InComTech: Jurnal Telekomunikasi dan Komputer
vol.16, no.2, Agustus 2026, 128-142
http://publikasi.mercubuana.ac.id/index.php/Incomtech
P-ISSN: 2085-4811 E-ISSN: 2579-6089

Topic Modeling of Indonesian News Articles on The Free Nutritious Meal Program: A Comparative Study of LDA, LSA, and NMF

Irmma Dwijayanti[1*], Muhammad Habibi[2]

[1]*Sistem Informasi, Universitas Jenderal Achmad Yani Yogyakarta, Jl. Siliwangi, Ringroad Barat, Sleman 55293, Indonesia*
[2]*Informatika, Universitas Jenderal Achmad Yani Yogyakarta, Jl. Siliwangi, Ringroad Barat, Sleman 55293, Indonesia*
*Email Penulis Koresponden: irmmadwijayanti@gmail.com

***Abstract*:**
The rapid expansion of digital journalism in Indonesia has resulted in a massive volume of long-form online news, posing significant challenges for large-scale information analysis and media monitoring. From an information and communication technology (ICT) perspective, this situation highlights the need for automated text analytics methods that are scalable, interpretable, and suitable for complex news streams. This study evaluates three classical topic modeling techniques—Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF)—for Indonesian online news analytics. The Free Nutritious Meal Program (Makan Bergizi Gratis, MBG) is employed as a real-world case study to represent sustained and multifaceted public discourse in digital media. The news corpus was processed using a standardized text analytics pipeline consisting of tokenization, stop-word removal, stemming, and normalization, followed by experiments with multiple topic configurations. Model performance was assessed using topic coherence scores and qualitative interpretability. The results show that NMF achieves the highest coherence and produces well-separated topic structures, while LDA offers stronger semantic interpretability through probabilistic topic representations. LSA demonstrates lower computational complexity but generates broader and less detailed themes. Temporal analysis further indicates that topic modeling is effective in detecting shifts in media attention over time, with a notable increase in MBG-related discourse beginning in 2024 and continuing into 2025. Overall, the findings demonstrate the applicability of classical topic modeling techniques for large-scale Indonesian news analytics and support their use in ICT-based media monitoring and decision support systems.

This is an open access article under the CC BY-NC license

***Kata Kunci*:**
Topic Modeling;
Indonesian News;
LDA;
LSA;
NMF

Riwayat Artikel:
Received Oct 31, 2025
Revised Jan 13, 2026
Accepted Jan 23, 2026

DOI:
10.22441/incomtech.v16i2.36809

1. INTRODUCTION

The rapid expansion of internet access has significantly increased digital news consumption in Indonesia. As of 2024, the Indonesian Internet Service Providers Association (APJII) reported 221.56 million internet users, covering 79.5% of the population [1]. As a result, online media has become a primary channel for public information dissemination. At the same time, the growing volume of digital content—particularly long-form news articles containing detailed reporting and analysis—has introduced challenges related to information overload. From an information and communication technology (ICT) perspective, this condition necessitates automated text analytics techniques capable of processing and summarizing large-scale news data efficiently.

Topic modeling has emerged as an unsupervised text analytics approach for discovering latent thematic structures within large document collections. Classical methods such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF) are widely used in large-scale text analysis [2]. LDA models documents as probabilistic mixtures of topics, LSA applies singular value decomposition to uncover latent semantic structures, and NMF decomposes text representations into additive and interpretable components. Comparative studies have shown that no single method consistently outperforms others across all datasets, as performance depends on corpus characteristics, preprocessing strategies, and evaluation criteria [3], [4], [5], [6], [7].

Recent research continues to refine these classical topic modeling techniques. Empirical studies demonstrate that preprocessing choices strongly influence topic coherence and interpretability [8]. Advances in NMF, including deep architectures and optimization using β-divergences, have improved topic separability and stability [6], [9]. In the context of Indonesian-language text, recent work has explored embedding-based approaches such as BERTopic while still employing LDA and NMF as baseline models for comparison [10]. Multilingual studies further confirm that linguistic properties play an important role in determining model effectiveness, particularly for morphologically rich languages [11]. However, progress in Indonesian natural language processing has largely focused on supervised tasks using models such as IndoBERT and IndoNLU , leaving limited empirical investigation of unsupervised topic modeling for long-form Indonesian news articles.

To address this gap, this study conducts a systematic comparison of LDA, LSA, and NMF for large-scale Indonesian online news analytics. The Free Nutritious Meal Program (Makan Bergizi Gratis, MBG) is employed as a real-world case study representing sustained and complex public discourse in digital media. Rather than focusing on policy evaluation, the MBG corpus serves as a testbed for assessing topic modeling performance on long-running news streams. The contributions of this paper are threefold: (1) a comparative evaluation of classical topic modeling techniques on Indonesian long-form news text, (2) a thematic mapping of large-scale media discourse using unsupervised text analytics, and (3)

practical recommendations for preprocessing strategies based on quantitative and qualitative analysis [3].

2. METHOD

The proposed method is designed as an ICT-based text analytics pipeline for scalable processing and analysis of large-scale Indonesian online news data. The study follows a sequential workflow. First is data acquisition by crawling major Indonesian news portals. Second is text preprocessing at a high level, which prepares the text for modeling. Third is topic modeling using Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF). Fourth is model evaluation using coherence and qualitative interpretability. Finally, topic analysis map's dominant themes and characterizes the framing of the news coverage. The complete pipeline is shown in Figure 1.

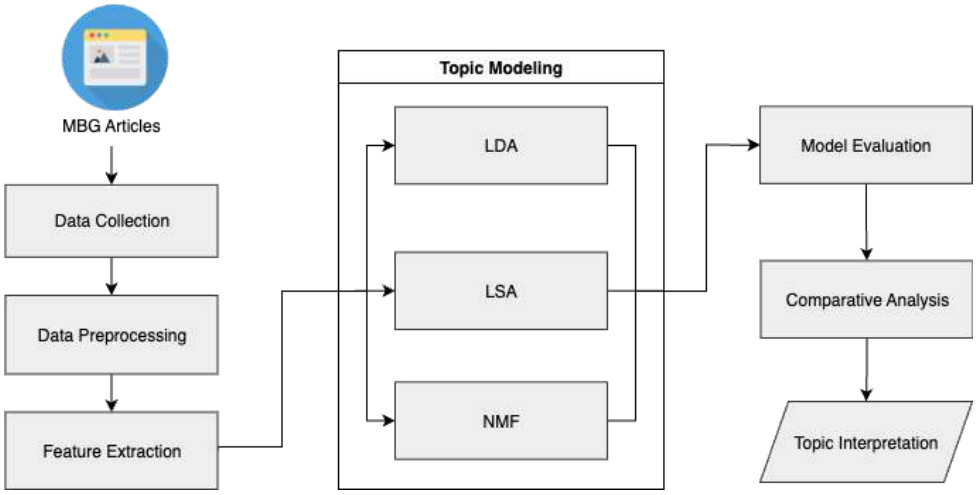

Figure 1. Research Stage

2.1 Data Collection

This data collection stage represents the ingestion layer of the analytics pipeline, enabling scalable acquisition of online news data. Data were collected from multiple Indonesian-language online news portals using web crawling with the query terms "makan bergizi gratis" and "mbg." Collection imposed no lower bound on publication date; all matching articles available through 4 September 2025 were included. This yielded 5,544 articles from seven portals. Table 1 reports counts and proportions.

Table 1. Data Distribution

No.	News portal	Count	Percentage
1	kompas	2170	39,14%
2	antaranews	1112	20,06%
3	metrotvnews	677	12,21%
4	detik	663	11,96%
5	okezone	348	6,28%
6	tribunnews	290	5,23%
7	cnbcindonesia	284	5,12%

2.2 Data Preprocessing

We use a light, three-step preprocessing pipeline to prepare the news articles for topic modeling.

- Tokenization, split text into tokens so terms can be counted and compared consistently. See a recent survey that details tokenization as a core step and discusses when to retain multi-word expressions; also, a scoping review that lists tokenization among the most common methods [12], [13], [14], [15].
- Punctuation and stray character cleanup, removing punctuation and non-letter artifacts reduces noise and standardizes the vocabulary before modeling. Both the Natural Language Engineering (NLE) survey and the scoping review document punctuation removal as standard practice [16], [17].
- Stop-word removal, filtering very common function words helps the model focus on content terms; journal work shows stop-word handling is a key preprocessing choice that can affect model outcomes. See the NLE survey, a PLOS ONE article focused on stopwords in technical language, and this study showing that preprocessing choices materially influence model performance [17], [18], [19].

This minimal setup keeps the text clean enough for modeling while preserving important terms and names that help define the topics.

2.3 Feature Extraction

Feature extraction turns cleaned text into numeric vectors that models can use. We represent each document with a term–document matrix whose cell values reflect how informative each term is for that document relative to the whole collection. Recent studies show that weighting schemes which reward document-specific terms and discount ubiquitous ones tend to improve separability for downstream modeling and classification [20], [21].

We use term frequency–inverse document frequency to weight terms. TF-IDF increases the weight of terms that appear often in a document but remain uncommon across the corpus, which helps highlight content words and suppress near-stop terms [21].

Let N be the number of documents, df(t) the number of documents containing term t, and tf(t,d) the frequency of term ttt in document d.

Inverse document frequency using equation (1):

$$\text{idf}(t) = \log\left(\frac{N}{df(t)}\right) \tag{1}$$

A widely used smoothed variant avoids zero divisions and stabilizes weights, equation (2):

$$\text{idf}_{\text{smooth}}(t) = \log\left(\frac{1+N}{1+df(t)}\right) + 1 \tag{2}$$

TF-IDF weight, equation (3):

$$\text{tfidf}(t,d) = tf(t,d) \times \text{idf}(t) \tag{3}$$

Modern applications often normalize the resulting vectors, so documents of different lengths remain comparable. Empirical work in the past five years continues to validate TF-IDF and its variants for topic modeling pipelines and text classification, including modified IDF formulations and n-gram extensions [22].

2.4 Topic Modeling

Topic modeling discovers latent themes in large text collections using unsupervised learning. In practice, a document–term matrix X (often TF–IDF weighted) is decomposed into a small set of "topics" and document–topic weights or is modeled probabilistically as mixtures over topics. Comparative studies in the last five years show that performance depends on the corpus and evaluation goals; no single method dominates across all settings [7], [23].

2.4.1 Latent Dirichlet Allocation (LDA)

LDA treats each document as a mixture of topics and each topic as a probability distribution over words. It is valued for interpretability because topic–word probabilities can be inspected directly and document–topic mixtures reflect topical breadth. Recent comparative work continues to use LDA as a transparent baseline against newer models [3], [7].

For document d and topic k, equation (4):

$$\begin{aligned} \theta_d &\sim \mathrm{Dir}(\alpha), \quad \phi_k \sim \mathrm{Dir}(\eta), \\ z_{dn} &\sim \mathrm{Cat}(\theta_d), \quad w_{dn} \sim \mathrm{Cat}(\phi_{z_{dn}}). \end{aligned} \tag{4}$$

where θ_d is the document–topic mixture, ϕ_k is the topic–word distribution, z_{dn} is the topic assignment for token n, and wdn is the observed word. Inference is typically via variational optimization or collapsed Gibbs sampling; evaluation often uses topic coherence.

2.4.2 Latent Semantic Analysis (LSA)

LSA (Latent Semantic Analysis) uncovers low-rank structure in the term–document matrix using linear algebra. The latent dimensions approximate co-occurrence structure and can serve as topics or semantic axes. Its performance is sensitive to preprocessing and weighting choices, which is why TF-IDF, and careful vocabulary curation are standard. Recent surveys include LSA as a core algebraic baseline[23].

Given $X \in R^{m \times n}$ (terms × documents), compute the singular value decomposition, equation (5):

$$X = U\Sigma V^{\top}. \tag{5}$$

A rank-k approximation retains the leading singular values and vectors, equation (6):

$$X_k = U_k \Sigma_k V_k^{\top}. \tag{6}$$

Columns of Uk (or rows of $V_k^{\top}$) provide the latent semantic dimensions used for analysis.

2.4.3 Non-Negative Matrix Factorization (NMF)

NMF factorizes a non-negative document–term matrix into non-negative "parts," often yielding sharper and more separated topics. Contemporary reviews and new variants such as deep NMF report strong separability and coherence on text tasks [5], [6].

Given $X \geq 0$, find $W \geq 0$ and $H \geq 0$ such that, equation (7):

$$X \approx WH, W \geq 0, H \geq 0. \quad (7)$$

often by minimizing a divergence such as the Frobenius norm, equation (8):

$$\min_{W,H \geq 0} \|X - WH\|_F^2. \quad (8)$$

Columns of W act as topics; rows of H give document–topic weights. Recent work refines objectives and optimization to improve stability and topic quality.

2.5 Model Evaluation

Topic modeling is a common method to uncover hidden themes in large text corpora by grouping words into topics [24]. However, not every set of words automatically forms a meaningful or interpretable topic. Therefore, evaluation is needed to ensure that the generated topics make sense not only mathematically but also semantically.

Topic coherence is one of the most widely used measures for this purpose [25], [26]. It evaluates the degree of semantic similarity among the most frequent words in a topic. If the top words within a topic are closely related in meaning, the topic is likely to be coherent and interpretable by humans. Conversely, if the words are unrelated, the topic will be less useful for interpretation.

In practice, coherence scores are often used to compare models with different numbers of topics and select the configuration that produces the most meaningful results [27]. Higher coherence generally indicates that topics are better aligned with human intuition. Because of this, coherence has become a standard evaluation metric in topic modeling research and has been shown to improve both topic quality and the reliability of subsequent analyses [28].

3. RESULTS AND DISCUSSION

3.1 Coherence Score Result

The evaluation of topic modeling performance in this study was conducted using the coherence score as a metric for measuring topic quality. Three algorithms were compared, namely Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix Factorization (NMF). The evaluation was performed for different numbers of topics (k) ranging from 1 to 20. The results of this evaluation are presented in Table 2.

Table 2. Coherence Score

k	Algorithm		
	LDA	LSA	NMF
1	0,2948	0,4502	0,4502
2	0,4365	**0,5925**	0,6278
3	0,3951	0,5395	0,5889
4	0,4722	0,5793	0,6569
5	0,4827	0,5517	0,7357
6	0,4962	0,5306	0,7188
7	0,4866	0,5362	0,6926
8	**0,5168**	0,5368	0,7207
9	0,4694	0,5203	0,7387
10	0,4788	0,5091	**0,7575**
11	0,4901	0,4977	0,7555
12	0,4985	0,4818	0,7423
13	0,5008	0,5022	0,7265

k	Algorithm		
	LDA	LSA	NMF
14	0,5001	0,4844	0,7368
15	0,4936	0,4824	0,7367
16	0,4823	0,4659	0,7404
17	0,5042	0,4658	0,7478
18	0,4847	0,4545	0,7362
19	0,472	0,4442	0,7395
20	0,4672	0,4495	0,7409

As presented in Table 2, the coherence scores demonstrate variations across algorithms and topic numbers. Among the three methods, NMF consistently produced the highest coherence scores. The maximum score for NMF was obtained at k=10 with a coherence value of 0.7575, indicating that this model generated the most coherent set of topics within the tested range.

For the LDA model (blue line), the coherence score exhibited a gradual improvement as the number of topics increased, reaching its optimal value at k=8 (0.5168). Although the performance of LDA was lower than NMF, the results suggest that LDA remains effective in producing reasonably coherent topics, particularly when a probabilistic representation of word distributions is required. The LSA model (orange line), in contrast, achieved its highest coherence score at k=2 (0.5925), after which the values generally declined as the number of topics increased. This indicates that LSA was less stable in maintaining topic coherence, particularly at higher values of k.

Meanwhile, the NMF model (green line) consistently outperformed both LDA and LSA, with coherence scores remaining relatively high across different values of k. The best result was obtained at k=10 (0.757), confirming the superiority of NMF in this experimental setting. The visualization of the coherence scores for the three models can be seen in Figure 2.

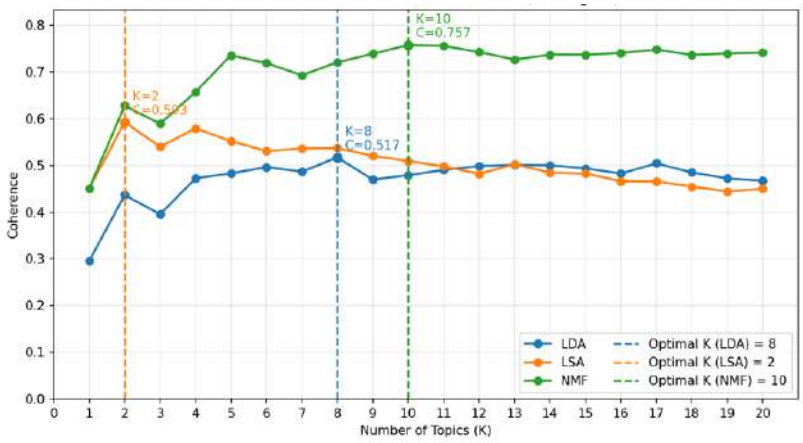

Figure 2. Coherence Score

In summary, the results indicate that NMF is the most suitable algorithm for topic modeling in this study, as it consistently produced the highest coherence values and demonstrated stability across different topic numbers. LDA achieved moderate performance and serves as a viable alternative when a probabilistic framework is required, while LSA produced the weakest results overall.

3.2 Topic Analysis

Based on the topic coherence evaluation, the optimal number of topics was determined as eight for LDA, two for LSA, and nine for NMF. To gain a deeper understanding of the generated topics, we conducted a detailed analysis of the topic distributions. The following subsections present the results of this analysis, supported by wordcloud visualizations and interpretation tables for each model.

3.2.1 LDA Topic Analysis

To facilitate topic interpretation, wordcloud visualizations were generated for each of the eight LDA topics, as presented in Figure 3. Each wordcloud highlights the most representative words for the respective topic, where larger words indicate higher importance within the topic distribution. This visualization simplifies the justification process for identifying coherent themes.

Figure 3. Wordcloud LDA

The interpretation of each topic was conducted by reviewing the dominant words in the wordclouds. The results are summarized in Table 3, which shows the eight extracted themes. The topics range from program implementation, budgetary issues, and economic impacts to food safety, logistics, and social concerns.

Table 3. LDA Topic Analysis

No	Topic Number	Topic
1	0	Partners & Program Implementation
2	1	Budget & Government Policy
3	2	Economy & MSMEs
4	3	Health Issues (Food Poisoning)
5	4	Politics & Children's Education
6	5	Distribution & Logistics
7	6	Social Welfare & Stunting
8	7	Public Support & Media

From these topics, it is evident that LDA successfully captured multiple dimensions of the *Free Nutritious Meal* Program, including its political framing,

operational aspects, and socio-economic consequences. The granularity of topics indicates that LDA provides a relatively comprehensive picture compared to the other models.

3.2.2 LSA Topic Analysis

For LSA, only two topics were generated. The corresponding wordclouds are shown in Figure 4. Similar to LDA, the visualization aids in identifying the semantic coherence of the topics. However, due to the reduced dimensionality inherent in LSA, the topics appear broader and less granular.

Figure 4. Wordcloud LSA

The two identified topics are summarized in Table 4. Both topics revolve around program implementation and budget allocation, with an additional emphasis on food safety concerns.

Table 4. LSA Topic Analysis

No	Topic Number	Topic
1	0	Program Implementation in Schools
2	1	Budget & Food Safety Issues

This outcome demonstrates that LSA captures the most dominant themes but lacks the ability to disentangle specific sub-topics such as logistics, stunting, or media narratives. While less granular, LSA provides a high-level overview of the discourse.

3.2.3 NMF Topic Analysis

The NMF model produced nine interpretable topics, supported by wordclouds in Figure 5. Compared to LSA, NMF is capable of generating more fine-grained topics, similar in richness to LDA, though with slightly different thematic boundaries.

The identified topics are summarized in Table 5. NMF successfully distinguishes between socioeconomic impacts, budget allocation, school-level implementation, and issues such as stunting, political figures, and public opinion.

Table 5. NMF Topic Analysis

No	Topic Number	Topic
1	0	Socioeconomic Impact & MSMEs
2	1	National Budget (APBN) & Fund Allocation
3	2	School Meals & Daily Menu
4	3	Distribution Mechanism & Beneficiaries
5	4	Food Poisoning Cases
6	5	Political Support & Public Figures

No	Topic Number	Topic
7	6	Ministries & Public Policy Role
8	7	Education & Schools
9	8	Nutrition & Stunting
10	9	Public Opinion & Media

Figure 5. Wordcloud NMF

The NMF results provide a nuanced understanding of the program, highlighting its multi-dimensional nature: political, economic, health-related, and social. Particularly, the presence of topics on *Nutrition & Stunting* and *Public Opinion* indicates the ability of NMF to capture both outcome-focused and perception-driven aspects.

3.3 Comparative Topic Analysis

A comparative analysis was conducted to identify overlapping and distinct themes across LDA, LSA, and NMF. While LDA generated eight topics and NMF generated nine topics, LSA produced only two broader themes. The results show that LDA and NMF captured a richer variety of dimensions such as distribution, stunting, political aspects, and media narratives, whereas LSA only highlighted the dominant themes of budget and food safety. The cross-model comparison is summarized in Table 6, which aligns similar themes identified across different models.

From this comparison, Budget & Policy, School Meals & Menus, and Food Poisoning Cases consistently emerged across all three models. In contrast, Economy & MSMEs, Distribution, Stunting, Politics, and Public Opinion appeared only in LDA and NMF but were absent in LSA. This suggests that LDA and NMF provide more nuanced representations of the discourse, while LSA offers a condensed high-level overview.

Table 6 Compararative Topic Analysis

Topic Number	LDA	LSA	NMF
Budget & Policy	1	1	1, 6
Economy & MSMEs	2	–	0
School Meals & Menus	3	0	2, 7
Food Poisoning Cases	3	1	4
Distribution & Beneficiaries	0, 5	–	3
Nutrition & Stunting	6	–	8
Politics	4	–	5
Public Opinion & Media	7	–	9

The triangulation of these findings strengthens the conclusion that the Free Nutritious Meal Program discourse in public media is multifaceted, covering not only financial and health aspects but also broader socio-political and economic dimensions.

3.4 Topic Trend Analysis

The temporal dynamics of topic prevalence offer critical insight into how the public discourse around the Free Nutritious Meal (MBG) program evolved in relation to political events, media attention, and program rollout. As shown in Figure 6 (LDA trend), Figure 7 (LSA trend), and Figure 8 (NMF trend), several consistent patterns are evident.

Although foundational discourse on nutritious meal programs exists since 2020, the volume of media attention remained relatively low until 2024. Around this period, a sharp increase in MBG-related news coverage is observed, coinciding with major national events and shifts in public attention. This pattern highlights the ability of topic modeling to capture sudden changes in discourse intensity within large-scale online news streams. After his election, the trend intensified further in 2025, as the MBG initiative transitioned from campaign rhetoric to actual policy implementation.

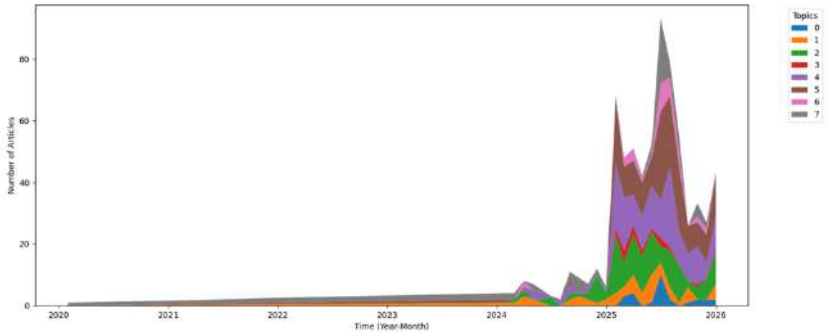

Figure 6. Topic Trends Over Time – LDA

In the LDA model Figure 6, topics such as Budget & Government Policy experienced a pronounced rise after 2024, reflecting intensified debates on funding

and state budget allocation. The Politics & Children's Education topic showed a sharp increase during the campaign period, illustrating how the program was framed within electoral narratives. The spike in Health Issues (Food Poisoning) in 2025 corresponds with extensive news coverage of food safety incidents, which attracted significant public scrutiny. Over time, the Social Welfare & Stunting and Public Support & Media topics also gained more prominence, indicating the growing importance of nutrition outcomes and the role of public sentiment in shaping the discourse.

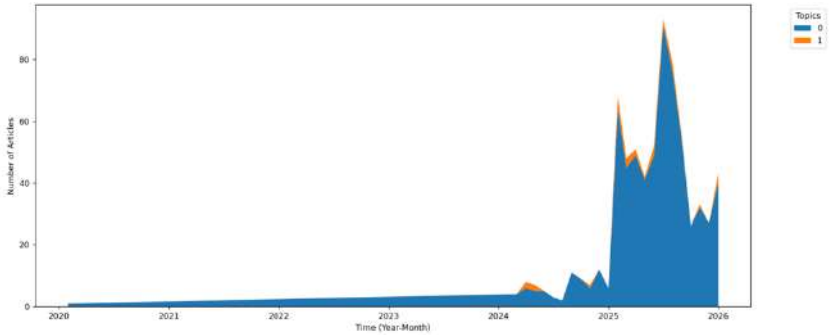

Figure 7. Topic Trends Over Time – LSA

Under LSA Figure 7, the two themes Program Implementation in Schools and Budget & Food Safety Issues display a parallel trend, with minimal mention prior to 2024, followed by rapid growth in 2024 and beyond. Although less granular than LDA or NMF, these themes reliably capture the dominant threads of discourse, particularly those focusing on feasibility and funding.

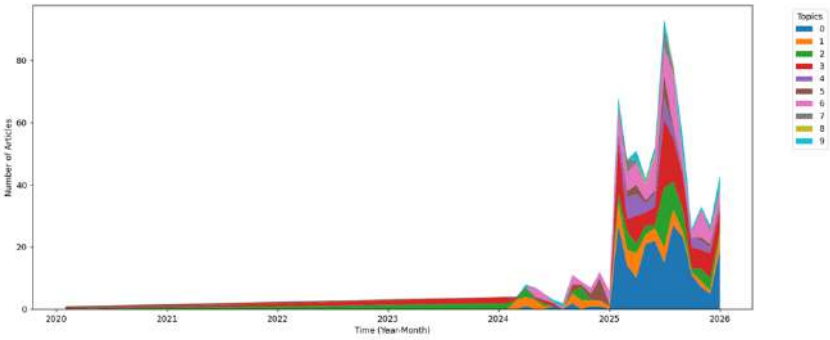

Figure 8. Topic Trends Over Time – NMF

In the NMF model Figure 8, a more detailed picture of the discourse is revealed. The Socioeconomic Impact & MSMEs topic rose significantly in 2025, reflecting the government's narrative about strengthening the role of local food suppliers and microenterprises. The Food Poisoning Cases topic peaked in a pattern similar to that of the LDA results, reinforcing the importance of food safety in shaping public concerns. The Nutrition & Stunting topic displayed a steady upward trajectory, linking the program with long-term health improvements and its relevance to Sustainable Development Goal agendas. In addition, the Public Opinion & Media topic increased during both the campaign and the early implementation phases, indicating strong media coverage and a high level of public debate.

Across all models, the trajectory is consistent: a low baseline from 2020 to 2023, a sharp surge beginning in 2024 (associated with electoral politics), and further

amplification into 2025 as the program moved toward execution. This pattern supports literature on media agenda-setting and policy salience during election cycles, where political events act as catalysts for topic visibility in the press. The case of MBG also fits with findings that media framing can influence public perception and policy evaluation (e.g. "Politics on a Plate" in Indonesia) [29]. Moreover, studies specifically on MBG's media framing (e.g. the narrative fractures study using NMF, and framing analyses of Detik.com and Kompas) corroborate that official narratives emphasize policy and nutrition benefits while public reactions often highlight skepticism, financial constraints, or implementation challenges [30], [31].

In sum, the trend analyses validate that the MBG discourse is deeply intertwined with political timing, implementation milestones, and media amplification. Topics that saw rising trends such as budget & policy, food safety, nutrition & stunting, and public opinion signify the evolving focal points of dialogue and align with broader theoretical and empirical work on how welfare programs and nutrition policies are debated in media and public arenas.

3.5 Implications for ICT-Based News Analytics Systems

The results of this study demonstrate the practical relevance of topic modeling as a core component of ICT-based news analytics systems. The comparative analysis of LDA, LSA, and NMF shows that unsupervised topic modeling can be effectively applied to large-scale Indonesian online news data to uncover dominant themes and track their evolution over time. Such capabilities are particularly valuable for media monitoring systems that must process continuously growing volumes of digital content.

The temporal topic trend analysis further highlights the potential of topic modeling for early issue detection and trend monitoring. Sudden increases in topic prevalence indicate shifts in public attention and emerging narratives, which can serve as early signals for policymakers, analysts, and organizations relying on timely information. From a system perspective, these methods enable automated surveillance of news streams without requiring manual labeling or prior topic definitions.

In addition, the scalability of classical topic modeling techniques makes them suitable for integration into decision support systems and large-scale information management platforms. By combining coherence-based model selection with qualitative interpretability, the proposed analytics pipeline supports both computational efficiency and human-centered analysis. Overall, this study confirms that topic modeling provides a robust and scalable foundation for ICT-based media analytics, particularly in environments characterized by high data volume, linguistic complexity, and rapidly evolving public discourse.

4. CONCLUSIONS

This study compared three classical topic modeling techniques—Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF)—for large-scale Indonesian online news analytics using a long-running news corpus as a case study. The results show that NMF achieved the highest topic coherence and produced stable topic structures, while

LDA offered stronger semantic interpretability through probabilistic topic representations. In contrast, LSA generated broader themes with lower granularity. The temporal analysis demonstrates that topic modeling is effective in identifying changes in media attention and tracking discourse evolution over time in large-scale news data. From an information and communication technology (ICT) perspective, these findings confirm that classical topic modeling techniques can be reliably integrated into scalable news analytics pipelines for media monitoring, trend detection, and decision support. Overall, this study highlights the continued relevance of topic modeling for analyzing high-volume Indonesian online news content.

REFERENCES

[1] A. Tri Haryanto, "APJII: Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang," detikinet. Accessed: Sep. 16, 2025. [Online]. Available: https://inet.detik.com/cyberlife/d-7169749/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003, doi: 10.1162/jmlr.2003.3.4-5.993.

[3] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, p. 886498, May 2022, doi: 10.3389/FSOC.2022.886498/BIBTEX.

[4] M. Hankar, M. Kasri, and A. Beni-Hssane, "A comprehensive overview of topic modeling: Techniques, applications and challenges," *Neurocomputing*, vol. 628, p. 129638, May 2025, doi: 10.1016/J.NEUCOM.2025.129638.

[5] Y. T. Guo, Q. Q. Li, and C. S. Liang, "The rise of nonnegative matrix factorization: Algorithms and applications," *Inf Syst*, vol. 123, p. 102379, Jul. 2024, doi: 10.1016/J.IS.2024.102379.

[6] J. Wang and X. L. Zhang, "Deep NMF topic modeling," *Neurocomputing*, vol. 515, pp. 157–173, Jan. 2023, doi: 10.1016/J.NEUCOM.2022.10.002.

[7] M. Rüdiger, D. Antons, A. M. Joshi, and T. O. Salge, "Topic modeling revisited: New evidence on algorithm performance and quality metrics," *PLoS One*, vol. 17, no. 4, p. e0266325, Apr. 2022, doi: 10.1371/JOURNAL.PONE.0266325.

[8] A. Rkia, A. Fatima-Azzahrae, A. Mehdi, and L. Lily, "NLP and Topic Modeling with LDA, LSA, and NMF for Monitoring Psychosocial Well-being in Monthly Surveys," *Procedia Comput Sci*, vol. 251, pp. 398–405, Jan. 2024, doi: 10.1016/J.PROCS.2024.11.126.

[9] V. Leplat, L. T. K. Hien, A. Onwunta, and N. Gillis, "Deep Nonnegative Matrix Factorization With Beta Divergences," *Neural Comput*, vol. 36, no. 11, pp. 2365–2402, Oct. 2024, doi: 10.1162/NECO_A_01679.

[10] M. F. Asnawi, M. Hanafi, N. F. Kurniawan, A. Suwondo, A. Nasrullah, and C. Setyawan, "Topic Modelling Analysis on Indonesian News Using BERT Topic Model," *2024 6th International Conference on Cybernetics and Intelligent System, ICORIS 2024*, 2024, doi: 10.1109/ICORIS63540.2024.10903779.

[11] D. Medvecki, B. Bašaragin, A. Ljajić, and N. Milošević, "Multilingual transformer and BERTopic for short text topic modeling: The case of Serbian," pp. 161–173, Feb. 2024, doi: 10.1007/978-3-031-50755-7_16.

[12] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.

[13] M. Nesca, A. Katz, C. K. Leung, and L. M. Lix, "A scoping review of preprocessing methods for unstructured text data to assess data quality," *Int J Popul Data Sci*, vol. 7, no. 1, p. 1757, 2022, doi: 10.23889/IJPDS.V6I1.1757.

[14] M. Risky and A. Aprillia, "Komparasi Algoritma Klasifikasi Dan Penerapan Ner pada Analisis Sentimen Bencana Alam Banjir," *InComTech : Jurnal Telekomunikasi dan Komputer*, vol. 14, no. 2, pp. 130–139, Aug. 2024, doi: 10.22441/INCOMTECH.V14I2.22962.

[15] R. Ariftiarno and M. Dwijayanti, "Analisis Sentimen Netizen Twitter Terhadap Pelayanan Provider Telkomsel: Komparasi Naive-Bayes dan K-Nearest Neighbors," *InComTech : Jurnal Telekomunikasi dan Komputer*, vol. 14, no. 1, pp. 60–77, May 2024, doi: 10.22441/INCOMTECH.V14I1.22597.

[16] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.

[17] M. Nesca, A. Katz, C. K. Leung, and L. M. Lix, "A scoping review of preprocessing methods for unstructured text data to assess data quality," *Int J Popul Data Sci*, vol. 7, no. 1, p. 1757, 2022, doi: 10.23889/IJPDS.V6I1.1757.

[18] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLoS One*, vol. 16, no. 8, p. e0254937, Aug. 2021, doi: 10.1371/JOURNAL.PONE.0254937.

[19] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf Syst*, vol. 121, p. 102342, Mar. 2024, doi: 10.1016/J.IS.2023.102342.

[20] R. N. Rathi and A. Mustafi, "The importance of Term Weighting in semantic understanding of text: A review of techniques," *Multimed Tools Appl*, vol. 82, no. 7, pp. 9761–9783, Mar. 2023, doi: 10.1007/S11042-022-12538-3/METRICS.

[21] Y. Setiawan, N. U. Maulidevi, and K. Surendro, "The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification," *Data Sci J*, vol. 23, no. 1, 2024, doi: 10.5334/DSJ-2024-031.

[22] M. Liang and T. Niu, "Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs," *Procedia Comput Sci*, vol. 208, pp. 460–470, Jan. 2022, doi: 10.1016/J.PROCS.2022.10.064.

[23] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf Syst*, vol. 112, p. 102131, Feb. 2023, doi: 10.1016/J.IS.2022.102131.

[24] A. Simonetti, A. Albano, A. Plaia, and M. Tumminello, "Ranking coherence in topic models using statistically validated networks," *J Inf Sci*, vol. 51, no. 3, pp. 744–765, Jun. 2025, doi: 10.1177/01655515221148369.

[25] Y. Wu, C. H. Tseng, J. Shang, S. Mao, G. Nenadic, and X. J. Zeng, "EDU-level Extractive Summarization with Varying Summary Lengths," *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*, pp. 1655–1667, 2023, doi: 10.18653/V1/2023.FINDINGS-EACL.123.

[26] R. Li, F. González-Pizarro, L. Xing, G. Murray, and G. Carenini, "Diversity-Aware Coherence Loss for Improving Neural Topic Models," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 1710–1722, May 2023, doi: 10.18653/v1/2023.acl-short.145.

[27] S. Syahrial, R. Perucha, and F. Afidh, "Fine-Tuning Topic Modelling: A Coherence-Focused Analysis of Correlated Topic Models," *Infolitika Journal of Data Science*, vol. 2, no. 2, pp. 82–87, Nov. 2024, doi: 10.60084/IJDS.V2I2.236.

[28] D. Xu *et al.*, "Mitigating Hallucinations of Large Language Models in Medical Information Extraction via Contrastive Decoding," *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, pp. 7744–7757, 2024, doi: 10.18653/V1/2024.FINDINGS-EMNLP.456.

[29] H. Roris and P. Sianturi, "Politics on a plate: equivocal communication in Indonesian presidential nutrition policy," *Front Commun (Lausanne)*, vol. 10, p. 1612652, Sep. 2025, doi: 10.3389/FCOMM.2025.1612652.

[30] G. Agarwal, S. K. Dinkar, and A. Agarwal, "Binarized spiking neural networks optimized with Nomadic People Optimization-based sentiment analysis for social product recommendation," *Knowl Inf Syst*, vol. 66, no. 2, pp. 933–958, Feb. 2024, doi: 10.1007/S10115-023-01956-W.

[31] V. N. N. Tsara and Y. Purbokusumo, "Framing Analysis of Free Nutritious Meal Program (MBG) News on Detik.com: Implications for Public Perception and Government Policy," Universitas Gadjah Mada, Yogyakarta, 2025.