JURNAL INFORMATIKA
PROXIES

Vol.9, No.2, January, 2026

Topic-Based Clustering and Social Network Analysis of News Headlines Using LDA and DBSCAN

Norli[1], Robertus Setiawan Aji Nugroho[2]

[1] Teknik Informatika, Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata, 22k10054@student.unika.ac.id
[2] Teknik Informatika, Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata, nugroho@unika.ac.id

Corresponding Author Email: nugroho@unika.ac.id

https://doi.org/10.18280/proxies.v9i2.15200

Received: 2026-04-06
Revised: 2026-07-20
Accepted: 2026-07-20
Available online: 2026-07-21

Keywords: *LDA, DBSCAN, SNA, preprocessing, online news clustering*

ABSTRACT

The rapid growth of online news has made it difficult for readers to find information that suits their interests and understand current issues in society. Therefore, automatic clustering is needed to facilitate efficient information analysis. The purpose of this study is to apply Latent Dirichlet Allocation (LDA) to extract topic representations from news headlines and evaluate its effectiveness for clustering using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method. The data used are only news headlines from three online media outlets: detik.com, kompas.com, and CNNIndonesia. The research stages include data preprocessing, text weighting, vector representation, LDA application for topic formation, and DBSCAN for density-based cluster formation. Model parameters are determined through experiments to obtain the optimal number of topics and the appropriate epsilon and MinPts values. Evaluation is carried out using a coherence score for topic quality and a silhouette score for cluster quality. In addition, Social Network Analysis (SNA) is used to analyze the relationship structure between news headlines within each cluster. The results of the study show that LDA is effective in producing topic representations, DBSCAN is able to cluster data, and SNA provides an understanding of the relationship structure within the cluster.

1. INTRODUCTION

The volume of online news continues to grow, making it difficult for readers to find information that meets their needs. Although categories such as sports, health, and economics are available, grouping is still done manually and is ineffective. Therefore, an automated approach is needed to group and analyze relationships between news items.

To meet the needs of readers, news is grouped based on topics with an automatic approach. Clustering is a method that is usually used to group news into one group that has similarities. For example, suppose there are several news titles such as "*Kenaikan kasus COVID-19 Asia bikin waswas, penjualan masker di Vietnam meroket"* and *"Benarkah menteri kesehatan wajibkan penumpang pesawat vaksin TBC? Ini faktanya"* can be automatically grouped into health topics. Therefore, it is necessary to use an approach such as Latent Dirichlet Allocation (LDA) [1] to find the dominant topic in a set of texts, while to group news titles based on density, without having to do the number of clusters at the beginning, you can use DBSCAN[2].

Previous studies have only used a combination of LDA and DBSCAN[1], but this approach has not explored the relationships between data in more depth. Therefore, this study combines LDA, DBSCAN, and SNA to produce more representative clustering and analyze the structure of relationships between news titles.

This automated approach helps filter out irrelevant information and program group news stories based on topical similarities. This grouping allows the structure of relationships between news stories to understand topic association patterns within the data. Therefore, this method not only improves data management efficiency but also supports data-driven analysis of news topic structure.

2. LITERATURE STUDY

Topic modeling such as LDA has been used in several previous studies. To analyze the narratives of COVID-19 patients in Italy using BERTopic and LDA[2]. The results showed that the cluster accuracy level reached 97.26% and the overall accuracy was 91.97%. Although the accuracy is very high, this study also has several shortcomings, such as using data that only comes from one particular period, so it greatly affects the clustering results and does not take into account changes in patient conditions over time.

To determine the topic in Korean drama reviews, LDA can be applied[3]. There are only 10 maximum topics that were successfully identified from the 20 topics generated. This study was successful in modeling topics using LDA, but there were also difficulties experienced in determining the most appropriate number of topics. LDA is indeed effective for use in identifying even hidden topics, but it remains a big challenge to choose the right number of topics in its application.

On the other hand, Gibbs sampling and LDA can also be used to group scientific journals based on titles[4]. They revealed that LDA is very suitable for handling sparse data. In their research, the challenge in training the model was because only a few words appeared repeatedly in the document. So, when the data used only contains a few words that appear repeatedly, this technique is less effective to apply, even though LDA has superior capabilities.

The DBSCAN algorithm is used for clustering twitter texts related to the regional head elections in Pekanbaru[5]. Only 31 clusters were successfully identified with a Silhouette Index (SI) value of 0.413, from 2,184 tweets analyzed. In their research, it was proven that DBSCAN is quite effective in grouping short texts. Although it is effective, it can still be developed through appropriate parameter adjustments and a thorough data cleaning process.

Hybrid framework proposed by[1] for news clustering using DBSCAN-Martingale combination. The performance shown is superior, when compared with 20 other methods, the highest Normalized Mutual Information (NMI) score. However, in the research conducted, the interpretation of the resulting topics was not studied thoroughly, which caused the semantic meaning to be less than optimal.

Meanwhile, in the study[6], text clustering from Instagram and X (Twitter) data regarding public views on the 2024 presidential candidates was carried out. They used the DBSCAN and LDA algorithms, and used two text representation approaches: TF-IDF and Sentence¬-BERT. The results showed that TF-IDF could produce more regular clusters, while Sentence¬-BERT could cause noise to appear and the resulting clusters were uneven. This situation shows that the text representation method greatly influences the effectiveness of clustering as well as the characteristics of the data and the suitability of the method.

Not only LDA and DBSCAN perform text clustering, but there are several other approaches that show effectiveness in identifying irregular data topics. To perform clustering of 20 newsgroups consisting of 20,000 documents, a combination of TF-IDF and K-Means methods can be used[7]. From their research, it shows that the formation of clusters has interrelated content, where each topic can be described through appropriate keywords. This approach has been proven to provide informative cluster results, but there are limitations because evaluation metrics are not included to assess the quality of clustering results, as a result, the objective truth of their findings can still be evaluated further.

Embedding-based representations of large language models such as BERT and GPT-3.5 can be used for text clustering[8]. In order to assess the quality of the resulting clustering, their research applied evaluation metrics including purity and silhouette score. GPT-3.5 showed superior clustering performance results in terms of accuracy. However, the challenge is that it requires large computing resources and high processing time, so this application is less effective for large scales without sufficient computing power support.

In addition, algorithms such as k-means and agglomerative clustering based on Word2Vec are used for clustering fake news [9]. In their research, agglomerative clustering showed less superior quality results when compared to K-means, evaluated from metrics such as the Adjusted Rand Index (ARI). However, when used to capture complex text contexts, the semantic representation of Word2Vec is considered to still have limitations. It is emphasized in this finding that although K-means clustering is quite effective, the quality of the text representation used greatly influences the success of the clustering.

On the other hand, there is a proposal to use a parallel approach for clustering and modeling news topics by applying big data technology through MapReduce and Apache Spark[10]. The purpose of this approach is for efficiency and scalability in processing large datasets collected from Kazakhstan news sites. The results of their research show high efficiency, although of course there are challenges they face such as limited computing resources, implementation complexity, and processing that requires a long duration. Therefore, the availability of hardware and sufficient technical expertise greatly influences the success of the method. In large-scale data scenarios, this approach provides a flexible alternative to classical clustering methods.

Topic modeling and Social Network Analysis (SNA) were combined in previous research to analyze health discourse on Twitter[11]. Several topic modeling methods were compared to identify key themes in text data. SNA was used to identify actors with significant influence on information dissemination. The analysis included identifying topics and the relationships between entities in text-based discourse.

Social Network Analysis (SNA) was previously used by researchers to analyze public discourse on Twitter regarding the COVID-19 pandemic[12]. The results of this study showed that conversation networks are fragmented and decentralized, which affects the spread of information. The SNA approach is used to visualize the structure of relationships between entities in text data. Network analysis allows the identification of relationship patterns that are not visible through content analysis alone. Therefore, SNA is a relevant approach for analyzing information relationships in text-based data.

3. RESEARCH METHODOLOGY

Data Collection

The data used in this study were obtained from the web scraping process from three online news media, namely CNNIndonesia, dertik.com and kompas.com. The focus of the data collection process is on the news title section, because it is able to show the main content of each article. This dataset contains the date of publication of the news, the news title and a link to the original news source. The data collection process is carried out periodically in order to obtain diverse topics and explain real issues that are currently of public concern. The number of datasets to be used is 36,504.

Data Preprocessing

The preprocessing stage includes several steps, namely data cleaning, case folding, tokenization, stopword removal, and stemming. Data cleaning is performed to remove duplicate data and empty data to maintain data quality. Case folding converts all letters to lowercase for consistency. Tokenization is used to separate text into words. Stopword removal removes common words that have no significant meaning, while stemming simplifies words to their basic form. These steps are carried out to improve the quality of text representation before it is used for further analysis.

Feature Extraction

Bag of Words (BoW) is used to transform news headline data into a numerical representation. This representation is used as input for the Latent Dirichlet Allocation (LDA) algorithm to generate topic distributions based on word frequency. This topic distribution is then used as the basis for the DBSCAN clustering process.

Latent Dirichlet Allocation (LDA)

After the feature extraction process, the news headline data was modeled using Latent Dirichlet Allocation (LDA) to extract latent topics. In LDA, each document is represented as a topic distribution, while each topic consists of a word distribution. The number of topics was determined experimentally, considering the coherence value to obtain optimal results.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

After obtaining the topic distribution from LDA, the DBSCAN algorithm is used to cluster the news headline data based on data density. Topic representations are used as input vectors, while the epsilon (ε) and minimum neighbor parameters are determined experimentally. This method generates clusters based on topic similarity and identifying data containing noise without the need to determine the number of clusters.

Social Network Analysis (SNA)

The next analysis used a Social Network Analysis (SNA) approach to analyze the relationship structure between news headlines within each cluster. In this approach, news headlines are represented as nodes, while edges indicate relationships between headlines within the same cluster. Network analysis was conducted using centrality metrics, namely degree, betweenness, and closeness centrality, as well as density to calculate the overall level of network connectivity. Through this analysis, we can identify network structure patterns and nodes that play a central role in each cluster.

Evaluation

Evaluation was conducted to assess the quality of the modeling and clustering results. LDA was evaluated using coherence scores, while DBSCAN used silhouette scores. Meanwhile, SNA was used to analyze the relationship structure between news headlines within each cluster and identify central actors within each cluster.

4. RESULT

The number of topics in LDA was determined using a hill climbing approach by testing 6–10 topics. The coherence value increased to a peak of 0.4835 at 8 topics, then decreased. Therefore, the optimal number of topics used was 8.

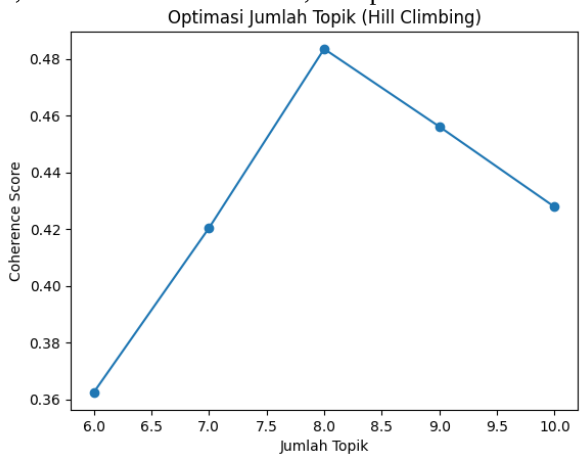

Figure 1 Optimizing the Number of Topic(Hill Climbing)

The ten most heavily weighted words in each topic illustrate the topic's main characteristics. Topic 5 is dominated by terms related to corruption and the Corruption Eradication Commission (KPK), indicating that law enforcement issues are a prominent theme in the data and reflecting the consistency of the LDA modeling results.

Table 1. Top Words per Topic LDA Result

Topic	**Top Words**
Topic 0	prabowo, sekolah, video, buka, perintah, menteri, rakyat, kepala, program, resmi
Topic 1	bal, nikmat, tarif, warga, seru, sehat, hut, indah, pulau, bandung
Topic 2	israel, kompas, com, video, serang, iran, trump, gaza, nadiem, cuaca
Topic 3	tewas, korban, temu, anak, video, bunuh, orang, tinggal, duga, mbg
Topic 4	jakarta, tangkap, polisi, mobil, warga, rumah, video, juta, motor, laku
Topic 5	dpr, demo, sangka, kpk, eks, korupsi, video, duga, kait, polisi
Topic 6	indonesia, harga, dunia, jadwal, daftar, timnas, ingat, cek, piala, main
Topic 7	baru, berita, kini, informasi, ekonomi, hari, didik, uang, teknologi, bank

DBSCAN parameter determination through a tuning process with eps = 0.20 and min_samples = 16. The clustering results show that most of the data is classified as noise, while the formed clusters have a relatively small and varied number of members. This condition indicates that the data has high variation, resulting in a less than optimal cluster pattern.

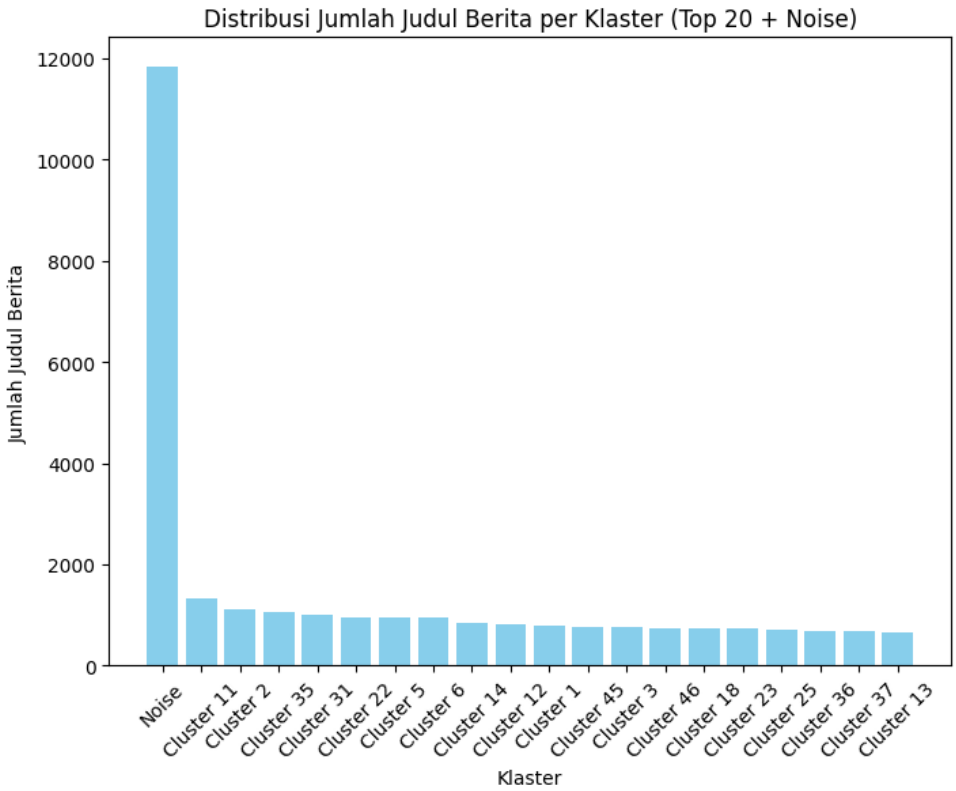

Figure 2 Distribution of News Titles per DBSCAN Cluster

Social network analysis was performed on a sample of 10,000 news headlines to maintain computational efficiency. The resulting network consisted of 6,721 nodes and 6,648 edges, with each node representing a news headline and edges indicating connections within the same cluster. The network visualization showed a structure fragmented into several separate components with no connections between clusters. This demonstrates that DBSCAN is able to clearly separate the data into distinct groups.

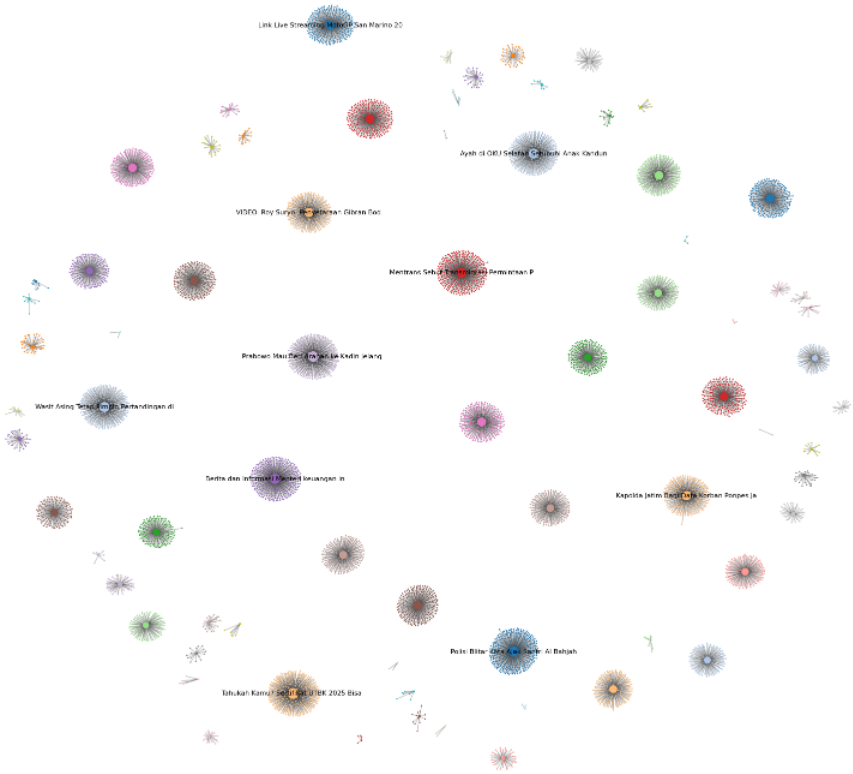

Figure 3 Network Graph of News Titles Based on SNA

The results of the network metric analysis show that the network has a low density (0.0003), indicating that the connectivity between nodes is relatively sparse. The low average values of degree centrality and betweenness centrality indicate that most nodes have only a few direct connections and do not act as primary links in the network. In addition, the low value of closeness centrality (0.0148) indicates that the distance between nodes is relatively large. Overall, these results indicate that the network is fragmented and dominated by connectivity within each cluster, with several nodes in each cluster acting as central actors who are the center of connections.

Table 2. Global Network Metrics Resulting from SNA Analysis

Network Metric	Value
Density	0.0003
Average Degree Centrality	0.0003
Average Betweenness Centrality	(4×10^{-6})
Average Closeness Centrality	0.0148

Degree centrality analysis indicates that some nodes have a higher degree of connectivity than others. These nodes act as central actors because they have more connections within the network. The presence of these central nodes indicates that the news headlines have a more dominant connection and serve as the center of connections within each cluster.

Table 3. Top 10 Nodes Based on Degree Centrality in the Network

TOP 10 AKTOR BERDASARKAN DEGREE CENTRALITY				
Node	Degree Centrality	Betweenness Centrality	Closeness Centrality	Title
2308	0.0466	0.002163	0.0466	Tahukah Kamu? Sertifikat UTBK 2025 Bisa Dipakai untuk Daftar Ujian Mandiri

2635	0.0443	0.001960	0.0443	Polisi Blitar Kota Ajak Santri Al Bahjah Tertib Lalu Lintas
18190	0.0443	0.001960	0.0443	Prabowo Mau Beri Arahan ke Kadin Jelang Retret di Magelang
17909	0.0421	0.001768	0.0421	Mentrans Sebut Transmigrasi Permintaan Pemda
33097	0.0418	0.001743	0.0418	Berita dan Informasi Menteri keuangan indonesia Terkini dan Terbaru Hari ini
17531	0.0406	0.001645	0.0406	Wasit Asing Tetap Pimpin Pertandingan di Super League 2025/2026
993	0.0402	0.001609	0.0402	Ayah di OKU Selatan Setubuhi Anak Kandung hingga Hamil 14 Minggu
32734	0.0333	0.001106	0.0333	Kapolda Jatim Bagi Data Korban Ponpes Jadi 3 Klaster
35803	0.0326	0.001057	0.0326	VIDEO: Roy Suryo: Penyetaraan Gibran Bodong
27148	0.0326	0.001057	0.0326	Link Live Streaming MotoGP San Marino 2025

Table 4.5 Top 10 Nodes Based on Degree Centrality in the Network

5. DISCUSSION

The results indicate that the topic modeling and clustering approach is capable of identifying thematic patterns in the news headline dataset. The LDA model yielded an optimal number of topics of 8 with a coherence value of 0.4835, indicating a fairly representative topic quality. Furthermore, the DBSCAN clustering stage produced several clusters with uneven distribution and a significant amount of noise. The silhouette score of 0.1566 indicates that clusters have been formed, although the separation is not yet fully optimal.

Analysis using SNA indicates that the network structure is fragmented, and the network analysis is limited to each cluster. Some nodes have a higher degree of connectivity and act as central actors in the network. These nodes become connection centers because they have more connections with other news titles discussing similar topics. On the other hand, low density values indicate that the connections between nodes are irregular and tend to be focused on specific nodes within the cluster.

Several limitations of this study include the use of data consisting only of news headlines, resulting in relatively limited information. Furthermore, the DBSCAN clustering results are heavily influenced by parameter selection, and the network model used in the SNA analysis is still relatively simple. Therefore, future research is recommended to use more comprehensive data and a more complex approach to obtain more optimal and representative results.

6. CONCLUSIONS

This study indicates that the combination of LDA, DBSCAN, and SNA is capable of identifying thematic patterns and relationship structures in news headline data. The eight LDA topics indicate a fairly representative distribution of topics, while DBSCAN formed clusters, albeit with a suboptimal level of separation. The SNA analysis shows that the network is fragmented and dominated by interconnectedness within clusters, with several nodes playing central roles. These results suggest that topic-based and network-based approaches can provide a deeper understanding of news connectivity patterns.

There are several limitations in this study, namely the use of data consisting only of news headlines, so the information obtained is relatively limited. On the other hand, although the parameters in LDA and DBSCAN have been determined through the optimization process, the clustering results still show suboptimal cluster separation. Therefore, future research is recommended to use complete news headline data with its contents and explore more complex clustering methods and network approaches to improve the quality of the analysis results.

REFERENCES

[1] A Hybrid Framework for News Clustering Based on the DBSCAN-Martingale and LDA. In: *Lecture Notes in Computer Science*. Cham: Springer International Publishing, pp. 170–184.

[2] Scarpino I, Zucco C, Vallelunga R, et al. Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration. *BioTech* 2022; 11: 41.

[3] Jannah AR, Kristi R, Habibi M. Metode Latent Dirichlet Allocation Untuk Menentukan Topik Pada Review Drama Korea.

[4] Santoso YO, Nugroho RSA. Pengelompokan Jurnal Ilmiah Berdasarkan Judul Menggunakan LDA. *proxies* 2021; 3: 32–42.

[5] Mustakim, Nurul Gayatri Indah R, Novita R, et al. DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru. *J Phys: Conf Ser* 2019; 1363: 012001.

[6] Hanifa A, Debora C, Hasani MF, et al. Analyzing Views on Presidential Candidates for Election 2024 Based on the Instagram and X Platforms with Text Clustering. *Procedia Computer Science* 2024; 245: 730–739.

[7] Sampath M. Text Clustering for Topic Identification: A TF- IDF and K-Means Approach Applied to the 20 Newsgroups Dataset.

[8] Petukhova A, Matos-Carvalho JP, Fachada N. Text Clustering with Large Language Model Embeddings. *International Journal of Cognitive Computing in Engineering* 2025; 6: 100–108.

[9] Statistics Program, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia, Muhammad Tianda I, Ubadah MN, et al. Clustering Fake News with K-Means and Agglomerative Clustering Based on Word2Vec. *IJMCR* 2024; 12: 3999–4007.

[10] Shomanov AS, Mansurova ME. Parallel news clustering and topic modeling approaches. *J Phys: Conf Ser* 2021; 1727: 012018.

[11] Ramamoorthy T, Kulothungan V, Mappillairaju B. Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India. *Front Artif Intell* 2024; 7: 1329185.

[12] Pascual-Ferrá P, Alperstein N, Barnett DJ. Social Network Analysis of COVID-19 Public Discourse on Twitter: Implications for Risk Communication. *Disaster med public health prep* 2022; 16: 561–569.