Jurnal Impresi Indonesia (JII)
p-ISSN: 2828-1284 e-ISSN: 2810-062x
website: https: //rivierapublishing.id/jii/index.php/
Vol. 5. No. 1. January. 2026

Application of Latent Dirichlet Allocation (LDA) and BERTopic Algorithms for Headline and Topic Analysis of Palestine–Israel Conflict News in Indonesian Online Media

Raden Gumilar Riyansyah*, Sajarwo Anggai, Tukiyat
Universitas Pamulang, Indonesia
Email: radengumilarr@gmail.com* , sajarwo@gmail.com, dosen02711@unpam.ac.id

Keywords
Topic Analysis, LDA, BERTopic, Model Combination, Model Comparison, Palestine–Israel Conflict, Online Media.

Abstract
Palestine–Israel conflict news coverage has become one of the most dominant international issues in Indonesian online media, necessitating a systematic analysis to understand the topic structures formed from the intensity and variation of the narratives presented. The main challenges arise from the high volume of text, differences in writing styles across media outlets, and the diversity of terminology—all of which hinder consistent manual topic identification. To address these challenges, this study proposes a combined and comparative approach using two topic modeling algorithms, Latent Dirichlet Allocation (LDA) and BERTopic, to obtain a more accurate, structured, and interpretable topic mapping. The modeling process begins with data collection through web scraping, followed by a preprocessing stage consisting of cleansing, case folding, tokenization, normalization, filtering, and stemming. The LDA model is developed by determining the optimal number of topics based on coherence score and perplexity, whereas BERTopic leverages transformer-based embeddings, UMAP dimension reduction, and HDBSCAN clustering. Evaluation is conducted using coherence score, perplexity, silhouette score, and visualizations such as intertopic distance maps and word clouds to assess topic quality. The results show that BERTopic achieves the highest coherence score of 0.99 and lower perplexity, producing semantically cohesive topics. Meanwhile, LDA remains advantageous in providing a stable and measurable probabilistic structure. The combination of both models yields a mapping of five main topics: attacks in Gaza, Indonesian diplomacy, international support, humanitarian issues, and global political dynamics. These findings demonstrate that integrating LDA and BERTopic enhances the quality of topic analysis on complex issues.

INTRODUCTION

The conflict between Palestine and Israel has persisted for decades and is recognized as one of the most complex and prolonged geopolitical disputes in the world (Suhayatmi et al., 2024). News coverage on this issue frequently addresses various dimensions, including military attacks, international diplomacy, and humanitarian crises (Manurung & Heriamsal, 2024). Mass media plays a crucial role in shaping public perceptions of the conflict through diverse and continuous reporting. As the volume of text data produced by media outlets continues to grow, effective analytical methods are required to understand the main topics being reported (Nuraliza et al., 2024).

The Palestine–Israel conflict is a longstanding dispute rooted in territorial claims, identity, and sovereignty dating back to the late nineteenth century. It has been further

complicated by colonial political promises, the establishment of Israel in 1948, and the ongoing occupation of Palestinian territories (Setiawan, 2024). Diplomatic efforts such as the Oslo Accords have failed to resolve key issues, including the status of Jerusalem, refugee rights, and illegal settlements (Mughni Sulubara et al., 2024). Studying this conflict through media narratives is essential because the narratives constructed influence public opinion, contain potential biases, and shape global understanding (Hartini et al., 2024).

The primary challenge in analyzing news topics related to the Palestine–Israel conflict lies in the complexity and large volume of data (Jelodar et al., 2019; Tran et al., 2019). The sheer number of articles published within a specific time frame makes it difficult to systematically extract meaningful information. Additionally, variations in writing styles across media outlets and the diversity of terminology used further complicated efforts to obtain consistent topic representations (Hajra Chaudhry, 2024; Linstead et al., 2008; Zahoor & Sadiq, 2021). Therefore, a systematic and advanced textual analysis approach is required to accurately capture semantic diversity and document structure. Various methods have been employed to analyze news texts, ranging from traditional manual content analysis to modern text-mining and machine-learning techniques. One of the most widely used methods for text analysis is Latent Dirichlet Allocation (LDA), which identifies hidden topics within a collection of documents (Helmayanti et al., 2023). In addition to LDA, the more recent BERTopic algorithm has demonstrated strong potential in topic analysis by leveraging transformer-based models and deep learning to produce more accurate and interpretable results (Nursyahrina et al., 2024).

This study proposes the use of two topic modeling algorithms, LDA and BERTopic, to analyze headlines and news topics related to the conflict between Palestine and Israel. LDA is employed to identify the main probabilistic topics based on word distributions within documents. In contrast, BERTopic utilizes transformer-based models to generate richer text embeddings and clusters documents according to semantic similarity. The comparison of these two methods is expected to provide a more comprehensive and in-depth understanding of the main topics reported in the news, as well as address challenges arising from complex and heterogeneous text data.

The Palestine–Israel conflict, rooted in the 1917 Balfour Declaration, remains a complex international issue shaped by historical, political, and religious factors (Syahroni et al., 2025). Indonesia consistently supports Palestine based on anti-colonial principles and Muslim solidarity, a stance reflected in government policy and strong public backing in international forums. Indonesian media also provide extensive coverage of the conflict, emphasizing Palestinian suffering and the resulting humanitarian crisis from Israeli attacks, thereby reinforcing public opinion that predominantly favors Palestine (Ramadani et al., 2024).

A news headline is the main title of an article that succinctly highlights the core message to capture readers' attention. In both digital and print media, headlines are crucial because they influence whether audiences proceed to the full story. Commonly written using the inverted-pyramid approach, they place key information at the beginning to enhance clarity and impact. Beyond their visual prominence, headlines also act as framing tools, where word choice and stylistic elements shape initial reader perceptions and influence how an issue is interpreted (Wastanti & Wiratama, 2024).

A news topic is the central theme or main issue that frames the focus of a news article. It provides the structural foundation of the report, ensuring that all presented information remains aligned with the core event or issue intended for the audience. A well-defined topic enables journalists to organize content coherently and helps readers understand the essence of the news

without losing context. In practice, news topics also reflect the media's agenda setting role, where certain issues are selected and emphasized based on relevance or public interest. This process shapes audience perceptions by influencing which information is highlighted and from what perspective. Repeated coverage of particular topics can indicate the priorities or biases of a media outlet, offering insights into how information is framed and presented (Faudzi & Fajri, 2024).

Topic modeling is a Natural Language Processing (NLP) technique used to identify hidden themes within large text collections by analyzing patterns of word co-occurrence, enabling researchers to uncover the underlying structure of information without manually reviewing each document. Methods such as Latent Dirichlet Allocation (LDA) and BERTopic group semantically related words into coherent topics and cluster documents based on shared thematic patterns. This technique is widely applied in academic research, market analysis, and media studies because it efficiently reveals dominant themes, supports the interpretation of complex textual data, and facilitates data-driven decision-making (Moch Erreza et al., 2024).

Latent Dirichlet Allocation (LDA) is a probabilistic topic modeling method that identifies hidden thematic structures within text collections by assuming that each document contains multiple topics and each topic consists of a distribution of words. Through iterative optimization, LDA estimates document topic and topicword distributions, allowing coherent topics to be extracted based on word co-occurrence patterns. The method is widely used due to its computational efficiency, scalability for large corpora, and ability to generate interpretable topics, making it suitable for analyzing news articles, academic texts, and other unstructured data. In practice, parameters such as the number of topics, alpha (document topic density), beta (topic word density), and the iteration count must be carefully configured to ensure stable, meaningful, and coherent topic outputs for research or decision making purposes (Abdurrazzaq, 2023).

BERTopic is an advanced topic modeling technique that utilizes transformer-based contextual embeddings, particularly BERT, to generate semantically coherent topics from text corpora. Unlike probabilistic models such as LDA, BERTopic creates high-dimensional embeddings that capture deep semantic relationships, which are then reduced with UMAP and clustered using HDBSCAN without requiring a predefined number of topics. The method integrates semantic embedding models, c-TF-IDF weighting, and density-based clustering, enabling the extraction of rich and precise topics from complex datasets. Although BERTopic provides strong contextual modeling, automatic topic discovery, and effective noise handling, it remains computationally intensive and can be less effective for very short texts while also posing interpretability challenges for users unfamiliar with embedding-based approaches (Samsir, Reagan Surbakti Saragih, Selamat Subagio, Rahmad Aditya, 2023).

This study aims to apply the Latent Dirichlet Allocation (LDA) algorithm and BERTopic in analyzing headline patterns and topics of reporting on the Palestinian-Israeli conflict in Indonesian online media to identify dominant themes, narrative dynamics, and trends in framing news that are developing. The benefits of this research are expected to make a theoretical contribution to the development of textual analysis and computational social science studies by enriching the application of topic modeling in the context of media and international conflicts, as well as providing practical benefits for academics, media practitioners, and policymakers in understanding the direction of news, potential narrative bias, and distribution

of issues that affect public opinion in Indonesia based on an analytical approach objective and data-driven.

METHODS

The extensive volume of news coverage on the Palestine–Israel conflict in Indonesian online media from 1 January to 31 March 2025 spanning military, diplomatic, and humanitarian dimensions makes manual identification of key topics challenging due to the complexity and scale of the data. Differences in writing style and terminology across media outlets further create inconsistencies in topic representation, limiting the ability of traditional content analysis to capture the full semantic diversity of the articles. Consequently, a technology-driven approach is necessary, with LDA and BERTopic offering effective solutions for automatically, accurately, and interpretively identifying and visualizing dominant topics on a large scale.

The selection of the research object ensures that the analyzed data aligns with the study's goals of identifying news patterns, mapping main topics, and evaluating the performance of LDA and BERTopic in large-scale text analysis. The object consists of Indonesian online news articles on the Palestine–Israel conflict published between 1 January and 31 March 2024 a period marked by intense political, diplomatic, military, and humanitarian reporting. These articles display substantial semantic variability due to differences in writing style, media perspectives, and terminology, making manual topic identification ineffective. This dataset is therefore suitable for applying LDA and BERTopic, which can process large text corpora, extract latent themes, and visualize topic structures systematically. The chosen object provides a strong foundation for analyzing issue framing in Indonesian media and assessing the effectiveness of both algorithms in capturing dominant topics during a critical escalation period of the conflict.

The data collection method employed in this study is web scraping. The dataset used consists of secondary data obtained online through a scraping technique utilizing the Scraper extension integrated into Python Flask and implemented using the Python programming language. This technique enables the automated extraction of data from websites without requiring manual copying.

This study collected online news articles and headlines related to the Palestine–Israel conflict published by Indonesian media from 1 January 2025 to 31 March 2025. The data were obtained through purposive sampling by selecting relevant news based on specific keywords. Web scraping was conducted using Python Flask with BeautifulSoup and requests to extract titles, descriptions, and publication dates. The dataset was then cleaned to remove duplicates and formatting errors. The final dataset consists of secondary news texts from Detik.com, Kompas.com, and CNNIndonesia.com, stored in CSV format with columns for title, description, and label. These data form the basis for topic analysis using LDA and BERTopic. Table 1 shows the total number of articles collected.

Table 1. Total Dataset of Online News on the Palestine–Israel Conflict

No	Media Name	Media Link	Total Dataset
1	Detik	https://www.detik.com	510
2	Kompas	https://www.kompas.com	506
3	CNN Indonesia	https://www.cnnindonesia.com	455

Topic modeling serves as a core stage of this study, aiming to extract latent themes from online news related to the Palestine–Israel conflict. Two algorithmic approaches were applied LDA and BERTopic integrated within a Flask-based system through the main function run_analysis_core(), which executes training and topic visualization. LDA identifies topic distributions using a probabilistic framework, while BERTopic generates contextual semantic representations using transformer-based embeddings.

The combined use of both models enables a comprehensive analysis from statistical and semantic perspectives. LDA was applied to identify latent topic patterns through CountVectorizer-based text vectorization, model training, and visualization using pyLDAvis and WordCloud. The parameter settings used in this study are shown in Table 2.

Table 2. LDA Model Parameter Configuration

Parameter	Value	Description
max_df	0.95	Excludes terms appearing in more than 95% of documents.
min_df	2	Removes terms appearing in fewer than two documents.
stop_words	'indonesian'	Eliminates common Indonesian stopwords.
n_components	10	Defines the number of topics generated.
random_state	42	Ensures reproducibility.

These parameters were selected to maintain stable topic separation, with max_df and min_df controlling term frequency thresholds and n_components set to capture the corpus's major thematic variations. Visual outputs from pyLDAvis and WordCloud effectively highlight semantic relationships and dominant keywords within each topic. BERTopic was used to generate semantically richer topic representations through SentenceTransformer embeddings, UMAP dimensionality reduction, and HDBSCAN clustering. The parameter configuration is summarized in Table 3.

Table 3. BERTopic Model Parameter Configuration

Component	Parameter	Value	Description
SentenceTransformer	model_name	"paraphrase-multilingual-MiniLM-L12-v2"	Multilingual model for contextual embeddings.
UMAP	n_neighbors	15	Number of nearest neighbors for local structure preservation.
	n_components	5	Reduced dimensionality.
	min_dist	0.0	Controls minimum point distance in reduced space.
	metric	'cosine'	Similarity metric for embeddings.
HDBSCAN	–	–	Clusters documents based on semantic proximity.
Evaluation	silhouette_score	–	Measures separation quality across clusters.

The BERTopic pipeline begins with embedding generation, followed by dimensionality reduction and topic clustering. Model evaluation uses the Silhouette Score, and results are visualized interactively via Plotly. This configuration enables BERTopic to capture deeper semantic relations, producing topic representations that align closely with substantive themes such as diplomacy, international responses, and humanitarian issues.

RESULT AND DISCUSSION

Results of Analysis

The study aimed to identify thematic patterns in Indonesian online news about the Palestine–Israel conflict using two topic-modeling methods: LDA and BERTopic. LDA applies a probabilistic generative approach, while BERTopic uses transformer-based embedding combined with UMAP and HDBSCAN clustering. Both algorithms were evaluated to determine the consistency and relevance of the topics generated from the news corpus.

LDA Analysis Results

The initial LDA analysis evaluated topic quality using coherence and perplexity. Coherence indicates the semantic relatedness of words within a topic, while lower perplexity reflects better predictive performance. Table 4 summarizes the performance of all topics at K=20, enabling an objective and systematic selection of the optimal topic. This table also supports identifying the topic that best represents the Palestine–Israel narrative, ensuring semantic consistency and model stability before interpreting the topic keywords.

Table 4. Coherence Score & Perplexity LDA

Topic	Coherence Score	Perplexity
Topic 1	0.5056	8.5179
Topic 2	0.3986	8.9439
Topic 3	0.3818	9.4117
Topic 4	0.5170	8.7889
Topic 5	0.2499	9.4557
Topic 6	0.4540	8.9630
Topic 7	0.4775	8.7417
Topic 8	**0.5829**	**8.3396**
Topic 9	0.4181	9.8704
Topic 10	0.3776	8.8366
Topic 11	0.3892	9.1123
Topic 12	0.2944	9.7323
Topic 13	0.3836	9.0372
Topic 14	0.4393	8.7575
Topic 15	0.4055	9.3760
Topic 16	0.5263	9.0846
Topic 17	0.3202	9.7577
Topic 18	0.3337	9.2254
Topic 19	0.4296	8.9047
Topic 20	0.4489	8.5464

Based on the coherence and perplexity values in Table 4, Topic 8 shows the best overall performance, reflected by the highest coherence and a relatively low perplexity, indicating strong semantic stability and predictive accuracy. These results position Topic 8 as the most representative topic for further analysis and confirm that the LDA model effectively captures consistent patterns in the news dataset. Figure 1 visualizes the coherence distribution across all topics, enabling quick identification of topics with stronger semantic associations and providing insight into the overall quality and stability of the model. This visualization supports

methodological transparency and strengthens the justification for selecting the most coherent topic.

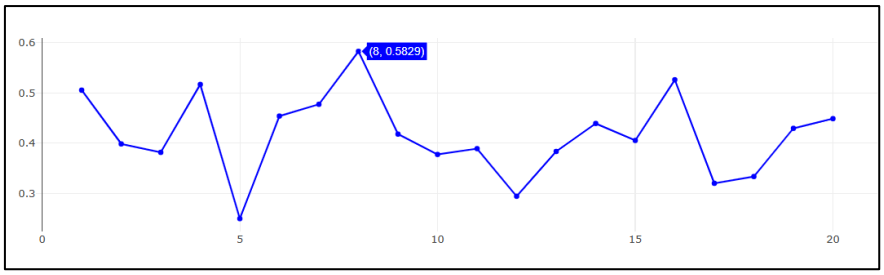

Figure 1. Comparison of K=20 and Coherence Scores (LDA)

Figure 1 shows clear variation in coherence values across topics, with Topic 8 achieving the highest score and demonstrating the strongest semantic consistency, reinforcing the numerical results in the table. The differing coherence levels indicate varying topic quality, with some topics appearing more generic. This visualization helps clarify the selection of the best topic by combining numerical and visual evidence. Figure 2 presents the distribution of perplexity values, illustrating the model's predictive accuracy in representing word distributions. Lower perplexity reflects more reliable topic formation, complementing the coherence assessment. Together, these visualizations provide a comprehensive evaluation of topic quality and support the decision to identify Topic 8 as the most representative topic.

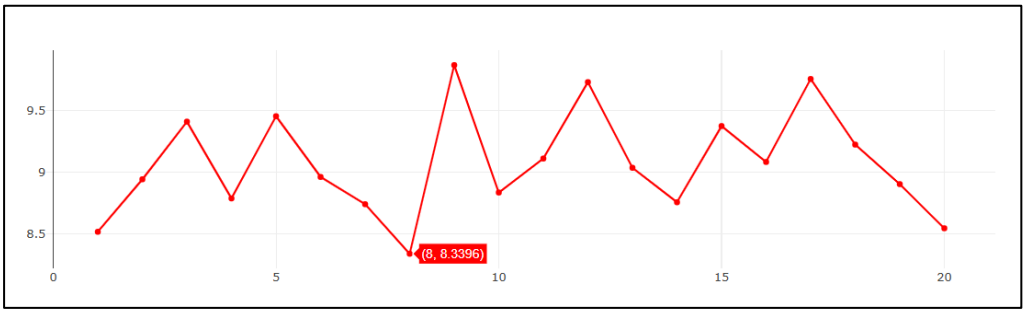

Figure 2. Comparison of K=20 and Perplexity Scores (LDA)

Figure 2 shows that perplexity values fall within a relatively uniform range, with some topics performing better through lower scores, reinforcing that Topic 8 is strong in both coherence and perplexity. The visual pattern indicates that the LDA model generates topics that are semantically coherent and predictively reliable, while higher perplexity in certain topics suggests broader or less focused themes. This visualization strengthens the numerical evaluation and supports balanced topic selection. Table 5 then details the keyword structure of Topic 8, providing essential insight into its dominant semantic content and the narrative patterns it represents. The keywords highlight high-frequency and high-relevance terms that shape the discourse on the Palestine–Israel conflict, serving as a foundation for validating Topic 8 before proceeding to further visualizations.

Table 5. Keywords of Topic 8 LDA

No	Keywords	Total Score
1	gencatan (170.37), senjata (167.91), gaza (120.60), israel (79.56), israelhamas (30.88), hamas (27.48), palestina (22.92), warga (21.40), trump (19.85), konflik (19.22), netanyahu (18.64), kesepakatan (14.50), ri (12.26), kabinet (12.12), biden (11.01), menlu (10.64), berlaku (10.12), resmi (9.90), menteri (9.55), sambut (8.66), sugiono (8.13), jelang (8.05), setujui (7.13), ancam (7.09), masjid (7.02), serang (6.88), desak (6.35), aqsa (6.13), januari (6.13), al (6.13)	866.60
2	palestina (46.54), israel (14.80), netanyahu (14.47), presiden (13.47), saudi (13.00), konflik (12.63), mahasiswa (12.12), trump (11.89), negara (10.05), ditangkap (9.75), tembak (9.12), hamas (8.64), mediator (8.13), arab (8.13), pria (8.12), kecam (8.06), terima (6.70), kasih (6.13), salah (6.12), usul (6.05), visa (5.99), orang (5.57), masjid (5.13), mati (5.13), darurat (5.13), papua (5.12), bawa (5.12), utusan (5.12), intelijen (5.12), umat (4.94)	286.29
3	gaza (210.01), warga (165.47), israel (162.48), palestina (105.88), trump (94.89), relokasi (53.12), tolak (49.82), tepi (41.29), barat (41.29), rencana (30.34), hamas (26.19), arab (25.85), negara (24.58), mesir (21.72), saudi (20.93), serang (20.92), ri (18.08), ide (17.12), as (16.38), bantuan (16.37), yordania (16.12), bikin (15.12), usir (14.99), militer (14.87), usulan (13.83), indonesia (13.42), klaim (13.16), jalur (13.00), iran (11.34), pindahkan (11.12)	1299.70
4	israel (177.94), serangan (111.12), gaza (94.06), tewas (89.84), orang (58.12), gempur (27.07), as (23.87), militer (23.53), konflik (23.13), yaman (23.12), tewaskan (22.12), houthi (21.47), hamas (20.59), akibat (20.33), trump (19.70), serang (18.96), korban (16.88), barat (15.12), tepi (15.12), udara (13.12), suriah (11.55), anak (10.97), kirim (10.72), tentara (9.32), klaim (8.91), lebanon (8.75), rudal (8.09), roket (7.97), perempuan (7.84), politik (7.13)	926.47
5	israel (184.02), hamas (152.75), sandera (126.12), gaza (80.19), gencatan (77.88), bebaskan (72.12), senjata (69.34), palestina (54.42), tahanan (45.90), trump (42.77), dibebaskan (42.53), as (38.76), serahkan (18.12), pembebasan (17.12), tunda (17.12), bantuan (14.54), ancam (13.83), tahap (13.62), jenazah (11.84), netanyahu (11.32), fase (10.12), daftar (9.05), tentara (8.87), perjanjian (8.70), dunia (8.49), langgar (8.13), bombardir (8.12), israelhamas (7.37), negosiasi (7.13), pulang (7.12)	1187.41
6	internasional (36.12), terpopuler (34.12), berita (34.12), gaza (31.03), palestina (26.05), pbb (22.04), israel (19.92), warga (14.72), as (13.61), trump (12.60), erdogan (11.13), prabowo (11.05), ham (10.12), bantuan (9.59), demo (9.02), perdamaian (8.01), menlu (7.89), columbia (7.21), ramadan (7.13), kemerdekaan (7.13), dewan (7.12), sekjen (7.10), universitas (6.60), ri (6.38), propalestina (6.37), buka (5.13), hnw (5.13), pasokan (5.13), papua (5.13), bela (5.12)	391.84
7	konflik (39.95), palestina (33.16), israel (32.10), negara (19.75), arab (15.40), gaza (12.74), warga (12.48), trump (9.33), tampung (9.12), gelar (9.00), netanyahu (8.51), perang (8.26), hamas (8.10), respons (7.13), selatan (7.12), dukung (6.56), harimau (6.13), manusia (6.10), solusi (5.92), indonesia (5.85), saudi (5.80), mesir (5.53), as (5.50), pemimpin (5.26), tolak (5.17), ihsg (5.13), al (5.13), singgung (5.12), kemlu (5.12), liga (5.12)	315.60
8	gaza (45.23), ambil (39.12), israel (38.18), trump (33.98), alih (33.12), hamdan (21.12), ballal (21.12), sutradara (20.12), rencana (19.90), diserang (18.12), palestina (17.90), oscar (15.12), other (14.12), land (14.12), no (14.12), netanyahu (13.36), konflik (11.33), hilang (9.13), bombardir (9.13), ngotot (9.12), pemukim (8.12), dunia (8.07), pm (7.79), jalur (7.63), perang (7.19), barat (6.39), tepi (6.39), peraih (6.12), pemenang (6.12), as (5.50)	486.84

Table 5 shows that keywords such as gaza, ambil, israel, and trump strongly shape Topic 8, indicating a narrative focused on geopolitical tension, conflict, and international reactions.

The appearance of terms like sutradara and oscar suggests that some articles link the conflict to cultural or media contexts, reflecting a broader narrative scope. These diverse keywords confirm that Topic 8 captures a multidimensional view of the Palestine–Israel issue. Figure 3, through the Intertopic Distance Map, visualizes the semantic distance and clustering of topics, revealing how distinct or overlapping they are. This mapping helps verify the thematic identity of the best-performing topic and offers an intuitive picture of topic density and dispersion, thereby strengthening the robustness of the LDA model.

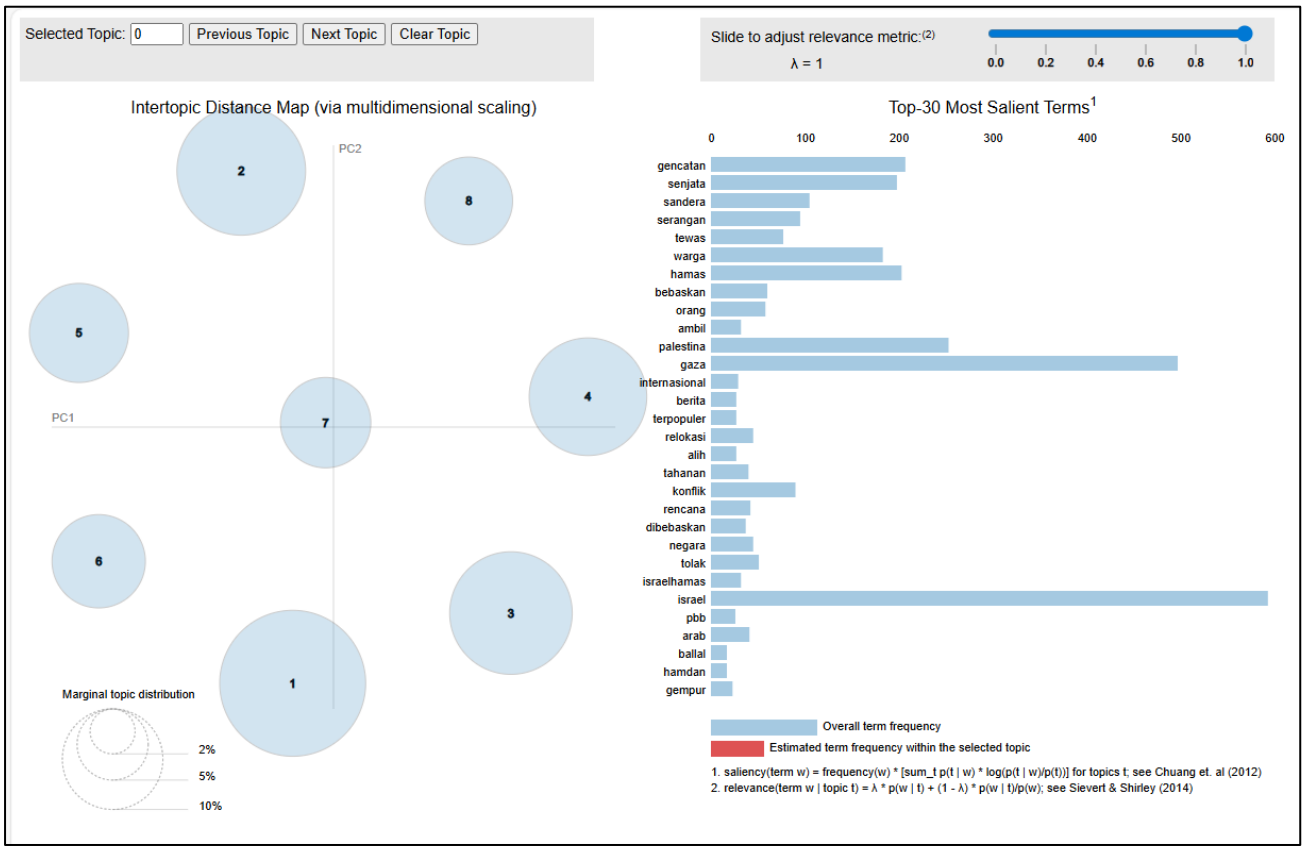

Figure 3. Intertopic Distance Map (LDA)

Figure 3 shows clear separation among several topics, indicating strong thematic differentiation, while Topic 8 occupies a stable position with minimal overlap, reinforcing its distinct semantic identity. Some clusters appear closer together, suggesting shared themes within certain segments of the news coverage and reflecting both focused and diverse narratives surrounding the Palestine–Israel conflict. This distribution confirms that the LDA model has produced a well-structured representation of topics. Figure 4 then presents a Word Cloud highlighting the most frequent words in each topic, allowing quick visual identification of dominant terms and offering an intuitive, non-technical understanding of topic structure. As an exploratory tool, the word cloud reveals patterns not easily captured in numerical tables and strengthens the interpretive process by providing visual context for word distribution.

Figure 4. LDA Word Cloud per Topic

Figure 4 shows that words such as *gaza*, *israel*, *palestina*, and trump dominate the word cloud, highlighting a strong emphasis on conflict and geopolitical issues. The varying word sizes reflect their relative importance in topic formation, while the presence of terms like oscar, universitas, and demo indicates that media coverage also touches on social, educational, and cultural dimensions. This suggests that the news narratives extend beyond direct conflict to include broader contextual themes. Overall, the word cloud strengthens the understanding of thematic diversity within the dataset.

BERTopic Analysis Results

The BERTopic analysis was conducted as a complement to LDA, leveraging Transformer-based embeddings combined with UMAP for dimensionality reduction and HDBSCAN for clustering, allowing topics to be formed based on semantic similarity rather than simple word frequency. Table 6 presents coherence and perplexity scores to evaluate the quality and stability of the topics generated by BERTopic. Coherence measures the semantic relatedness of key terms within each topic, while perplexity reflects the model's predictive reliability in mapping word distributions. Assessing both metrics together enables an objective comparison of topic quality and helps determine whether BERTopic provides a more stable topic distribution than LDA. Overall, this table serves as a crucial benchmark for systematically interpreting and validating the significance of the topics produced in the media text analysis.

Table 6. Coherence Score and Perplexity Results Using BERTopic

Topic	Coherence Value	Perplexity
Topic 0	0.446	3.4729
Topic 1	0.4235	2.8104
Topic 2	0.3871	3.0784
Topic 3	0.4438	2.8376
Topic 4	0.2865	2.9173
Topic 5	0.5086	2.6397
Topic 6	0.5762	2.3891
Topic 7	0.3041	2.6967
Topic 8	1.0	2.9268
Topic 9	0.3254	2.4475
Topic 10	0.9376	1.7851
Topic 11	0.6481	2.1353
Topic 12	0.3321	2.1163
Topic 13	0.6108	2.0458
Topic 14	0.1754	2.3623
Topic 15	0.4793	2.455
Topic 16	0.5179	2.1096
Topic 17	0.3237	2.2782
Topic 18	0.2202	2.2782

Based on Table 6, Topic 8 shows the highest coherence value, while Topic 10 records the lowest perplexity, indicating strong semantic quality and high predictive accuracy in BERTopic's results. The variation across both metrics demonstrates that BERTopic generates a diverse thematic structure reflecting the complexity of Palestine–Israel news coverage and the differing degrees of semantic consistency among topics. Overall, the table confirms that BERTopic delivers competitive and stable topic quality compared to LDA. Figure 5 visualizes the coherence values for all topics, allowing readers to identify semantically robust topics quickly and observe whether topic quality is dominated by a few topics or distributed more evenly. This visualization highlights the reliability of BERTopic's semantic embeddings and supports the methodological validation of topic selection for further analysis.

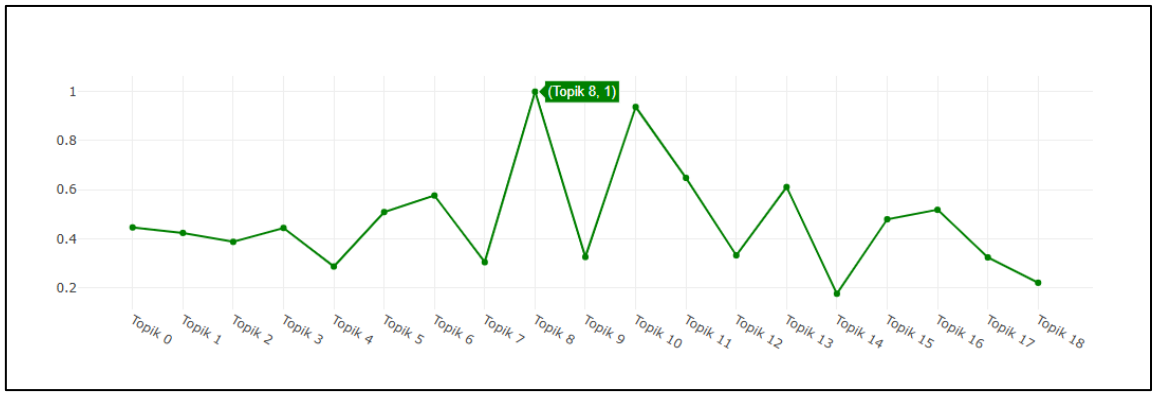

Figure 5. Comparison of 19 Topics Based on Coherence Score (BERTopic)

Figure 5 shows considerable variation in BERTopic coherence scores, with Topic 8 achieving the highest value, indicating strong semantic connectedness among its keywords. Several topics display lower coherence, suggesting broader or less focused themes, yet the overall trend confirms that BERTopic effectively captures multiple semantically solid clusters reflecting the complex narrative landscape of the news coverage. Figure 6 then visualizes perplexity values across topics to assess the model's predictive reliability, revealing whether certain topics demonstrate stronger or weaker predictive performance. This distribution helps identify outliers and evaluate the stability and accuracy of BERTopic's topic structures. Together, these visualizations complement the numerical metrics and strengthen the overall validation of BERTopic prior to deeper thematic interpretation.

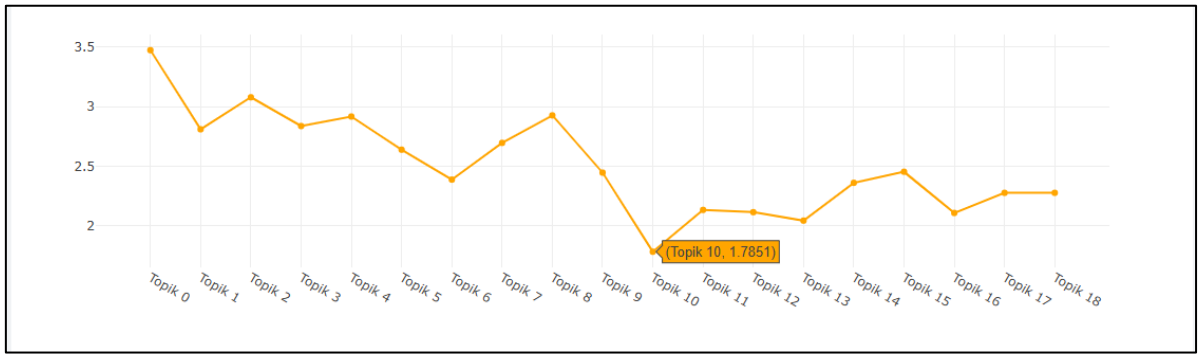

Figure 6. Comparison of 19 Topics Based on Perplexity (BERTopic)

Figure 6 shows that Topic 10 has the lowest perplexity value, indicating the strongest predictive performance and a more focused semantic structure, while other topics with higher perplexity reflect broader or more complex themes captured by the model. This pattern confirms that BERTopic generates several predictively stable topics suitable for deeper analysis. Table 7 then provides a comprehensive overview of the keywords forming all 19 BERTopic topics, produced through transformer embeddings, UMAP dimensionality reduction, and HDBSCAN clustering. By examining these keywords and their contribution weights, readers can understand each topic's semantic structure and thematic focus whether global, political, or social making the table a key foundation for narrative interpretation and a benchmark for comparing BERTopic and LDA results.

Table 7. BERTopic Topic Keywords

Topic ID	Keywords	Total Score
Topic 0	konflik (0.07), trump (0.05), gaza (0.04), as (0.03), warga (0.03), relokasi (0.03), tolak (0.02), ambil (0.02), alih (0.02), houthi (0.02)	0.34
Topic 1	tewas (0.13), orang (0.11), serangan (0.08), israel (0.07), gaza (0.06), akibat (0.05), tewaskan (0.05), bunuh (0.04), palestina (0.04), korban (0.03)	0.66
Topic 2	palestina (0.13), warga (0.06), presiden (0.05), trump (0.05), tolak (0.04), kemerdekaan (0.04), gaza (0.04), erdogan (0.04), relokasi (0.03), usir (0.03)	0.51
Topic 3	senjata (0.15), gencatan (0.15), israelhamas (0.13), hamas (0.05), israel (0.04), perjanjian (0.04), kesepakatan (0.04), gaza (0.04), isi (0.03), berlaku (0.03)	0.71
Topic 4	israel (0.07), tepi (0.07), barat (0.07), listrik (0.06), pasokan (0.06), gaza (0.06), militer (0.04), putus (0.04), bantuan (0.04), blokir (0.04)	0.55
Topic 5	sandera (0.21), bebaskan (0.14), hamas (0.13), dibebaskan (0.10), israel (0.08), tawanan (0.06), gencatan (0.04), sabtu (0.04), senjata (0.04), jenazah (0.04)	0.87
Topic 6	tahanan (0.28), bebaskan (0.14), palestina (0.14), sandera (0.10), dibebaskan (0.09), israel (0.07), penjara (0.06), ratusan (0.06), ditukar (0.06), hamas (0.06)	1.06
Topic 7	netanyahu (0.25), menteri (0.08), gencatan (0.06), trump (0.06), bertemu (0.06), lanjutkan (0.06), alih (0.06), ambil (0.05), senjata (0.05), shin (0.05)	0.78
Topic 8	terpopuler (0.94), berita (0.94), internasional (0.93), (0.00), (0.00), (0.00), (0.00), (0.00), (0.00), (0.00)	2.81
Topic 9	ramadan (0.18), masjid (0.17), boikot (0.09), terafiliasi (0.09), aqsa (0.08), batasi (0.07), protes (0.07), al (0.07), dmi (0.06), demonstran (0.06)	0.94
Topic 10	sutradara (0.24), ballal (0.21), hamdan (0.21), no (0.18), other (0.18), land (0.18), oscar (0.18), diserang (0.15), pemukim (0.11), hilang (0.10)	1.74
Topic 11	saudi (0.32), negara (0.19), netanyahu (0.17), arab (0.15), normalisasi (0.12), palestina (0.11), usul (0.09), pembentukan (0.08), dirikan (0.07), kecam (0.06)	1.34
Topic 12	lebanon (0.21), rudal (0.21), yaman (0.16), roket (0.16), cegat (0.13), hizbullah (0.13), balas (0.08), houthi (0.08), israel (0.07), klaim (0.07)	1.31
Topic 13	mahasiswa (0.25), columbia (0.19), propalestina (0.16), visa (0.14), universitas (0.14), indonesia (0.13), demo (0.12), palestina (0.09), kampus (0.08), as (0.08)	1.38
Topic 14	muslim (0.18), umat (0.12), when (0.11), phone (0.11), rings (0.11), the (0.10), sekjen (0.07), kasih (0.07), aqsa (0.07), islamofobia (0.06)	0.99
Topic 15	bombardir (0.12), gencatan (0.12), serang (0.12), senjata (0.11), bom (0.10), israel (0.08), gaza (0.07), langgar (0.07), chinarusia (0.07), murka (0.05)	0.89
Topic 16	arab (0.25), rekonstruksi (0.17), liga (0.17), saudi (0.17), komite (0.10), biayai (0.10), cs (0.09), rencana (0.09), sepakat (0.08), otoritas (0.08)	1.31
Topic 17	indonesia (0.39), ri (0.14), relokasi (0.13), trump (0.10), warga (0.08), rs (0.08), anak (0.08), wujudkan (0.07), pemantapan (0.07), pengaruhnya (0.07)	1.22
Topic 18	malaysia (0.22), thailand (0.15), turis (0.14), bisnis (0.12), onar (0.12), amm (0.07), melunak (0.07), asean (0.07), blacklist (0.07), dilancarkan (0.07)	1.11

Table 7 shows that each BERTopic topic has distinct keyword characteristics, reflecting a wide thematic range in Palestine–Israel news coverage for instance, Topic 6 focuses on detainees and hostages, while Topic 13 highlights students, campuses, and international protests. These patterns demonstrate that BERTopic identifies more specific and granular themes than LDA, with contribution weights indicating the dominance and relevance of each keyword. The themes span political, social, humanitarian, and diplomatic narratives, reinforcing the model's effectiveness in mapping media discourse in detail. Figure 7 presents

the Intertopic Distance Map, illustrating semantic proximity among topics and revealing whether themes are closely related or clearly differentiated. This visualization helps assess topic clustering, thematic fragmentation, and the overall structure produced by transformer-based embeddings, serving as a key basis for evaluating BERTopic's modeling quality relative to LDA.

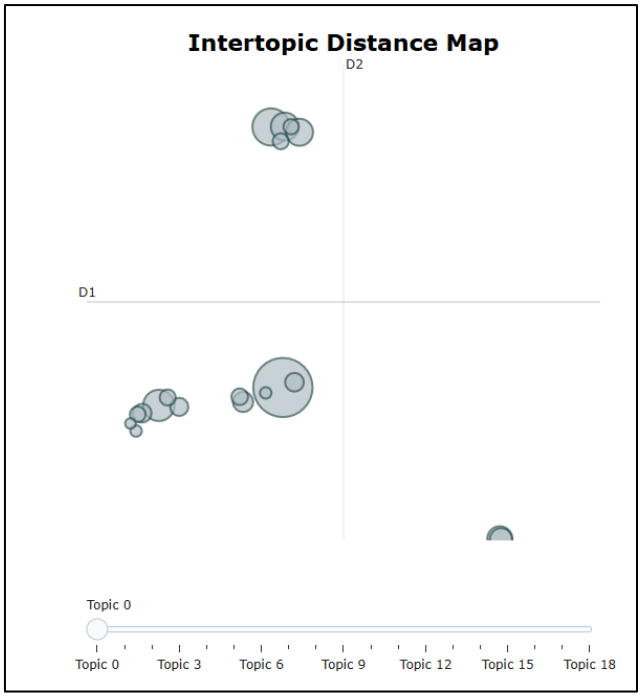

Figure 7. Intertopic Distance Map (BERTopic)

Figure 7 shows that BERTopic topics form several well-defined clusters, with some positioned closely indicating thematic similarity while others appear farther apart, reflecting highly specific and distinct themes. This distribution demonstrates strong thematic differentiation and confirms that the model captures substantial variation within the media discourse. Figure 8 then presents the Word Cloud for each topic, visually highlighting dominant words based on size and complementing the keyword table by reinforcing each topic's semantic structure. Visualization helps reveal word relationships and thematic intensity that may not be fully visible in numerical data, making it a valuable tool for quickly understanding narrative patterns and enhancing the interpretation of the topic modeling results.

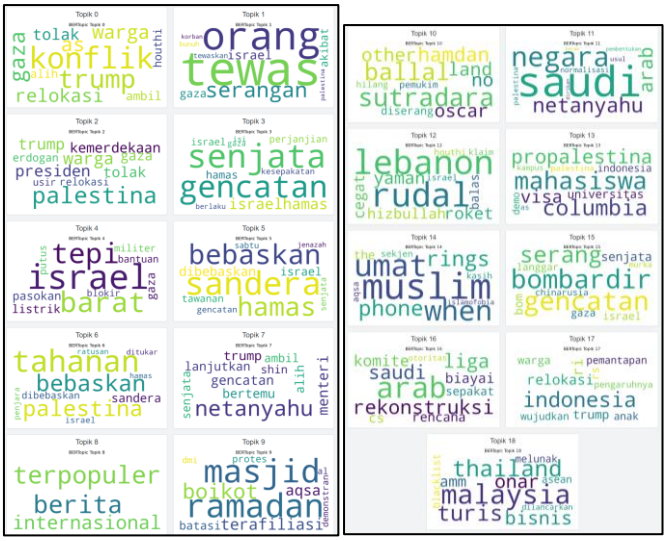

Figure 8. BERTopic Word Cloud Per Topic

Figure 8 shows that terms such as *gaza*, *israel*, *palestina*, and *serangan* dominate multiple BERTopic topics, highlighting a strong focus on conflict, humanitarian issues, and diplomatic dynamics, while more specific terms like *lebanon*, *oscar*, and *columbia* reflect political, social, and cultural diversity in the news coverage. This visualization reinforces that BERTopic captures a broader and more detailed thematic range than LDA and helps clarify the semantic relevance of terms within each topic. Figure 9 then presents the Word Scores, displaying contribution weights that more precisely define each topic's composition. This visualization enables researchers to evaluate semantic strength, compare the influence of dominant terms, and understand how key words shape cluster formation. The varying score intensities further highlight differences in topic depth, making this figure an essential component for analytically validating the semantic quality of the BERTopic model.

That each BERTopic topic contains keywords with varying contribution scores, reflecting clear structural differences across topics. High-scoring terms such as *tahanan*, *bebaskan*, and *mahasiswa* highlight the dominant themes within specific clusters and demonstrate strong semantic consistency. The variation in score intensity further indicates that the model captures thematic depth according to the complexity of the underlying narratives. These results confirm that BERTopic provides more granular and detailed topic representations than LDA. Overall, the graph plays an essential role in supporting accurate and valid thematic interpretation.

The evaluation section requires a more structured comparison to objectively assess model performance; therefore, a table is developed to compare LDA and BERTopic across key metrics such as coherence, perplexity, topic depth, granularity, and interpretability. This table provides a concise yet comprehensive overview of each model's strengths and limitations, allowing readers to directly observe differences without relying solely on narrative explanation. It also offers an empirical basis for determining which algorithm is more suitable for analyzing large-scale topics like the Palestine–Israel conflict. By presenting the evaluation in tabular form, the assessment process becomes clearer, more transparent, and academically rigorous, thereby strengthening the validity and accountability of the analytical findings.

Table 8. Comparative Evaluation of LDA and BERTopic

Evaluation Aspect	LDA	BERTopic
Highest Coherence Score	0.5829 (Topic 8)	1.0 (Topic 8)
Lowest Perplexity Score	8.3396	1.7851 (Topic 10)
Number of Topics Generated	20 Topics	19 Topics
Word Connectivity Within Topics	Consistent but more general	Sharper and more detailed due to transformer-based embeddings
Topic Granularity	Moderate (broader themes)	High (more specific and segmented)
Ease of Interpretation	Very easy due to simple statistical structure	Good, but requires deeper semantic interpretation
Suitability for Complex Datasets	Adequate but less sensitive to contextual nuance	Highly suitable due to strong contextual embedding representation
Primary Strength	High interpretability	High semantic accuracy and topic depth
Primary Limitation	Limited ability to capture full contextual meaning	Requires higher computational resources and interpretation effort

The comparative table shows that BERTopic offers clear advantages in generating higher-quality topics, particularly in coherence and thematic granularity, due to its use of transformer-based embeddings capable of capturing richer semantic nuances. However, LDA remains a strong alternative when simplicity, faster interpretation, and computational efficiency are the primary goals, making it suitable for studies requiring a straightforward probabilistic framework. In contrast, BERTopic provides more detailed and specific thematic distinctions, enabling deeper analytical insights into complex discourse. Thus, researchers can choose the model that best fits their methodological needs, with BERTopic performing better for datasets containing diverse and multilayered narratives, while LDA remains valuable for concise and easily interpretable topic modeling.

CONCLUSION

This study shows that the combination of LDA and BERTopic forms a complementary analytical framework in examining the reporting of the Palestinian-Israeli conflict in Indonesian online media. LDA produced 20 top-performing topics in Topic 8 that had high coherence (0.5829) and low perplexity (8.3396), while BERTopic produced 19 top-performing topics, characterized by the highest coherence in Topic 8 (1.0) and the lowest perplexity in Topic 10 (1.7851). Both models consistently identify five main thematic clusters, namely military strikes in Gaza, Indonesian diplomacy, international responses, humanitarian issues, and global political dynamics. Overall, BERTopic is more effective at capturing detailed and contextual topics, while LDA provides a stable probabilistic structure, thus integrating the two improves the accuracy and interpretability of topic mapping. Further research is suggested to develop a more structured hybrid approach, expand data sources across platforms and languages, and integrate sentiment analysis, media bias, and expert validation to enrich the depth of interpretation and thematic relevance.

REFERENCES

Abdurrazzaq, M. A. (2023). Analisis Ulasan Aplikasi MyPertamina Menggunakan Topic Modeling dengan Latent Dirichlet Allocation. *KALBISCIENTIA Jurnal Sains Dan Teknologi*, *10*(1), 1–6. https://doi.org/10.53008/kalbiscientia.v10i1.694

Faudzi, A., & Fajri, M. (2024). Manajemen Produksi Siaran Berita Di Padang Tv. *RETORIKA : Jurnal Kajian Komunikasi Dan Penyiaran Islam*, *6*(2), 59–74. https://doi.org/10.47435/retorika.v6i2.3125

Hajra Chaudhry, D. M. R. (2024). Ethical Dilemmas In Media Coverage Of Israel-Palestine Conflict: Analysis Of The New York Times. *Migration Letters*, *21*(S11), 72–83.

Hartini, T., Puspasari, C., Jafaruddin, Harinawati, & Hasan, K. (2024). Framing Pemberitaan Konflik Israel-Palestina dalam Harian (E-Paper) Kompas. *CENDEKIA: Jurnal Hukum, Sosial & Humaniora*, *2*(4), 775–788.

Helmayanti, S. A., Hamami, F., & Fa'rifah, R. Y. (2023). Penerapan Algoritma Tf-Idf Dan Naïve Bayes Untuk Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Flip Pada Google Play Store. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, *4*(3), 1822–1834. https://doi.org/10.35870/jimik.v4i3.415

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.

Linstead, E., Lopes, C., & Baldi, P. (2008). An application of latent Dirichlet allocation to analyzing software evolution. *2008 Seventh International Conference on Machine Learning and Applications*, 813–818.

Manurung, F., & Heriamsal, K. (2024). 19 Strategi Diplomasi Indonesia Dalam Upaya Mewujudkan Perdamaian Pada Konflik Terbaru Hamas-Israel. *Jurnal Hubungan Luar Negeri*, *9*(1), 30.

Moch Erreza, M. E., Kartini, K., & Rizki, A. M. (2024). Pencarian Topik Penelitian Pada Studi Kasus Jurnal Jifti Menggunakan Teknik Hiearchical Dirichlet Processes. *Spirit*, *16*(1), 170–182. https://doi.org/10.53567/spirit.v16i1.325

Mughni Sulubara, S., Lestaria, E., Sempena, D., Humaira, D., Mawaddah, L., Ulan Dari, M., Ulfa, N., Wulandari, N., Bahgia, P., Roza, R., Supia, R., Dewi, Y., & Muhammadiyah Mahakarya Aceh Jalan Gayo Simpang Bireun Nomor, U. I. (2024). Perlindungan Hukum Internasional Tentang Konfik Perang Lintas Negara Antara Palestina Dan Israel. *Maret*, *2*(1), 358–366.

Nuraliza, V., Andhi Nur Rahmadi, Alvan Mubaroq, Kristiyono Kristiyono, Alisyia Putri Melani, & Anila Ifana. (2024). Peran Komunikasi Politik Dalam Membentuk Opini Publik Menghadapi Pemilu 2024. *CENDEKIA: Jurnal Ilmu Sosial, Bahasa Dan Pendidikan*, *4*(1), 245–261. https://doi.org/10.55606/cendikia.v4i1.2514

Nursyahrina, Defit, S., & Sovia, R. (2024). Metode BERTopic dan LDA untuk Analisis Tren Penelitian Bidang Ilmu Komputer. *Jurnal KomtekInfo*, *11*(4), 332–341. https://doi.org/10.35134/komtekinfo.v11i4.580

Ramadani, Mutiara. S., Khaerudin Kurniawan, & Ahmad Fuadin. (2024). Menguak Bias Media dalam Pemberitaan Konflik Israel-Palestina: Sebuah Analisis Konten Kritis. *Jurnal*

Onoma: Pendidikan, Bahasa, Dan Sastra, *10*(1), 887–905. https://doi.org/10.30605/onoma.v10i1.3392

Samsir, Reagan Surbakti Saragih, Selamat Subagio, Rahmad Aditya, R. W. (2023). Jurnal Media Informatika Budidarma Bertopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory. *Jurnal Media Informatika Budidarma*, *7*(3), 1514–1520. https://doi.org/10.30865/mib.v7i3.6426

Setiawan, I. (2024). Eskalasi Konflik Palestine-Israel di Tahun 2023: Perspektif Kebijakan Luar Negeri Indonesia. *Jurnal Hubungan Internasional*, *17*(1), 248–263. https://doi.org/10.20473/jhi.v17i1.52392

Suhayatmi, Rahmatulummah, A., & Resky, S. A. (2024). Eskalasi Konflik Iran-Israel di Damaskus: Implikasi terhadap Stabilitas Keamanan Regional dan Global. *Jurnal Hubungan Luar Negeri*, *9*(1), 49–68. https://doi.org/10.70836/jh.v9i1.49

Syahroni, R., Rusmana, D., Thohir, A., Salam, M., & Hidayat, A. A. (2025). Respons KH. Saifuddin Zuhri Terhadap Pendudukan Yahudi di Palestina Tahun 1936-1948. *Jurnal Ilmiah Multidisiplin*, *1*(4), 2407–2424.

Tran, B. X., Nghiem, S., Sahin, O., Vu, T. M., Ha, G. H., Vu, G. T., Pham, H. Q., Do, H. T., Latkin, C. A., & Tam, W. (2019). Modeling research topics for artificial intelligence applications in medicine: latent Dirichlet allocation application study. *Journal of Medical Internet Research*, *21*(11), e15511.

Wastanti, R., & Wiratama, A. (2024). Peran Redaktur dalam Menentukan Headline Halaman Utama pada Surat Kabar Harian Lampu Hijau: Editorial Decision-Making in Headline Selection for the Front Page of the Lampu Hijau Daily Newspaper. *Indonesian Scholar Journal of Communication (ISJC)*, *2*(02), 62–70.

Zahoor, M., & Sadiq, N. (2021). Digital public sphere and Palestine-Israel conflict: A conceptual analysis of news coverage. *Liberal Arts and Social Sciences International Journal (LASSIJ)*, *5*(1), 168–181.