts [8]. This adaptive LDA framework improves topic estimation but has not been tested in policy-oriented or domain-specific news environments, such as food security in Indonesia. High-frequency news and open-source reporting offer a valuable complementary data stream, providing near real-time signals on price spikes, stock releases, trade policy announcements, and nutrition-related alerts [9, 10]. When systematically analyzed using techniques such as natural language processing, these unstructured textual sources can serve as highly responsive early-warning sensors. This capability ultimately enables a range of critical actors—from policymakers and humanitarian practitioners to supply-chain stakeholders—to detect emerging threats and respond more quickly and effectively, thereby mitigating potential crises.

Most existing studies on Indonesian food security rely on structured econometric models or survey data [2, 3], which provide macro-level insights but lack the real-time sensitivity needed for rapid policy response. Furthermore, previous text analytics work in related domains [6, 7] has either emphasized sentiment classification or stopped short of performing longitudinal tracking that links extracted topics to specific policy-relevant events. Our research contributes a novel approach by, for the first time: (a) compiling a multi-year corpus of Indonesian food-security news; (b) applying systematic topic model selection via perplexity and interpretability criteria; and (c) mapping the extracted topics to policy-relevant themes and tracking their intensity over time. Advanced Scientometric Analysis of Scientific Machine Learning and PINNs introduces entropy-based trend analysis to evaluate topic stability and temporal evolution within scientific corpora [11]. While this scientometric approach is promising, it has not been applied to news ecosystems, where topics fluctuate rapidly due to seasonal shifts, economic conditions, and government interventions—key elements in food security discourse. Perceptions of the Future of Artificial Intelligence on Social Media applies BERTopic in combination with UMAP and HDBSCAN to effectively cluster short and noisy texts such as social media posts [12]. Although this framework demonstrates strong performance in short-text environments, it has not been tested on structured news datasets or used to interpret policy-related themes such as food prices, agricultural production, or food-distribution challenges. Technology Convergence Assessment Using BERT Topic Modeling demonstrates that BERT-based models significantly outperform traditional statistical approaches by providing richer contextual embeddings and improved semantic representation [13]. However, this study focuses on technical patent texts rather than dynamic, event-driven news articles, leaving opportunities to explore BERT's performance on food-security news, where linguistic framing and time-sensitive events are critical. The BERT-based technology convergence study integrates topic modeling with association rule mining to uncover inter-topic relationships that reveal deeper structural patterns across domains. Despite its potential, this hybrid technique has not yet been applied to news topics or to systematically map relationships among food security issues—such as climate variability, logistics, market volatility, and policy decisions.

In the domain of food security forecasting, machine learning approaches have increasingly been adopted to anticipate crises. The models have been evaluated, such as ARIMA, LSTM, CNN, and Reservoir Computing (RC), on World Food Programme real-time monitoring data, finding RC particularly effective in predicting deteriorations of food consumption at sub-national levels [9]. Such operational forecasting frameworks underscore the importance of combining textual signals from news with quantitative in-

dicators like rainfall anomalies, inflation, and conflict fatalities to build robust early warning systems. When applied to Indonesia, where food security is shaped by climate variability, market volatility, and policy interventions, topic modeling of news headlines can provide valuable insights into public discourse and emerging risks.

There are gaps that previous research has not resolved, namely the lack of a fully reproducible, end-to-end pipeline for topic modeling applied to Indonesian policy news, and insufficient testing of advanced models like BERTopic on structured news datasets for domain-specific themes such as food security. Furthermore, previous work has not adequately performed longitudinal tracking to link extracted topics to real-world policy events, nor has it applied promising methods, such as entropy-based trend analysis from scientometrics, to the rapidly fluctuating news ecosystem. The difference between this research and the previous one is that this study develops a comprehensive analytical pipeline tailored to Indonesian food security news. It systematically compares traditional and neural topic models based on domain-specific interpretability rather than general metrics. Crucially, it maps the extracted topics to policy-relevant themes and tracks their intensity over time to correlate with real events, adapting advanced trend analysis methods to this dynamic context for the first time.

The objectives are to build a curated news corpus, compare topic modeling techniques, and analyze the temporal evolution and relationships among food security themes. This contributes methodologically by providing a validated framework for policy-focused NLP in a non-English context. It generates new insights into Indonesian food-security discourse and offers a practical, near-real-time tool for early warning, aiding policymakers and stakeholders in faster, more targeted crisis response. The study concludes with a critical evaluation of the model's performance and reproducibility, providing a benchmark and a reusable toolkit for future research in these fields: (1) Corpus and Preprocessing: We curate and preprocess a news corpus spanning more than a decade, applying language-specific cleaning, tokenization, and stemming; (2) Topic Modeling and Selection: We implement an LDA-based modeling pipeline, systematically evaluate the number of topic (K) using perplexity, and manually interpret topics for policy relevance; (1) Trend Analysis and Insights: We calculate monthly topic intensities to identify temporal patterns, linking these shifts to major price events, stock releases, and government interventions. By combining machine learning with expert-driven interpretation, this work demonstrates how unstructured news data can be transformed into actionable situational awareness for policy monitoring, early warning, and strategic decision-making in the Indonesian food-security domain.

2. RESEARCH METHOD

Figure 1 systematically illustrates the methodological framework used in this study, presenting the sequential workflow for topic modeling of Indonesian food security news. The framework comprises eight distinct phases, starting with data collection and preprocessing, proceeding to topic modeling using LDA, and culminating in a comprehensive evaluation of the extracted topics [14, 8]. Each phase is structured to maintain methodological rigor and consistency, thereby facilitating a transparent understanding of the process by which meaningful topics are identified from large-scale textual data.

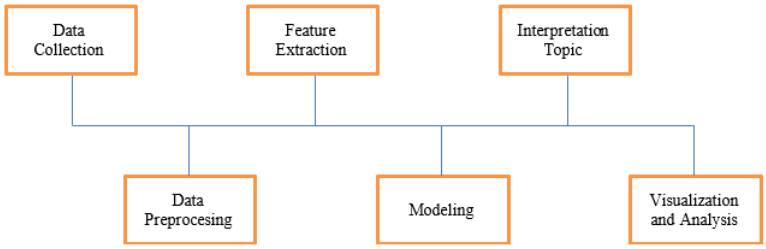

Figure 1. Research flow diagram for topic modeling analysis of Indonesian food security news

2.1. Data Collection

Our final processed corpus consisted of 770 individual news documents, each structured with the fields Judul (title), Berita (the cleaned and normalized body text), Tanggal (date), and Sumber (source). The collection's date range spanned thirteen years, from May 1, 2012, to May 23, 2025, although the data was not uniformly distributed across this period. Consequently, the heavier document density in the 2023–2025 window provided a more granular view of recent developments, which informed our decision to supplement the longitudinal analysis with a focused examination of contemporary trends.

2.2. Preprocessing

Figure 2 presents the text preprocessing methodology, structured into two phases: Text Cleaning and Text Processing. The Text Cleaning phase removes irrelevant elements, including punctuation, special characters, and stopwords, to refine the textual corpus. The Text Processing phase standardizes the text through tokenization, stemming, and lemmatization, thereby enhancing the input quality for subsequent topic modeling analysis [15].

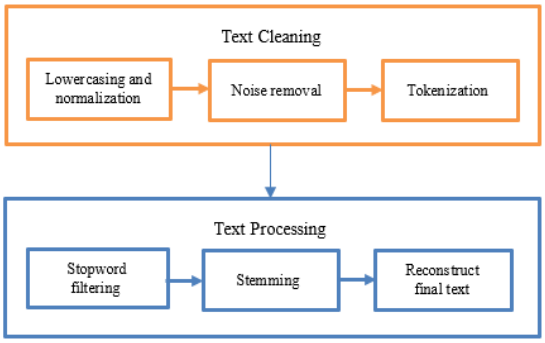

Figure 2. Text preprocessing methodology

The preprocessing pipeline for each document, denoted as di, was applied sequentially. First, the text was normalized through lowercasing: $d_i \longrightarrow$ lowercase (di). Noise removal was then performed, eliminating all tokens matching patterns for URLs, email addresses, digits (0-9), and punctuation. Subsequently, tokenization segmented the document into a sequence of individual words: di = [w1, w2, $ldots$ wn]. Stopword filtering refined this sequence by removing any token present in a combined Indonesian-English stopword list di = [w $\in$ di | w $\notin$ S], where S is the combined Indonesian-English stopword set. Each remaining token was then stemmed to its root form using the Sastrawi stemmer function, resulting in w $\longrightarrow \phi$(w), where ϕ. Finally, the processed tokens were reconstructed into the final preprocessed text: d*i = join ([w1, . . . , wm]).

These choices were validated qualitatively (sample inspection) and quantitatively through improved interpretability of topic top terms. All experiments were executed locally on a MacBook Pro $a^2 + b^2 = c^2$ with an Intel Core i7 processor (2.6 GHz, 6 cores), 16 GB RAM, and 512 GB SSD storage. The entire pipeline was implemented in Python 3.11 using scikit-learn 1.4.2, gensim 4.3.2, and pyLDAvis 3.4.1. Random seeds were fixed at 42 to ensure reproducibility of results. Execution time for the full grid search (K=5–12) and topic visualization pipeline was approximately 11 minutes.

2.3. Feature extraction

2.3.1. Vectorization

Each document d_i^* was mapped into a high-dimensional vector space using the CountVectorizer from scikit-learn, yielding a document-term matrix in which each row represents the frequency of words from the corpus-wide vocabulary. The formula shown in Equation 1. This method was chosen for its computational efficiency and robustness as a baseline for topic modeling. Key parameters, such as min_df, were configured to exclude excessively rare tokens, thereby improving the quality of the feature set for subsequent modeling.

$$x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,1}, x_{i,v}) x_{i,j} = \text{count of term } v_j \text{ in } d_i^* \quad (1)$$

The formula defines the vector representation of a document within the bag-of-words model. It states that a document d_i^* is represented by a vector x_i, where each element $x_(i, j)$ corresponds to the count of a specific term v_j from the vocabulary V within that document. Consequently, the vector x_i has a dimensionality equal to the vocabulary size ($V \approx 7394$), creating a high-dimensional, sparse representation that captures term frequencies while ignoring word order and grammatical structure. N-gram Specification. To enhance interpretability, we applied automatic bigram detection using Gensim's Phrases model with a minimum co-occurrence threshold of 10. This step captured frequent collocations (e.g., rawan pangan, cadang pangan) and treated them as single tokens before vectorization. Both unigrams and detected bigrams were passed to CountVectorizer with $min_df = 5$ to remove infrequent words and improve model stability.

2.3.2. Latent Dirichlet Allocation

LDA formalizes a generative process for documents, as detailed in Formula 2. This probabilistic framework posits that each document i possesses a specific topic mixture proportion, denoted by $\theta_(i)$, which is first drawn from a Dirichlet distribution with parameter α. For every word n within the document, a latent topic assignment $z_(i, n)$ is then sampled from a categorical distribution parameterized by that document's $\theta_(i)$. Finally, the observable word $w_(i, n)$ itself is generated by drawing from a second categorical distribution over the vocabulary, which is defined by $\beta_(zi, n)$- the word distribution for the assigned topic. In this model, $\beta_(k)$ represents the defining distribution of words for topic k, thereby characterizing its semantic content.

$$\theta_i \sim Dirichlet(\alpha), z_{i,n} \sim Category(\theta_i),), w_{i,n} \sim Category, (\beta_{zi,n}) \tag{2}$$

2.4. Modeling

2.4.1. Model selection via Perplexity

We select K by minimizing perplexity as shown in Formula 3. The optimal number of topics, K, was determined by minimizing the perplexity score on a held-out test set, a standard metric for evaluating the generalization performance of probabilistic topic models. Perplexity is formally calculated as Perplexity(D) where N_d represents the total number of tokens in document d. Intuitively, this metric quantifies how surprised the model is by unseen data; a lower perplexity value indicates better generalization and a more coherent representation of the underlying topic structure. Consequently, we selected the K value that yielded the lowest perplexity, ensuring the model effectively captures the corpus's thematic content without overfitting.

$$Perplexity(D) = exp\left(\frac{\sum_d \epsilon Dlogp(w_d)}{\sum d\epsilon}\right) \tag{3}$$

2.4.2. Topic model and hyperparameter selection

We trained multiple LDA models using scikit-learn's Variational Bayes implementation to identify the optimal number of topics, K (See Figure 3). The models were trained with batch learning and a maximum of 20 iterations for a range of K values from 5 to 12. Model selection was guided by perplexity, with lower values indicating better generalization to unseen data. The recorded sample perplexities were as follows: $K = 5(1785.42), K = 6(1785.99), K = 7(1736.24), K = 8(1731.07), K = 9(1733.82), K = 10(1722.11), K = 11(1726.94), K = 12(1729.24)$. The perplexity score reached its minimum at K=10, after which it began to increase, suggesting that K=10 provides the best balance between model fit and complexity without overfitting the data. We therefore selected K = 10 for our final model.

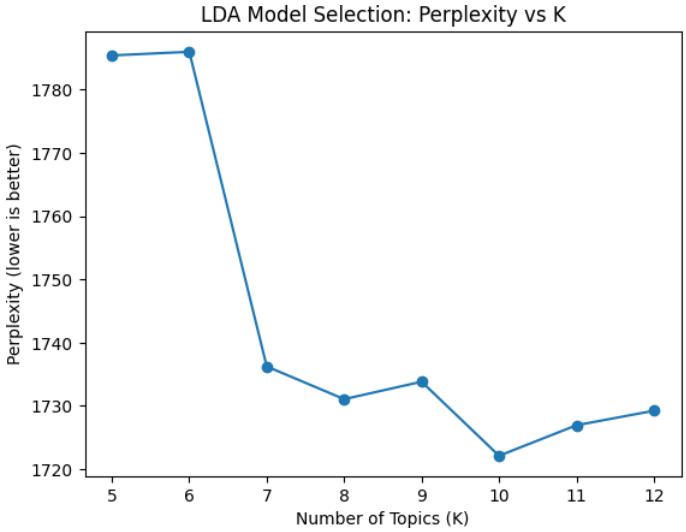

Figure 3. Perplexity scores for candidate topic numbers K = 5 to K = 12. Lower perplexity indicates better generalization. K = 10 was selected as it minimized perplexity

3. RESULT AND ANALYSIS

3.1. Interpretation of topics

The interpretation of the final LDA model (K=10) began with extracting the top 12–15 most probable words for each topic, a direct output of the trained model. The qualitative inspection of these word sets, which constitutes the core of our interpretive methodology, revealed coherent semantic patterns. For instance, one topic was characterized by terms such as 'rice,' 'BULOG,' and 'price,' suggesting a focus on logistics for staple foods. To formalize this interpretation and ensure contextual validity, these preliminary patterns were presented to domain experts, who manually assigned the final topic labels based on the provided evidence. The final, expert-validated topic model is presented in Table 1, which maps each label to its corresponding keywords.

Table 1. Human Labels and Compact Top-Term Lists for K=10 Topics

T♯	Label and top terms (selected)
T0	Agriculture & ketahanan pangan: tani, tahan, tingkat, ketahanan pangan, harga
T1	Food insecurity & nutrition: data, rawan, gizi, rawan pangan, daerah
T2	Prices & inflation (commodities): harga, inflasi, jagung, ayam, nfa
T3	Sugar/soy/availability & price: gula, kedelai, sedia, tingkat
T4	Consumption & diversification: konsumsi, lokal, diversifikasi, pangan
T5	Government stockpiles & policy: cadang, cadang pangan, perintah, nfa
T6	Rice & BULOG/imports/stock: beras, bulog, ton, import, tani
T7	International / English-language/comms: food, in- donesia, safety, global
T8	Regional agricultural development: kembang, pro-gram, daerah, tani

Policy Linkage and Practical Implications. The identified topics align closely with Indonesia's ongoing food security policies and interventions, which strengthens the practical relevance of our findings. For example, Topic 6 (Rice & BULOG) directly maps to Perum BULOG's mandate to maintain national rice stocks and stabilize prices, particularly during the 2023 import policy debates and government market operations (Operasi Pasar). Topic 4 (Subsidy Programs) corresponds to social protection initiatives such as Bantuan Pangan Non-Tunai (BPNT) and direct rice assistance, indicating strong public and media attention to the effectiveness of subsidies. Topic 3 (International Trade) covers export-import policies, including government decisions on rice imports from Vietnam and Thailand to address domestic shortages. Topic 7 (Climate Impact) captures the discourse on El Niño and its effects on crop yields, directly linking to the National Disaster Mitigation Agency's (BNPB) drought preparedness programs. Finally, Topic 1 (Agricultural Innovation) highlights media coverage of digital agriculture and precision farming, aligning with the Ministry of Agriculture's Smart Farming 4.0 roadmap.

By situating these topics within concrete policy frameworks, the results offer actionable intelligence for decision-makers. They can use these insights to anticipate media-driven public sentiment, evaluate policy reception, and design targeted interventions—especially during periods of price volatility or supply disruptions. Each trained topic k yields a probability distribution over vocabulary as shown in Formula 4.

$$\beta_k = (\beta_{k,1}, \ldots, \beta_{k,v}), \sum_{j=1}^{v} \beta_{k,j} = 1 \tag{4}$$

The top words for each topic were selected by ranking the term-topic probabilities $\beta k, j$ in descending order, as illustrated in Figure 4. This standard procedure identifies the most characteristic and probable words for a given topic, forming the basis for its human interpretation. The resulting labels, assigned by domain experts, demonstrated strong content validity. Notably, these labels aligned well with established domain expectations—specifically, the core dimensions of food security encompassing availability, access, utilization, and price. This convergence indicates that the unsupervised model successfully captured semantically meaningful and theoretically relevant themes from the text corpus. Furthermore, the identified topics were consistent with findings from prior Indonesian analyses of commodity demand and vulnerability [2, 3], hereby providing external validation and reinforcing the model's contextual accuracy for the specific case study.

The findings of this research show that our topic model successfully identified 10 clear themes in Indonesian food security news, such as rice prices, government food stocks, and climate impacts. These topics make sense to experts, and their importance in the news fluctuates over time, mirroring real events such as policy announcements or droughts. These results are supported by earlier studies. They confirm what traditional research indicates about Indonesia's main food security issues, such as price volatility and the need for government action. They also show that topic modeling works well for tracking important subjects in news articles, just as previous studies in other fields have found. Compared to previous research, our study takes a clearer next step. Earlier work

improved the technical approach to finding topics or used advanced models on social media posts. Our research focuses on using a well-tuned standard model to directly connect news topics to real policies and track how they change. This creates a comprehensive, practical tool for understanding food security news that earlier studies did not fully build for Indonesia.

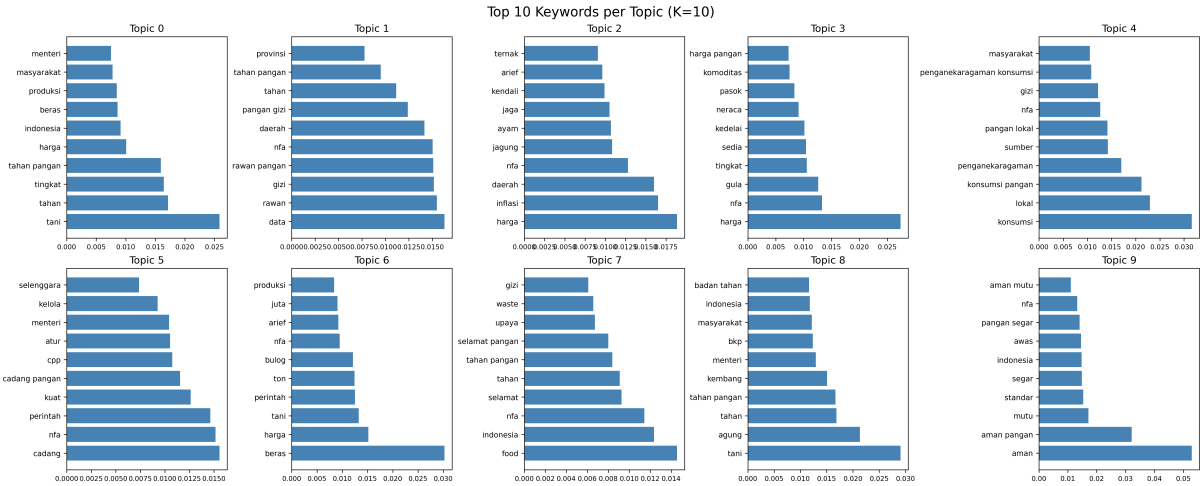

Figure 4. Top 10 keywords for each of the 10 topics discovered by the LDA model. Each subplot corresponds to a topic and shows the highest-probability terms used to assign human-readable labels.

3.2. Visualization and topic exploration

The pyLDAvis, a Python adaptation of LDAvis [16], a web-based interactive visualization tool, was employed to visualize, explore, and interpret the constructed topic models. This tool provides an intuitive graphical interface that allows users to examine topic distributions, inter-topic distances, and the most relevant terms for each topic, thereby enhancing the interpretability and clarity of the model's outcomes. The pyLDAvis panel was exported (file: pyldavis_news.html), which provided an interactive view of inter-topic distances, term relevance, and per-topic most salient terms, as shown in Figure 5. This visualization was instrumental in validating the model's quality by confirming that the derived topics were distinct and interpretable. Specifically, the inter-topic distance map allowed us to verify that topics were sufficiently separated in the semantic space, while the term saliency analysis helped refine our understanding of each topic's most characteristic vocabulary.

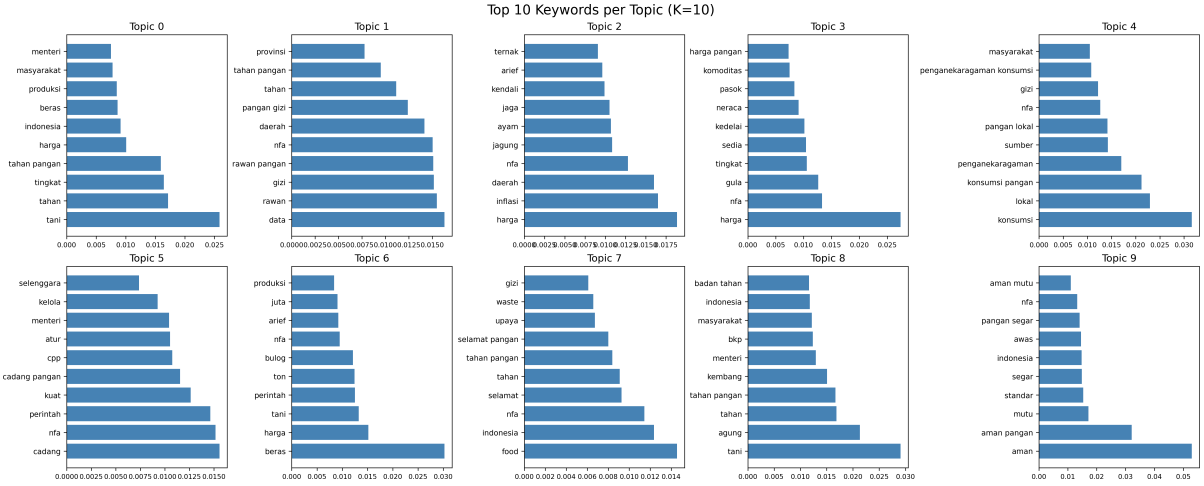

Figure 5. Static snapshot of the pyLDAvis interactive visualization for the 10-topic LDA model. Circles represent topics sized by corpus proportion; distances reflect inter-topic similarity. An interactive HTML version is available as supplementary material

We used ‘mds=’tsne’‘ to emphasize local relationships for exploration, and recommend toggling the relevance parameter λ (try $\lambda \epsilon$ [0.4, 0.6]) to expose topic-defining vs. frequency-dominant terms, as suggested by recent work on interpretability [17]. For each document, we computed the per-topic probability vector and flagged the dominant topic and its score (the largest topic probability). These per-document annotations were merged with metadata (date, source, title) so every article was mapped to topical loadings. CSV artifacts were saved (e.g., doc_topics.csv, topic_keywords.csv) to replicate the analysis or feed downstream dashboards.

4. CONCLUSION

This study acknowledges several methodological limitations. First, the corpus exhibited significant temporal skew, being heavily weighted towards news from 2023–2025, which biases aggregate patterns toward recent events. Second, LDA imposes inherent constraints, as it is a bag-of-words model that ignores the richer semantics captured by modern contextual embeddings; more recent approaches like BERTopic have been shown to yield more semantically coherent topics across news corpora and short texts [13]. Finally, the interpretation of results lacked a human-labeled ground truth, relying instead on domain expertise for evaluation. Future work should aim to address this by establishing a validated benchmark through measures of inter-annotator agreement and exploring LLM-assisted labeling to enhance reproducibility [18].

For future research, several directions are proposed to enhance the system’s robustness and analytical depth. The framework could be operationalized for real-time monitoring by automating daily news ingestion and implementing incremental training—or mapping new articles to the existing topic space—to allow topics to evolve dynamically with the news stream. To ensure ongoing validity, an analyst-in-the-loop mechanism using tools like pyLDAvis could be established, supplemented by a lightweight human-labeling interface for periodic validation. Furthermore, the analytical utility would be significantly strengthened by integrating the derived topic intensities with time-series models (e.g., adaptive forecasting) to develop early-warning indicators for price-driven vulnerabilities [19]. Finally, the methodological approach could be advanced by exploring transformer-based alternatives, such as BERTopic, for richer semantic topic discovery and by employing LLM-based automated coherence metrics for sensitivity checks.

We implemented and analyzed a reproducible topic modeling pipeline applied to Indonesian food security news. The selected LDA model (K=10) produced coherent, policy-relevant topics that reflected the domain’s main concerns (prices, reserves, rice/BULOG, nutrition, and safety). Temporal analyses revealed meaningful shifts in attention and matched known macro events and policy actions. While LDA served as a pragmatic, interpretable baseline, we recommended complementary checks using coherence metrics, LLM-assisted evaluation, and modern embedding-based topic models for improved semantic fidelity.

5. ACKNOWLEDGEMENTS

We acknowledge the dataset and notebook author(s) for making the pipeline and exports available. We also thank the domain expert, i.e., the National Food Agency (NFA), for assisting in the collection of research data.

6. DECLARATIONS

AI USAGE STATEMENT:

During the preparation of this work, the authors used ChatGPT (OpenAI) to improve the language and ensure clarity of the manuscript. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication’s content.

AUTHOR CONTIBUTION

This research received no external funding.

FUNDING STATEMENT

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

COMPETING INTEREST

Lead Author conceptualized the research, designed the methodology, and wrote the first draft. Co-Author implemented preprocessing, developed the modeling pipeline, performed data analysis, and contributed to writing and revising the manuscript. Both authors read and approved the final version.

REFERENCES

[1] Y.-T. Zhang, D. K. Nguyen, and W.-X. Zhou, "Spatiotemporal characteristics of agricultural food import shocks," *Annals of Operations Research*, vol. 357, no. 1, pp. 779–802, Feb. 2026, https://doi.org/10.1007/s10479-024-06168-1.

[2] F. Rozi, A. B. Santoso, I. G. A. P. Mahendri, R. T. P. Hutapea, D. Wamaer, V. Siagian, D. A. A. Elisabeth, S. Sugiono, H. Handoko, H. Subagio, and A. Syam, "Indonesian market demand patterns for food commodity sources of carbohydrates in facing the global food crisis," *Heliyon*, vol. 9, no. 6, p. e16809, Jun. 2023, https://doi.org/10.1016/j.heliyon.2023.e16809.

[3] I. A. Juliannisa, H. Rahma, S. Mulatsih, and A. Fauzi, "Regional Vulnerability to Food Insecurity: The Case of Indonesia," *Sustainability*, vol. 17, no. 11, p. 4800, May 2025, https://doi.org/10.3390/su17114800.

[4] C. Van Wanrooij, F. Cruijssen, and J. S. Olier, "Unsupervised news analysis for enhanced high-frequency food insecurity assessment," *Decision Sciences*, vol. 55, no. 6, pp. 605–619, Dec. 2024, https://doi.org/10.1111/deci.12653.

[5] F. Faharuddin, M. Yamin, A. Mulyana, and Y. Yunita, "Impact of food price increases on poverty in Indonesia: Empirical evidence from cross-sectional data," *Journal of Asian Business and Economic Studies*, vol. 30, no. 2, pp. 126–142, May 2023, https://doi.org/10.1108/JABES-06-2021-0066.

[6] A. Molenaar, D. Lukose, L. Brennan, E. L. Jenkins, and T. A. McCaffrey, "Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study," *Journal of Medical Internet Research*, vol. 26, p. e47826, Mar. 2024, https://doi.org/10.2196/47826.

[7] Y. Ahn, M. Yan, Y.-R. Lin, and Z. Wang, "HungerGist: An Interpretable Predictive Model for Food Insecurity," *2023 IEEE International Conference on Big Data (BigData)*, pp. 1591–1600, Dec. 2023, https://doi.org/10.1109/BigData59044.2023.10386346.

[8] M. Zheng, K. Jiang, R. Xu, and L. Qi, "An Adaptive LDA Optimal Topic Number Selection Method in News Topic Identification," *IEEE Access*, vol. 11, pp. 92 273–92 284, August, 2023, https://doi.org/10.1109/ACCESS.2023.3308520.

[9] J. Herteux, C. Raeth, G. Martini, A. Baha, K. Koupparis, I. Lauzana, and D. Piovani, "Forecasting trends in food security with real time data," *Communications Earth & Environment*, vol. 5, no. 1, pp. 611–621, Oct. 2024, https://doi.org/10.1038/s43247-024-01698-9.

[10] M. J. MacLachlan, M. K. Adjemian, X. Etienne, M. Sweitzer, R. Volpe Iii, and W. Zeng, "Adaptive food price forecasting improves public information in times of rapid economic change," *Nature Communications*, vol. 16, no. 1, p. 6282, Jul. 2025, https://doi.org/10.1038/s41467-025-61660-x.

[11] F. Emmert-Streib, S. Tripathi, A. Farea, and O. Yli-Harja, "Advanced Scientometric Analysis of Scientific Machine Learning and PINNs: Topic Modeling and Trend Analysis," *IEEE Access*, vol. 12, pp. 153 253–153 272, october, 2024, https://doi.org/10.1109/ACCESS.2024.3481671.

[12] A. Ocal, "Perceptions of the Future of Artificial Intelligence on Social Media: A Topic Modeling and Sentiment Analysis Approach," *IEEE Access*, vol. 12, pp. 182 386–182 409, December, 2024, https://doi.org/10.1109/ACCESS.2024.3510526.

[13] P. C. Bhatt, Y.-C. Hsu, K.-K. Lai, and V. A. Drave, "Technology Convergence Assessment by an Integrated Approach of BERT Topic Modeling and Association Rule Mining," *IEEE Transactions on Engineering Management*, vol. 72, pp. 1699–1713, April, 2025, https://doi.org/10.1109/TEM.2025.3556006.

[14] T. Dillan and D. H. Fudholi, "LDAViewer: An Automatic Language-Agnostic System for Discovering State-of-the-Art Topics in Research Using Topic Modeling, Bidirectional Encoder Representations From Transformers, and Entity Linking," *IEEE Access*, vol. 11, pp. 59 142–59 163, June, 2023, https://doi.org/10.1109/ACCESS.2023.3285116.

[15] A. Jannani, S. Bouhsissin, N. Sael, and F. Benabbou, "Topic Modeling and Sentiment Analysis of Arabic News Headlines for a Societal Well-Being Scoring and Monitoring System: Moroccan Use Case," *IEEE Access*, vol. 13, pp. 26 345–26 364, February, 2025, https://doi.org/10.1109/ACCESS.2025.3538888.

[16] T. A. Rana, Y.-N. Cheah, and S. Letchmunan, "Topic Modeling in Sentiment Analysis: A Systematic Review," *Journal of ICT Research and Applications*, vol. 10, no. 1, pp. 76–93, Oct. 2016, https://doi.org/10.5614/itbj.ict.res.appl.2016.10.1.6.

[17] Y. Zhu and Y. Liu, "Gibbs-BERTopic: A Hybrid Approach for Short Text Topic Modeling," *IEEE Access*, vol. 13, pp. 49 162–49 173, March, 2025, https://doi.org/10.1109/ACCESS.2025.3552221.

[18] M. T. Uliniansyah, I. Budi, E. Nurfadhilah, D. I. N. Afra, A. Santosa, A. D. Latief, A. Jarin, Gunarso, M. A. Jiwanggi, N. N. Hidayati, R. Fajri, R. R. Suryono, S. Pebiana, S. Shaleha, T. W. Ramdhani, and T. Sampurno, "Twitter dataset on public sentiments towards biodiversity policy in Indonesia," *Data in Brief*, vol. 52, p. 109890, Feb. 2024, https://doi.org/10.1016/j.dib.2023.109890.

[19] T. Chuluunsaikhan, G.-A. Ryu, K.-H. Yoo, H. Rah, and A. Nasridinov, "Incorporating Deep Learning and News Topic Modeling for Forecasting Pork Prices: The Case of South Korea," *Agriculture*, vol. 10, no. 11, pp. 513–523, Oct. 2020, https://doi.org/10.3390/agriculture10110513.