REKAYASA
Journal of Science and Technology
https://journal.trunojoyo.ac.id/rekayasa

REKAYASA ,19(1): 2026 pp.96-108
ISSN: 0216-9495 (Print)
ISSN: 2502-5325 (Online)

Analysis of Public Topics on The Indonesian House of Representatives in The 17+8 Demonstration Demands in News and Social Media Using BERTopic

Anugrah Marta Wijaya[1*)], Doni Abdul Fatah[2)] Yeni Kustiyahningsih[3)]
[1, 2, 3] Universitas Trunodjoyo Madura, Indonesia
*email: 200441100130@student.trunojoyo.ac.id

DOI: https://doi.org/10.21107/rekayasa.v19i1.33699
Submitted March, 6th 2026, Accepted April, 27th 2026, Published May, 8th 2026

Abstract

Social media has become the main space for the public to voice their opinions and criticisms of government institutions. One issue that has received a lot of attention is the 17+8 demonstration demanding accountability from the Indonesian House of Representatives (DPR RI) as a legislative body. This study aims to map public opinion of the DPR RI during this momentum using a BERTopic-based topic modeling approach. The research data consists of 21,195 comments from YouTube, TikTok, Instagram, and Facebook. The analysis process included scraping, text preprocessing, semantic representation formation using BERT embedding, dimension reduction with UMAP, and clustering using HDBSCAN. The modeling results produced more than 90 topics covering key issues such as corruption, the role of students, legislative performance, and security forces. Model evaluation showed an intra-topic cohesion value of 0.778, inter-topic separation of 0.412, and topic diversity of 0.749, indicating that the topics formed were coherent, diverse, and contextual. This study proves that BERTopic is quite effective in analyzing informal public opinion and can be used as a basis for developing digital political communication studies in Indonesia.

Keywords: 17+8 Demonstrations, BERTopic, Indonesian House of Representatives, Public Opinion, Topic Modeling Analysis.

© 2026 Authors

INTRODUCTION

The Indonesian government system adheres to the principle of trias politica, which divides state power into three institutions, namely the executive, legislative, and judicial branches, with sovereignty residing in the hands of the people as mandated in the 1945 Constitution of the Republic of Indonesia (Mazjah & Kurdi, 2025). In this constitutional structure, the principle of separation of powers places the House of Representatives of the Republic of Indonesia (DPR RI) as one of the main pillars of the government system, namely as a legislative body that represents the sovereignty of the people (Setiawan et al., 2025) (Edwin Hendrik et al., 2024) (Jayadi et al., 2022). DPR RI also has the function of forming laws, overseeing the implementation of government, determining the State Revenue and Expenditure Budget, and channeling the aspirations of the people as their representatives (Mengko, 2024) (Sugiman, 2020). However, in practice, there is often a gap between public expectations and the reality of the institution's performance, which triggers public dissatisfaction. Various actions and behaviors of DPR members are considered to be increasingly distant from the interests of the people, causing widespread anger that is exploited by various parties, culminating in a wave of demonstrations in a number of major cities. These prolonged protests even escalated into incidents of arson of public facilities, clashes with the police, and looting of certain political symbols. The climax of these protests, with demands known as 17+8, became a moment of crystallization of the accumulated public dissatisfaction

with the DPR RI (Anantari, 2025). This movement reflects the failure of the legislative body to represent the aspirations of its constituents and marks a crisis of public confidence in representative institutions.

The decline in public trust in formal channels of representation has led to the emergence of social media as an influential alternative public space for Indonesians to express their political dissatisfaction and obtain information about demonstrations (Anggiawan, 2023) (Rizqi Nandadita Pamungkas et al., 2024) (Hendra & Halbadika Fahlevi, 2024). This digital space not only serves as a channel for aspirations that transcend institutional boundaries, but also functions as an arena for contestation between various actors such as political buzzers, provocateurs, and influencers who actively frame narratives, mobilize public emotions, and shape mass opinion regarding the performance of the DPR RI. Through various content formats, ranging from videos on TikTok and YouTube to posts on Instagram and Facebook, complex issues from the 17+8 demonstrations are often summarized with emotional content and disseminated rapidly, creating an exaggerated effect. As a result, the center of public debate has shifted from formal spaces to narrative battles on social media. (Sari et al., n.d.). In this digital space, facts are often mixed with emotions and fragmented information, creating an increasingly polarized political communication environment (Dinda Aurellia & Katimin, 2025). Although complex, this phenomenon is unavoidable and has a significant impact on public perception of the performance of the DPR RI.

In dealing with the complexity and large volume of text data generated in the digital space, Topic Modeling is one of the computational techniques in the field of Natural Language Processing (NLP) designed to extract hidden themes or topics (latent themes) from a text corpus automatically and unsupervised (Hankar et al., 2025) (Romero et al., 2024). Unlike manual content analysis, which is time-consuming, labor-intensive, and prone to subjectivity, this technique works by analyzing word distribution patterns in a set of documents to group texts with similar meanings into coherent topics. Each topic is then represented by a set of the most prominent keywords (EKİNCİ & İLHAN OMURCA, 2020) (Silva et al., 2021). The ability of Topic Modeling to map broad discourse landscapes and identify underlying thematic structures makes it an important tool for analyzing conversation trends, issue evolution, and public discussion focus in large data sets such as social media comments, online news, and policy documents. Thus, this approach provides a deep understanding that is difficult to achieve through traditional analysis methods (Kousis & Tjortjis, 2023) (Wahyuni et al., 2025).

Although traditional methods such as Latent Dirichlet Allocation (LDA) have long been the standard in topic modeling, this approach has fundamental weaknesses because it operates at the word-level statistics with a bag-of-words that treats each word as an independent unit without considering the linguistic context and semantic relationships between words (Zoya et al., 2021) (Venugopalan & Gupta, 2022). This statement causes LDA to often produce topics that are semantically incoherent (Rosalinda et al., 2023) (Cici Suhaeni, 2025) (Nursyahrina et al., 2024), especially when dealing with data in informal language such as social media comments, where the meaning of words is highly dependent on the context in which they are used. In response to these limitations, BERTopic has emerged as a cutting-edge method that leverages the revolution in contextual language models such as BERT (Bidirectional Encoder Representations from Transformers). Through the embedding process, each document is converted into a high-dimensional vector that represents its contextual meaning, so that words with similar meanings will have representations that are close to each other in semantic space. These vectors are then grouped using the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm, which is capable of identifying clusters with varying densities and shapes while separating data considered to be outliers. The combination of BERT's semantic understanding and the power of HDBSCAN clustering produces topics that are more coherent, interpretive, and resilient to noise and language variation in contemporary texts.

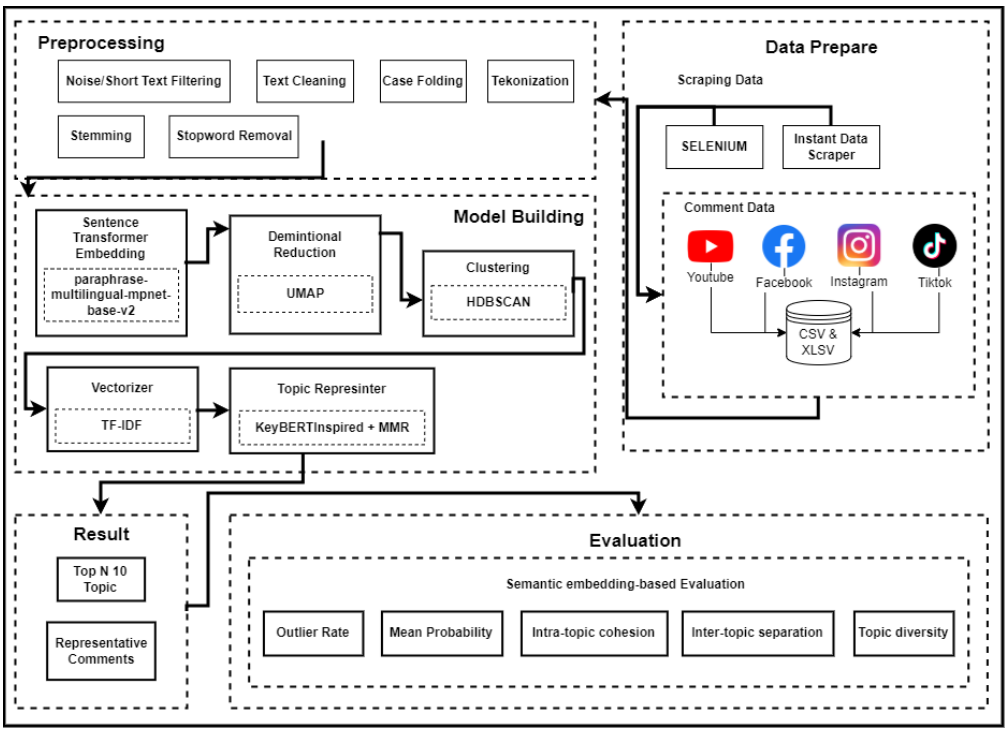
In the study by Maria H. M. et al., the authors explicitly state that LDA performs poorly on short texts because the topics generated are merely a multinomial distribution of words, and topic identification relies on "top words" without context, often resulting in a loss of semantic meaning. This makes LDA less reliable for social media comments (Hernandez-Mendoza et al., 2022). This finding is in line with Ika W. et al., who

reported that the baseline LDA coherence was only 0.336 and increased to 0.431 after refinement, an increase of 28.27%, confirming that LDA easily deteriorates in short texts and requires semantic enrichment to be interpretable (Widiastuti & Yong, 2025). On the other hand, the BERTopic variant (BERT + kernelPCA + k-means) in the study by Bayode O. et al. produced a coherence of 0.76 for 10 topics and 0.8463 when the number of topics/terms was not limited, demonstrating high performance on industrial corpora (Ogunleye et al., 2023). Tong Li et al. In a large corpus with n=4,504, they extracted 9 topics with inter-topic similarity ranging from 0.27 to 0.91 and evidence of sharp thematic representation through c-TF-IDF scores, for example, 'Fire' 0.118 in Topic 7, which indicates BERTopic's ability to capture both semantic proximity and separation (T. Li et al., 2025). Another advantage highlighted by Chenchen Li et al. is that BERTopic can automatically identify the number of topics, unlike LDA, which requires manual setting. Thanks to BERT, it captures deep semantic and contextual information, thereby improving the accuracy of text representation (C. Li & Hu, 2025). Consistent with this statement, Li Ma et al. show UMass metrics where LDA with 18 topics ranges from −3.0060 to −1.3128, while BERTopic with 21 topics ranged from −1.0587 to −0.0531, closer to zero and therefore more coherent, providing evidence of BERTopic's superiority over LDA (Ma et al., 2025).

Based on this background, this study aims to map the landscape of public opinion regarding the Indonesian House of Representatives at the peak of the 17+8 protests in a more in-depth and contextual manner, by computationally analyzing a collage of comments from various social media platforms, namely YouTube, TikTok, Instagram, and Facebook, using a cutting-edge topic modeling approach, namely BERTopic. More specifically, this study not only seeks to identify and characterize the dominant themes that emerge, but also assesses the coherence and distribution of these topics, as well as interpreting the meaning and semantic nuances of each opinion cluster that is formed. By utilizing a combination of the BERT contextual language model to understand the meaning of words in sentence context, UMAP dimension reduction to preserve the semantic structure of the data, and the HDBSCAN clustering algorithm, which is capable of grouping data based on density and identifying outliers, this discourse mapping is expected to capture a deeper and more complex discourse structure, going beyond traditional approaches that rely solely on word frequency. The final results of this study are expected to provide a comprehensive, systematic, and context-rich picture of key issues of public concern regarding the performance of legislative institutions amid high political pressure, thereby providing an empirical basis for stakeholders to gain a more complete understanding of public aspirations.

METHOD

This study uses a quantitative-exploratory approach based on text mining, which begins with the collection of public comments from various social media platforms, followed by a series of text data cleaning and transformation processes to prepare semantic embeddings. The methodological flow then continues to dimension reduction and clustering to group comments into coherent topics, whose quality and diversity are ultimately evaluated comprehensively through a series of semantic metrics. Overall, this process is designed to systematically map the structure of public discourse, as illustrated in the Research Method Flowchart in Picture 1 below.



Picture 1. Research Flowchart

A. *Data Collection*

	Author	Comment	Source
0	nareswari15	cari tau gaes tanggapan benix dan Silvia tjan ...	TikTok
1	mpuuts45	Siapa sih yg bikin ini?? Kok aneh dan ga masuk...	TikTok
2	aliphildan	bismillaah ayo kawal gaiis 🔥 ✨	TikTok
3	baiqmitawardani	kukira tuntutan salsa itu.. ternyata bu kaaann...	TikTok
4	farhanzubedi	Lets go! 🔥 🔥 🔥	TikTok

Picture 2. Example Dataset

The initial stage of the research was to collect data on comments related to the 17 + 8 demonstration demands made on various social media platforms, namely YouTube, Instagram, Facebook, and TikTok. The data was collected using web scraping techniques, which is the process of automatically retrieving information from web pages without the need for manual copying (Khder, 2021) (Ayuso et al., 2024) (Uzun, 2020). Data collection was performed using the Selenium library in the Python programming language to extract comments from dynamic pages (Çiplak & Yildirim, 2026) (Teotia et al., 2023), supported by the Instant Data Scraper extension to accelerate the process of collecting data from pages with simple structures.

The comments collected focused on the content of the 17 + 8 demands in the form of videos, news posts, and public discussion posts. The data sources are divided into two main categories, namely news media such as Kompas TV, CNN Indonesia, tvOne News, iNews TV, and others that broadcast coverage or public opinion, as well as Indonesian public figures and influencers such as Ferry Irwandi, Jerome Polin, Salsa Erwina, and other content creators who actively discuss social and political issues.

To maintain the relevance and consistency of Indonesia's socio-political context, the data collected was limited to comments from users in Indonesia. The scraping results produced 21,195 comments, which were then saved in CSV and XLSX formats for easy data processing. This dataset includes various forms of informal language, abbreviations, and emotional expressions typical of social media. An example of the dataset can be seen in Picture 2 above.

B. ***Preprocessing***

The preprocessing stage is carried out to convert the data into clean text that is ready for processing. Since the data comes from social media comments, much of the text is varied and informal. Therefore, the following stages are carried out.

1) *Noise / Short Text Filtering*

Filtering out comments that are very short by deleting them (Helan & Sultani, 2023).

2) *Text Cleaning*

Remove irrelevant characters or elements such as URLs, hashtags (#), mentions (@username), numbers, excessive punctuation, and non-alphanumeric characters (Marcelina et al., 2025) (Sejati et al., 2024).

3) *Case Folding*

All letters are converted to lowercase to ensure uniformity in word form (Asani et al., 2024) (Oktavianto et al., 2024)

4) *Tokenization*

Sentences are broken down into word units (tokens), which will be used in the next stage of analysis (Rai & Borah, 2021) (Vayadande et al., 2024).

5) *Stopword Removal*

Removing common words that have no significant contextual meaning, such as "yang" (which), "dan" (and), 'di' (in), "ke" (to), as well as common conjunctions or particles in Indonesian (Rahimi & Homayounpour, 2023) (Alonso-Fernandez et al., 2021).

6) *Stemming*

Return the word to its base form using the Sastrawi library (Nugroho et al., 2023) (Benali, 2024).

The result of this preprocessing stage is clean, uniform text in the form of root words and without noise. The processed data is ready for use in the next stage.

C. ***BERTopic Modeling***

The formation of semantic representation (sentence embedding) is the first step, using the paraphrase-multilingual-mpnet-base-v2 multilingual language model from Sentence Transformers. This model converts each comment into a 768-dimensional vector that represents the contextual meaning of the sentence (Devika et al., 2021) (Polignano et al., 2020). As a result, words with similar meanings have vectors that are close to each other in semantic space. Next, dimensionality reduction by UMAP (Uniform Manifold Approximation and Projection) is used to project high-dimensional vectors into lower-dimensional space without losing much semantic structure (Mittal et al., 2024) (Wani, 2025). The reduction results will be clustered using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which is capable of grouping data with varying densities and cluster shapes, as well as automatically identifying outliers or data that do not belong to any topic (Fuchs & Höpken, 2022) (Stewart & Al-Khassaweneh, 2022), its use is more stable against noise, which is commonly found in social media texts (Hoti et al., 2025) (Asyaky & Mandala, 2021). After clustering, continue vectorization using TF-IDF (Term Frequency–Inverse Document Frequency) to assess the level of importance of words in each topic (Susanto et al., 2023) (Yang et al., 2022). This word vector is then used in the Topic Representer module with a KeyBERTInspired + MMR (Maximal Marginal Relevance) approach to sort out non-redundant keywords for each topic that not only have high weight but are also semantically diverse, so that the resulting topics are easy to interpret (Jiang et al., 2025) (Isoaho et al., 2021) . The final stage produces a list of topics along with key words and representative comments that describe each cluster of public opinion.

D. ***Evaluation***

The evaluation is useful for assessing the quality of the model in terms of semantic depth, cluster stability, and representation of meaning between topics. These values help assess the extent to which the BERTopic model successfully captures the structure of public discourse. The evaluation is based on semantic embedding with several metrics, including intra-topic cohesion, inter-topic separation, topic diversity, mean probability, and outlier rate (Parlak et al., 2025) (Scaringi et al., 2025) (Blair et al., 2020) (Feng et al., 2025).

Through these metrics, model quality is assessed in terms of semantic stability, clarity of boundaries between topics, and variety of themes produced. This embedding-based evaluation approach is more appropriate than traditional metrics such as topic coherence, as it is able to assess the closeness of meaning between documents and between topics based on semantic vector representations resulting from BERTopic modeling rather than LDA with Bag of Words.

RESULTS

After the process was completed, public responses to the demands of the 17+8 demonstrations and the performance of the Indonesian House of Representatives were analyzed to identify the main topics that emerged and the semantic relationships between them. Picture 3 shows the WordCloud results of preprocessing, which illustrates the frequency of words appearing in the comment corpus. Words such as "rakyat", "DPR", "negara", "tuntut", and "Indonesia" appear in the largest size, indicating that public arguments are dominated by discussions about the relationship between the people and the legislative body and demands on the government. Other words such as "aset", "ampas", and "koruptor" indicate strong criticism of corruption and abuse of power.



Picture 3. Wordcloud

	Topic	Count	Name	Representation	keyBERT	MMR	Representative_Docs
0	-1	4243	-1_dpr_rakyat_orang_perintah	[dpr, rakyat, orang, perintah, tuntut, negara,...	[anggota dpr, politik, wakil rakyat, presiden,...	[demo, anggota dpr, menteri, hukum, kuasa, kor...	[kalau duga provokator lalu sebut dukung presi...
1	0	1271	0_indonesia_rakyat_seluruh_baik	[indonesia, rakyat, seluruh, baik, negara, leb...	[indonesia, kaya, bangsa, juang rakyat, negeri...	[indonesia, seluruh rakyat, mahasiswa, negara,...	[tugas rakyat indonesia awal tuntut sebut sela...
2	1	660	1_salsa_mba_mbk_kak	[salsa, mba, mbk, kak, dukung, wakil, erwina, ...	[juang rakyat, wakil suara, rakyat rakyat, wak...	[wakil rakyat, sehat selalu, wakil suara, bera...	[butuh orang kaya mbk salsa had ketua dpr mbk ...
3	2	575	2_koruptor_korupsi_aset_ampas	[koruptor, korupsi, aset, ampas, uu, sahkan, r...	[aset koruptor, koruptor, korupsi, prampasan a...	[aset koruptor, korupsi, koruptor, ampas aset,...	[uu ampas aset koruptor, uu ampas aset korupto...
4	3	480	3_mahasiswa_muda_juang_anak	[mahasiswa, muda, juang, anak, tangsel, rakyat...	[mahasiswa, anak bangsa, rakyat juang, masyara...	[mahasiswa, surat, generasi muda, wakil suara,...	[oih jgn ceritain mahasiswa, terimakasih banya...
...	...	...	...	...	...	...	...
95	94	42	94_jalan_sakit_jelek_cengkraman	[jalan, sakit, jelek, cengkraman, bogor, wlw, ...	[bikin rusuh, ngeri, sakit, rusuh, putus, sahk...	[sakit, ngeri, normal, bikin rusuh, jgn sampe,...	[sangat miris klu benar ungkap air alir cari j...
96	95	36	95_ngeyel_warung_ciri_kwitang	[ngeyel, warung, ciri, kwitang, kapolri, nepal...	[biang kerok, kapolri, kerok, gantung, uud amp...	[klw, kapolri, gantung, pingin, kerok, biang k...	[klw rusus tangkap lgi jarah warung dagang kak...
97	96	29	96_earned_not_given_respect	[earned, not, given, respect, is, emosi, ngera...	[respect, rakyat tuntut, tuntut rakyat, bikin ...	[not, respect, emosi, empati, batu, cerdas, bo...	[trust is earned not given, trust is earned no...
98	97	25	97_mafia_riza_chalid_sarang	[mafia, riza, chalid, sarang, siudin, minyak, ...	[koruptor, rezim, korupsi, rakyat bubar, riza,...	[minyak, proyek, korupsi, prabowo, siapa, pres...	[efek pak presiden incar mafia besar riza chal...
99	98	30	98_chaos_fatantis_plan_twist	[chaos, fatantis, plan, twist, mulus, nge, hur...	[chaos, kacau, kabur, parah, paksa, kalau, dar...	[chaos, rakyat mau, tanggung jawab, darah, kab...	[pantess isi tuntut ngaco segala tni suruh die...

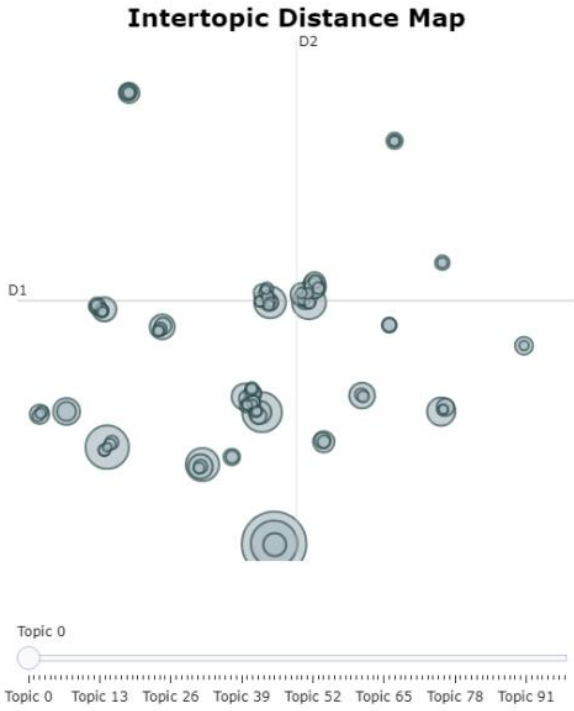
Picture 4. Topic Info

Furthermore, Picture 4 above shows that the model generated 100 topics, with one large cluster labeled topic -1 (outlier) containing ungrouped comments, amounting to approximately 4,243 documents or 28.19% of the total data. Meanwhile, other topics such as Topic 0, 1, 2, and 3 have the most comments and form the core of public discourse.

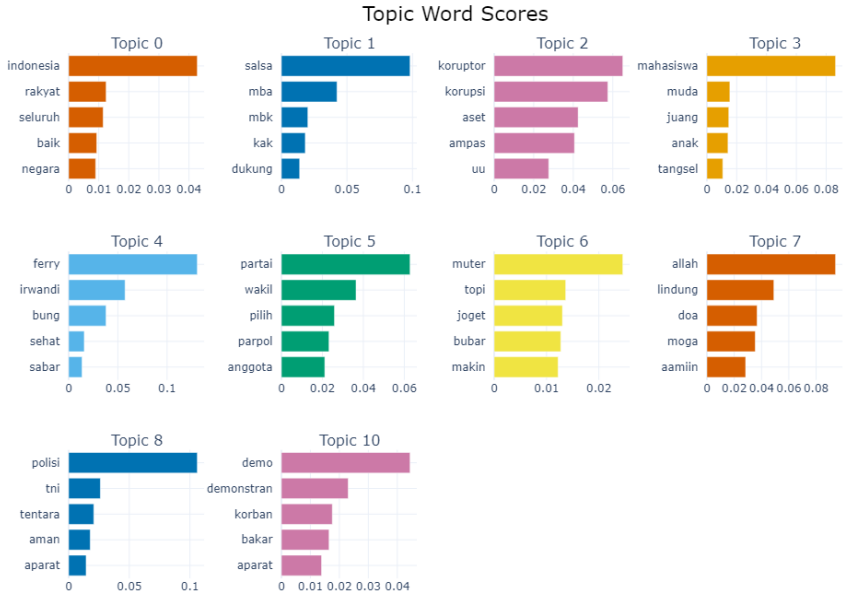
The Intertopic Distance Map visualization in Picture 5 below shows the distribution and semantic distance between topics resulting from modeling. Larger circles indicate a higher proportion of data, while the distance between circles illustrates the semantic differences between topics. This map shows that most topics are well

distributed, indicating a clear semantic separation between major themes such as "rakyat–pemerintah", "korupsi–aset negara", and "aksi–demonstrasi".

To observe the main words that form each topic, Picture 6 below shows the Topic Word Scores for the top ten topics. Topic 0 focuses on the words "indonesia", "rakyat", "seluruh", and "baik"; topic 2 on the words "koruptor", "korupsi", and "aset"; while topic 3 features words such as "mahasiswa", "muda", and "juang" which describe the sentiment of the younger generation's struggle. This combination of words shows that public opinion is divided into three major axes, namely.



Picture 5. Intertopic Distance Map



Picture 6. Topic Word Scores

Table 1 below presents the document probability values for each topic, which quantitatively reflect the model's level of confidence in grouping comments into appropriate semantic clusters. The results show that most comments have a topic probability value above 0.4, indicating that the BERTopic model has a high degree of confidence in determining the semantic classification of each comment. Overall, these classification results show solid performance with an average document-topic probability value of 0.488, indicating that the model successfully grouped comments with an adequate level of semantic consistency. Comments with high probabilities such as those in index 8, which reached a value of 1.0 generally have clear meanings and strong relevance to the main theme of the topic, while comments with low probabilities or those included in topic -1 (outliers) reflect linguistic variations, the use of sarcasm, non-standard abbreviations, or natural noise commonly found in social media data. Thus, these findings not only confirm that the resulting topic mapping is valid and semantically consistent, but also reflect the model's ability to distinguish between core discourse and complex and ambiguous linguistic variations.

Table 1. Document Probability Value For Topics

Index	Cleanned Comment	Topic	Topic Probability
0	cari tau gaes tanggap benix silvia tjan ngusut...	-1	0.381308
1	siapa bikin aneh masuk akal tuntut hati2 jgn a...	34	0.052100
2	bismillaah ayo kawal gais	-1	0.180472
3	kira tuntut salsa nyata bu kaann	1	0.541417
5	mau tanya dok tuntut tni barak jelas dok	4	0.398805
6	presiden kga ngerti kga bca poin tunttutan pid...	12	0.181144
7	tuntut fokus dpr si kocak	24	0.037414
8	lebih tuju usul salsa lebih masuk akal minggu ...	1	1.000000
9	koq bahas pajak naik banyak cuy	21	0.455406
10	jerome unfollowe anies curiga takut tahu kalo ...	-1	0.985073

The topic coherence value (C_v) presented in Table 2 above reveals the semantic quality of each topic cluster generated. It can be seen that Topic 2 ranks highest with a C_v score of 0.798818, followed by Topic 81 (0.788841) and Topic 29 (0.756789). The C_v scores approaching 0.8 for these three leading topics indicate that the sets of keywords in these topics have very close semantic relationships and high contextual consistency, making them easy to understand as a cohesive thematic whole. Meanwhile, the range of coherence values varying from medium to high across all topics with most still above 0.65 confirms that overall, the model successfully generates topics that are not only relevant but also have a logical and easily interpretable internal structure.

Table 2. Topic Coherence Score

Topic	Coherence Value (C_v)
2	0.798818
81	0.788841
29	0.756789
13	0.747067
23	0.730517
8	0.725303
0	0.678954
11	0.648498
71	0.647865
45	0.646957

Further analysis of the keywords in Table 3 below shows that topics with high coherence often contain words such as “koruptor”, “aset”, “ampas”, “uu”, and “sahkan”. Meanwhile, other topics such as Topic 8 feature a combination of words such as “polisi”, “tni”, “tentara”, and “aman”, which point to a discourse on security forces and the handling of demonstrations.

Table 3. Topic Keyword

Topic	KeyWord
Topic 2	0.065*"koruptor" + 0.057*"korupsi" + 0.042*"aset" + 0.041*"ampas" + 0.028*"uu" + 0.022*"sahkan" + 0.017*"ruu" + 0.015*"korup" + 0.015*"sah" + 0.013*"bahas"
Topic 81	0.788841 0.067*"ps" + 0.062*"kelompok" + 0.03*"jkw" + 0.03*"mafia" + 0.029*"sigit" + 0.029*"parte" + 0.026*"society" + 0.024*"civil" + 0.024*"jokowi" + 0.024*"sasar"
Topic 29	0.278*"pensiun" + 0.072*"hapus" + 0.062*"umur" + 0.053*"tunjang" + 0.046*"dana" + 0.04*"dpt" + 0.039*"kerja" + 0.037*"hidup" + 0.033*"uang" + 0.032*"gaji"
Topic 13	0.117*"aset" + 0.117*"ampas" + 0.088*"uu" + 0.07*"sah" + 0.051*"sahkan" + 0.049*"ruu" + 0.035*"asset" + 0.034*"segera" + 0.032*"uud" + 0.026*"utama"
Topic 23	0.142*"anarkis" + 0.045*"demo" + 0.027*"demonstran" + 0.027*"bebas" + 0.021*"warga" + 0.019*"provokator" + 0.018*"aparat" + 0.018*"jarah" + 0.017*"umum" + 0.016*"provokasi"
Topic 8	0.106*"polisi" + 0.026*"tni" + 0.021*"tentara" + 0.018*"aman" + 0.014*"aparat" + 0.014*"tangkap" + 0.012*"intel" + 0.011*"tugas" + 0.01*"laku" + 0.01*"aksi"
Topic 0	0.043*"indonesia" + 0.012*"rakyat" + 0.011*"seluruh" + 0.009*"baik" + 0.009*"negara" + 0.009*"lebih" + 0.009*"bangsa" + 0.008*"satu" + 0.008*"juang" + 0.008*"semua"
Topic 11	0.105*"gaji" + 0.042*"umr" + 0.033*"naik" + 0.029*"tunjang" + 0.029*"inflasi" + 0.027*"bulan" + 0.022*"kerja" + 0.021*"dpr" + 0.015*"jt" + 0.015*"pokok"
Topic 71	0.147*"koruptor" + 0.147*"mati" + 0.138*"hukum" + 0.07*"aset" + 0.065*"ampas" + 0.041*"uu" + 0.041*"ruu" + 0.03*"uud" + 0.029*"undang" + 0.026*"vonis"
Topic 45	0.077*"aset" + 0.06*"ampas" + 0.05*"ruu" + 0.034*"uu" + 0.032*"tdk" + 0.031*"tunggu" + 0.028*"sahkan" + 0.027*"sah" + 0.026*"bahas" + 0.021*"bubar"

Table 4 below presents several representative comments that serve as concrete examples for each topic cluster. For example, Topic 2 (corruption–assets) contains harsh comments such as “The law is a waste of corruptors' assets...”, while Topic 23 (anarchists–demonstrations) contains more aspirational statements such as “Demonstrations should be elegant expressions of aspiration, not anarchist acts that damage public facilities...”. This variation in language style shows that the BERTopic model successfully grouped comments based on their semantic similarity, even though the expressions used came from different backgrounds and writing styles. This proves the model's resilience in capturing the essence of meaning behind differences in diction and sentence structure.

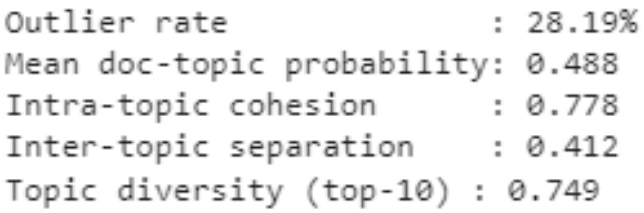
Table 4. Representative Comment

Topic	Representative Comment
Topic 2	UU ampas aset koruptor...
Topic 23	Demo benar aspirasi elegan jangan anarkis jgn rusak fasum...
Topic 8	Tuntut TNI padahal nabrak polisi...
Topic 0	Tugas rakyat Indonesia awal tuntut selamat rakyat Indonesia...

Overall, BERTopic successfully revealed the complex structure of public discourse, which reflected a number of key issues of public concern. Prominent issues include public dissatisfaction with the performance of legislative institutions, issues of corruption and state asset management, the role of students and the people in social action, and narratives of disappointment with the attitude of the authorities during the demonstrations. This

mapping provides a clear and structured picture of the sentiments and concerns of the public developing in the digital space, making it a valuable source of insight.

The results of the model evaluation in **Error! Reference source not found.** below provide concrete, quantitative evidence that BERTopic's performance can be confidently categorized as both stable and semantically representative. A closer look at the specific metrics reveals the robustness of the model: an Outlier Rate of 28.19%, while indicating room for minor improvement, fundamentally shows that the vast majority of comments were successfully and correctly classified into their relevant thematic topics. The Mean Document–Topic Probability, which stands at 0.488, suggests a moderate level of confidence from the model in its assignments, a common and acceptable characteristic when dealing with the inherent noise and ambiguity of natural language, particularly in informal online discourse. More impressively, the high Intra-topic Cohesion score of 0.778 demonstrates that the comments grouped within each individual topic share a very strong and coherent semantic meaning, confirming that the topics are internally consistent and meaningful. This internal clarity is effectively balanced by a sufficient Inter-topic Separation of 0.412, which confirms that the boundaries between different topics are distinct enough to prevent significant overlap and confusion, ensuring each topic captures a unique theme. Finally, the high Topic Diversity score of 0.749 proves that the keywords generated to label each topic are varied, comprehensive, and non-redundant, effectively capturing the essence of each cluster without repetition. When synthesized, this powerful combination of metrics conclusively demonstrates that BERTopic possesses a reliable and adaptable ability to map the complex and nuanced landscape of public discourse on social media in a contextual manner, successfully navigating the challenges posed by short, informal, and often unstructured language texts such as online comments.



Picture 7. Evaluation Model

CONCLUSIONS

This study successfully identified and mapped public opinion towards the Indonesian House of Representatives (DPR RI) at the moment of the 17+8 demands using the BERTopic method based on semantic embedding. From the analysis of 21,195 social media comments, more than 90 topics were obtained that represented major issues such as criticism of the DPR's performance, allegations of corruption and misuse of state assets, student involvement in the protests, and public views on security forces. The model evaluation results showed an intra-topic cohesion value of 0.778, inter-topic separation of 0.412, topic diversity of 0.749, and an outlier rate of 28.19%, indicating that BERTopic was able to produce a topic structure that was sufficiently coherent, diverse, and contextual. This conclusion confirms that the transformer embedding-based approach is effective for understanding public discourse in short, informal texts such as social media comments. Further research is recommended to develop approaches using other methods such as Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), or even Deep Clustering and Graph-based Topic Modeling approaches, in order to compare performance and broaden perspectives in topic analysis and public discourse mapping in Indonesia.

REFERENCES

Alonso-Fernandez, F., Belvisi, N. M. S., Hernandez-Diaz, K., Muhammad, N., & Bigun, J. (2021). Writer Identification Using Microblogging Texts for Social Media Forensics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *3*(3), 405–426. https://doi.org/10.1109/TBIOM.2021.3078073

Anantari, A. (2025). *Unjuk rasa dan kerusuhan Indonesia Agustus–September 2025*. Wikipedia. https://id.wikipedia.org/wiki/Unjuk_rasa_dan_kerusuhan_Indonesia_Agustus–September_2025

Anggiawan, F. V. (2023). *Analisis Pengaruh Media Sosial Pemerintah Terhadap Masyarakat di Kabupaten Sukoharjo*. Institut Pemerintahan Dalam Negeri.

Asani, E. O., Babalola, O. S., Akinola, A. E., Barnabas, A. S., Adams, O. V., & Odumesi, J. O. (2024). A Server-side Phishing detection API using Long-Short Term Memory (LSTM) and Multi-layer Perceptron (MLP). *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, 1–8. https://doi.org/10.1109/SEB4SDG60871.2024.10629792

Asyaky, M. S., & Mandala, R. (2021). Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP. *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA53211.2021.9640285

Ayuso, E., Dumfeh Brogya, M. S., Kumar Ahlawat, V., & Sain, M. (2024). From Manual to Machine: How AI is Redefining Web Scraping for Superior Efficiency: A Literature Review. *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*, 1–9. https://doi.org/10.1109/CCIS63231.2024.10931912

Benali, M. (2024). *Exploring Arabic Hotel Reviews: Sentiment Insights Through Deep Learning and BERT Transformers Models* (pp. 3–13). https://doi.org/10.1007/978-3-031-68650-4_1

Blair, S. J., Bi, Y., & Mulvenna, M. D. (2020). Aggregated topic models for increasing social media topic coherence. *Applied Intelligence*, *50*(1), 138–156. https://doi.org/10.1007/s10489-019-01438-z

Cici Suhaeni, B. S. S. N. S. A. K. L. Z. (2025). Pemodelan Topik pada Komentar YouTube Arra: Komparasi LDA dan K-Means Menggunakan Fitur Leksikal dan Semantik. *Jurnal Informatika: Jurnal Pengembangan IT*, *10*(3), 694–707. https://doi.org/10.30591/jpit.v10i3.8763

Çiplak, Z., & Yildirim, S. (2026). *The Use of Selenium Library in the Academic World* (pp. 411–424). https://doi.org/10.1007/978-981-96-8901-9_36

Devika, R., Vairavasundaram, S., Mahenthar, C. S. J., Varadarajan, V., & Kotecha, K. (2021). A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data. *IEEE Access*, *9*, 165252–165261. https://doi.org/10.1109/ACCESS.2021.3133651

Dinda Aurellia, & Katimin, K. (2025). Arah Baru Politik di Era Digital (Perspektif Generasi Milenial di Media Sosial). *SOSMANIORA: Jurnal Ilmu Sosial Dan Humaniora*, *4*(2), 231–236. https://doi.org/10.55123/sosmaniora.v4i2.5149

Edwin Hendrik, Subekti Subekti, & M. Syahrul Borman. (2024). Kebijakan Batasan Kekuasaan Eksekutif dan Legislatif dalam Amademen Undang-Undang Dasar 1945 pada Sistem Pemerintahan. *Majelis: Jurnal Hukum Indonesia*, *1*(3), 122–140. https://doi.org/10.62383/majelis.v1i3.345

EKİNCİ, E., & İLHAN OMURCA, S. (2020). NET-LDA: a novel topic modeling method based on semantic document similarity. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, *28*(4), 2244–2260. https://doi.org/10.3906/elk-1912-62

Feng, Y., Chen, Z., Zhang, Y., Huang, W., Zhang, X., & He, S. (2025). BERTopic_Teen: a multi-module optimization approach for short text topic modeling in adolescent health. *Frontiers in Public Health*, *13*. https://doi.org/10.3389/fpubh.2025.1608241

Fuchs, M., & Höpken, W. (2022). *Clustering* (pp. 129–149). https://doi.org/10.1007/978-3-030-88389-8_8

Hankar, M., Kasri, M., & Beni-Hssane, A. (2025). A comprehensive overview of topic modeling: Techniques, applications and challenges. *Neurocomputing*, *628*, 129638. https://doi.org/10.1016/j.neucom.2025.129638

Helan, A., & Sultani, Z. N. (2023). *Topic modeling methods for text data analysis: A review*. 040006. https://doi.org/10.1063/5.0118679

Hendra, H., & Halbadika Fahlevi, A. (2024). Implementation of Good Corporate Governance (GCG) Principles in PDAM Tirta Ogan, Ogan Ilir District. *Iapa Proceedings Conference*, 187. https://doi.org/10.30589/proceedings.2024.1052

Hernandez-Mendoza, M., Aguilera, A., Dongo, I., Cornejo-Lupa, J., & Cardinale, Y. (2022). Credibility Analysis on Twitter Considering Topic Detection. *Applied Sciences*, *12*(18), 9081. https://doi.org/10.3390/app12189081

Hoti, M. H., Rexhepi, A., Hoti, A. H., & Rexha, B. (2025). Challenges and Solutions in Clustering Low-Resource Language Social Media Text: An Evaluation Using Unsupervised Algorithms. *International Journal of Interactive Mobile Technologies (IJIM)*, *19*(20), 151–167. https://doi.org/10.3991/ijim.v19i20.56307

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Studies Journal*, *49*(1), 300–324. https://doi.org/10.1111/psj.12343

Jayadi, H., Basniwati, A., & Sofwan, S. (2022). Lembaga-lembaga Negara Sederajat Dalam Struktur Kelembagaan Negara Berdasarkan Undang-Undang Dasar 1945. *Journal Kompilasi Hukum*, *7*(2). https://doi.org/10.29303/jkh.v7i2.122

Jiang, J., Ying, F., & Dhuny, R. (2025). Unveiling Technological Evolution with a Patent-Based Dynamic Topic Modeling Framework: A Case Study of Advanced 6G Technologies. *Applied Sciences*, *15*(7), 3783. https://doi.org/10.3390/app15073783

Khder, M. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 145–168. https://doi.org/10.15849/IJASCA.211128.11

Kousis, A., & Tjortjis, C. (2023). Investigating the Key Aspects of a Smart City through Topic Modeling and Thematic Analysis. *Future Internet*, *16*(1), 3. https://doi.org/10.3390/fi16010003

Li, C., & Hu, X. (2025). Medical Artificial Intelligence in Scholarly and Public Perspective: BERTopic-Based Analysis of Topic-Sentiment Collaborative Mining. *Data Science and Informetrics*. https://doi.org/10.1016/j.dsim.2025.05.001

Li, T., Cui, L., Wu, Y., Pandey, R., Liu, H., Dong, J., Wang, W., Xu, Z., Song, X., Hao, Y., Cui, X., Du, J., Zhang, X., & Wang, Y. (2025). Unveiling and advancing grassland degradation research using a BERTopic modelling approach. *Journal of Integrative Agriculture*, *24*(3), 949–965. https://doi.org/10.1016/j.jia.2024.11.008

Ma, L., Chen, R., Ge, W., Rogers, P., Lyn-Cook, B., Hong, H., Tong, W., Wu, N., & Zou, W. (2025). AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Experimental Biology and Medicine*, *250*. https://doi.org/10.3389/ebm.2025.10389

Marcelina, J., Tosida, E. T., & Aryani, A. S. (2025). Implementation of Bidirectional Long Short Term Memory (BiLSTM) Algorithm with Embedded Emoji Sentiment Analysis of Covid 19 Anxiety Level and Socio Economic Community. *International Journal of Quantitative Research and Moadeling (IJQRM)*, *Vol. 5 No.* https://doi.org/https://doi.org/10.46336/ijqrm.v5i2.682

Mazjah, I., & Kurdi. (2025). Pemisahan Kekuasaan Dalam Sistem Pemerintahan: Studi Atas Implementasi Trias Politica Di Indonesia. *(JIHHP) Jurnal Ilmu Hukum, Humaniora Dan Politik*. https://doi.org/https://doi.org/10.38035/jihhp.v5i3

Mengko, V. A. (2024). Fungsi Dewan Perwakilan Rakyat Daerah Dalam Pengawasan Anggaran Pendapatan Belanja Daerah Di Kota Manado. *LEX PRIVATUM*, *13*(3).

Mittal, M., J, P. G., M S, G. P., Devadas, R. M., Ambreen, L., & Kumar, V. (2024). Dimensionality Reduction Using UMAP and TSNE Technique. *2024 Second International Conference on Advances in Information Technology (ICAIT)*, 1–5. https://doi.org/10.1109/ICAIT61638.2024.10690797

Nugroho, R. A., Warsito, B., & Hakim, A. R. (2023). Performance Comparison of Support Vector Machine and Random Forest Using Delta TF-IDF on Sentiment Classification of Tiket.com Reviews. *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 632–637. https://doi.org/10.1109/ISITIA59021.2023.10221142

Nursyahrina, Defit, S., & Sovia, R. (2024). Metode BERTopic dan LDA untuk Analisis Tren Penelitian Bidang Ilmu Komputer. *Komtekinfo*, *Vol. 11 No*. https://doi.org/10.35134/komtekinfo.v11i4.580

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences*, *13*(2), 797. https://doi.org/10.3390/app13020797

Oktavianto, H., Ayun, Q., Azizah Al Faruq, H., Handayani, L., Lusiana, D., & Al Hikam Tirta Bayu Aj, M. F. (2024). The Effect of Laplace Smoothing on Naive Bayes for Sentiment Analysis of the Impact of the AstraZeneca Vaccine. *2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C)*, 625–630. https://doi.org/10.1109/BTS-I2C63534.2024.10941727

Parlak, I. B., Şahin, M. Ş., Acarman, T., Adel, M., & Bourennane, S. (2025). Quantification of Overlapping and Network Complexity in News: Assessment of Top2Vec and Fuzzy Topic Models. *Applied Sciences*, *15*(17), 9627. https://doi.org/10.3390/app15179627

Polignano, M., de Gemmis, M., & Semeraro, G. (2020). *Contextualized BERT Sentence Embeddings for Author Profiling: The Cost of Performances* (pp. 135–149). https://doi.org/10.1007/978-3-030-58811-3_10

Rahimi, Z., & Homayounpour, M. M. (2023). The impact of preprocessing on word embedding quality: a comparative study. *Language Resources and Evaluation*, *57*(1), 257–291. https://doi.org/10.1007/s10579-022-09620-5

Rai, A., & Borah, S. (2021). *Study of Various Methods for Tokenization* (pp. 193–200). https://doi.org/10.1007/978-981-15-6198-6_18

Rizqi Nandadita Pamungkas, Permadi, D., & Florina, I. D. (2024). Strategi Humor Gibran Rakabuming dalam Komunikasi Politik di Media Sosial X (Twitter). *Jurnal Pemerintahan Dan Politik*, *9*(3), 175–182. https://doi.org/10.36982/jpg.v9i3.4057

Romero, J. D., Feijoo-Garcia, M. A., Nanda, G., Newell, B., & Magana, A. J. (2024). Evaluating the Performance of Topic Modeling Techniques with Human Validation to Support Qualitative Analysis. *Big Data and Cognitive Computing*, *8*(10), 132. https://doi.org/10.3390/bdcc8100132

Rosalinda, G., Santoso, R., & Kartikasari, P. (2023). Pemodelan Topik Ulasan Aplikasi Netflix Pada Google Play Store Menggunakan Latent Dirichlet Allocation. *Jurnal Gaussian*, *11*(4), 554–561. https://doi.org/10.14710/j.gauss.11.4.554-561

Sari, E. L., Widyastuti, L., Rato, D., Mazjah, R. M. I., Kushidayati, M. R. D. L., Pratama, K. J., & Benuf, K. (n.d.). *MAJALAH HUKUM NASIONAL*.

Scaringi, R., Stea, G., Fanelli, N., Vessio, G., & Castellano, G. (2025). Multimodal Artwork Topic Modeling via Fine-Tuned Clip and Knowledge-Driven Prompts. *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. https://doi.org/10.1109/MLSP62443.2025.11204345

Sejati, P. T., Alzami, F., Marjuni, A., Indrayani, H., & Puspitarini, I. D. (2024). Aspect-Based Sentiment Analysis for Enhanced Understanding of "Kemenkeu" Tweets. *Journal of Applied Informatics and Computing*, *8*(2), 487–498. https://doi.org/10.30871/jaic.v8i2.8558

Setiawan, A., Sari, S. M., & Rahmah, N. (2025). Lembaga Negara Pemegang Kekuasaan Legislatif. *Jurnal Hukum Dan Kewarganegaraan*, *Vol 15 No*. https://doi.org/doi.org/10.3783/causa.v2i9.2461

Silva, C. C., Galster, M., & Gilson, F. (2021). Topic modeling in software engineering research. *Empirical Software Engineering*, *26*(6), 120. https://doi.org/10.1007/s10664-021-10026-0

Stewart, G., & Al-Khassaweneh, M. (2022). An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences*, *12*(5), 2405. https://doi.org/10.3390/app12052405

Sugiman, S. (2020). Fungsi Legislasi Dpr Pasca Amandemen UUD NKRI 1945. *Jurnal Ilmiah Hukum Dirgantara*, *10*(2). https://doi.org/10.35968/jh.v10i2.468

Susanto, A. D., Andrian Pradita, S., Stryadhi, C., Setiawan, K. E., & Fikri Hasani, M. (2023). Text Vectorization Techniques for Trending Topic Clustering on Twitter: A Comparative Evaluation of TF-IDF, Doc2Vec, and Sentence-BERT. *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)*, 1–7. https://doi.org/10.1109/ICORIS60118.2023.10352228

Teotia, H., Shishodia, G., Tyagi, E., Prakash, A., & Avasthi, S. (2023). Instagram Analysis and Activity Automation: Using Python and Selenium Automation Tools. *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, 522–526. https://doi.org/10.1109/CICTN57981.2023.10140356

Uzun, E. (2020). A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages. *IEEE Access*, *8*, 61726–61740. https://doi.org/10.1109/ACCESS.2020.2984503

Vayadande, K., Kulkarni, A. M., Yadav, G. B., Kumar, R., & Sawant, A. R. (2024). A Comprehensive Analysis of Various Tokenization Techniques and Sequence-to-Sequence Model in Natural Language Processing. In *How Machine Learning is Innovating Today's World* (pp. 1–11). Wiley. https://doi.org/10.1002/9781394214167.ch1

Venugopalan, M., & Gupta, D. (2022). An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. *Knowledge-Based Systems*, *246*, 108668. https://doi.org/10.1016/j.knosys.2022.108668

Wahyuni, W., Lestari, T. P., Apriliana, M., & Gumelta, R. (2025). Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments. *Journal of Applied Informatics and Computing*, *9*(4), 1964–1971. https://doi.org/10.30871/jaic.v9i4.9754

Wani, A. A. (2025). Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions. *PeerJ Computer Science*, *11*, e3025. https://doi.org/10.7717/peerj-cs.3025

Widiastuti, I., & Yong, H.-S. (2025). TR-GPT-CF: A Topic Refinement Method Using GPT and Coherence Filtering. *Applied Sciences*, *15*(4), 1962. https://doi.org/10.3390/app15041962

Yang, X., Yang, K., Cui, T., Chen, M., & He, L. (2022). A Study of Text Vectorization Method Combining Topic Model and Transfer Learning. *Processes*, *10*(2), 350. https://doi.org/10.3390/pr10020350

Zoya, Latif, S., Shafait, F., & Latif, R. (2021). Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. *IEEE Access*, *9*, 127531–127547. https://doi.org/10.1109/ACCESS.2021.3112620