Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction

Dewi Fatmarani Surianto[1*], Dewi Fatmawati Surianto[2]
[1]Universitas Negeri Makassar, Indonesia, [2]Universitas Bakrie, Indonesia
[1]dewifatmaranis@unm.ac.id, [2]dewi.fatmawati@bakrie.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 17-12-2024
Accepted: 23-12-2024
Published: 30-12-2024

Keywords:
Clustering; feature extraction; k-means; LDA; TF-IDF

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Clustering is a fundamental technique in data analysis, particularly in unsupervised learning, to group data with similar characteristics. However, the effectiveness of the K-Means algorithm in text clustering heavily depends on proper feature extraction. This study proposes an enhanced feature extraction approach by integrating Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) to improve clustering performance on journal article datasets. The dataset consists of 427 journal article abstracts collected from Google Scholar. The preprocessing steps include tokenization, stopword removal, and TF-IDF vectorization, followed by topic extraction using LDA, which serves as input features for the K-Means clustering algorithm. The optimal number of clusters is determined using the Silhouette Score, with the best result obtained at k=9, achieving a score of 0.6806. The practical implications of this study include improved accuracy in academic document clustering, with applications in journal recommendation systems, digital library indexing, and research trend analysis. The results demonstrate that the combination of TF-IDF and LDA produces more informative text representations, significantly enhancing clustering quality. This study contributes to text mining and data science by proposing a systematic preprocessing framework for document clustering. Future research could explore its application to full-text articles, hierarchical clustering, or deep learning-based models to further improve clustering performance.

INTRODUCTION

Clustering refers to a basic technique in data analysis, mainly in the context of unsupervised learning, where the goal of clustering is to group similar data without labels or classes. Among the various clustering algorithms, K-Means is one of the most widely used methods due to its simplicity and efficiency in grouping data into different clusters based on similarity measures. However, the effectiveness of k-means clustering is highly dependent on the choice of features used for clustering. This has led to extensive research into feature extraction methods that can improve the performance of clustering algorithms (Goh et al., 2022). When applied to textual data, such as journal article datasets, clustering faces additional challenges due to the high dimensionality and sparsity of text representation.

The challenges associated with clustering journal article datasets are particularly pronounced due to the high dimensionality and often sparse nature of textual data. Traditional clustering methods are often unable to deal with these characteristics, resulting in suboptimal clustering results. Therefore, to overcome these problems, researchers have explored various feature extraction techniques, including the use of Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA). TF-IDF is a statistical measure that reflects the importance of a word in a document relative to the document set, while LDA focuses on a generative probabilistic model that identifies latent topics in a text corpus (Chen et al., 2019; B. Liu et al., 2021; Pokharel et al., 2021).

Recent advancements in clustering methodologies have the potential to significantly enhance the outcomes of data analysis tasks. However, a notable gap persists in the literature concerning the systematic integration of Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA) as a preprocessing step for K-Means clustering. While prior studies have investigated the individual applications of these techniques, there remains a lack of concentrated research on their combined synergistic effects to optimize clustering results. For instance, Nugroho et al. explored the integration of TF-IDF with LDA for aspect extraction in e-commerce, demonstrating that the combination can yield improved hyperparameter tuning and feature selection, which is crucial for effective clustering (Nugroho et al., 2022). This suggests that using TF-IDF to preprocess textual data before applying LDA could enhance the quality of the clusters formed by K-Means.

Moreover, the work of Rashid et al. highlights the importance of addressing redundancy in data, which can significantly affect the performance of clustering algorithms. Their study compared multiple kernel fuzzy topic modeling techniques against traditional LDA, emphasizing the necessity of preprocessing steps like TF-IDF to mitigate redundancy issues in biomedical documents (Rashid et al., 2022). This indicates that the preprocessing of textual data

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

through TF-IDF could not only improve the performance of LDA but also enhance the overall clustering results when subsequently applying K-Means. The potential for TF-IDF to serve as a preprocessing step is further supported by the findings of Muhajir et al., who demonstrated that improved term weighting techniques, including TF-IDF, can lead to more precise topic modeling outcomes (Muhajir et al., 2024).

In addition, the comparative analysis conducted by Dewald et al. revealed that while LDA and TF-IDF performed similarly in certain contexts, their combined application could yield superior results in specific datasets, particularly in medical texts (Dewald et al., 2023). This reinforces the notion that a systematic approach combining TF-IDF and LDA could optimize clustering tasks by leveraging the strengths of both methodologies. Furthermore, the study by Al-Obaydy et al. on document classification using TF-IDF and K-Means clustering supports the idea that integrating these techniques can enhance the classification and clustering of research papers, thereby addressing common challenges faced in information retrieval (Al-Obaydy et al., 2022).

The exploration of TF-IDF as a preprocessing step for LDA in clustering tasks is not only timely but also crucial for advancing the field. The findings from Istiqomah et al. further corroborate this, as they employed K-Means clustering with TF-IDF weighting to analyze user application reviews, demonstrating the effectiveness of this approach in deriving meaningful insights from textual data (Istiqomah et al., 2023). Similarly, the research by Mutawalli et al. on automating word and topic extraction using TF-IDF and LDA underscores the practical applications of combining these methodologies for enhanced clustering performance (Mutawalli et al., 2024).

Integration of the two techniques has the potential to improve clustering performance by providing a more informative representation of the data. Integrating LDA as a feature extraction method before applying K-Means clustering is expected to improve clustering performance by providing a more informative representation of the data. This approach leverages the strengths of TF-IDF and LDA, where TF-IDF is used to process the data first, and then utilizes LDA to extract meaningful topics that serve as features for k-means clustering. The combination of these methods is expected to produce better clustering results, as it allows the identification of underlying structures in the data that may not be apparent when using only raw text features (Karthik et al., 2019; Simoes et al., 2022).

Furthermore, the integration of clustering techniques with advanced feature extraction methods, particularly in the context of journal article datasets, has become an interesting research topic. Various studies have examined the application of each of these techniques separately, but few have explored their synergistic potential in improving clustering performance. In particular, there is still an important gap in the literature regarding the systematic combination of Term Frequency-Inverse Document Frequency (TF-IDF) with Latent Dirichlet Allocation (LDA) as a pre-processing step for k-means clustering.

LITERATURE REVIEW

Recent studies have underscored the efficacy of Latent Dirichlet Allocation (LDA) in topic modeling, showcasing its ability to extract latent topics from diverse textual datasets. For instance, Zhao's work on analyzing Reddit discussions illustrates how LDA can effectively identify and categorize topics within user-generated content, thereby enhancing our understanding of community discussions (Zhao, 2022). Similarly, the research conducted by Ghosh demonstrates the successful application of LDA in classifying scientific articles, where LDA was used in conjunction with TF-IDF to improve the representation of abstracts (Ghosh, 2021). These findings highlight the utility of LDA in uncovering underlying themes within documents, yet they often overlook the potential benefits of preprocessing textual data using TF-IDF prior to LDA application, which could enhance the quality of the extracted features. These studies underline the usefulness of LDA in identifying underlying themes in documents, but they do not discuss the pre-processing of textual data through TF-IDF before applying LDA, which could potentially improve the quality of the extracted features.

Moreover, although LDA has been widely recognized for its capabilities in topic modeling, its integration with clustering algorithms such as k-means remains unexplored. Recent advances in clustering methodologies, such as by George and Sumathy, who combined BERT with LDA for better topic modeling, show the growing interest in hybrid approaches (George & Sumathy, 2023). However, they did not specifically focus on the pre-processing step of using TF-IDF before applying LDA, which can be an important refinement in the clustering process.

Previous research has also revealed that while LDA is effective in extracting topics, it can also be computationally intensive, especially when applied to high-dimensional data sets (Aggarwal, 2024). Aggarwal's study on online reviews emphasized the need for efficient topic modeling techniques, showing that integrating TF-IDF can alleviate some of the computational burden by reducing dimensionality prior to the application of LDA (Aggarwal, 2024). This gap highlights the need for a more comprehensive exploration of how TF-IDF can serve as a pre-processing step to optimize LDA performance in clustering tasks.

The application of LDA in various domains, such as sentiment analysis and document retrieval, has been widely documented (Im et al., 2019). However, these studies often focus on LDA results without considering the impact of pre-processing techniques such as TF-IDF. For example, Liu et al. discussed the use of LDA for dynamic topic detection but did not investigate the potential benefits of integrating TF-IDF as a preprocessing step (H. Liu et al., 2020). This oversight provides an opportunity for further research to explore how such integration can improve clustering results.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, Number 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

Moreover, while some investigations have examined the integration of LDA with other machine learning techniques, such as deep learning models, the specific combination of TF-IDF and LDA for K-Means clustering remains largely unexplored. For example, the work of Istiqomah et al. emphasizes the significance of feature extraction methods in document retrieval systems, yet it does not delve into the synergistic integration of TF-IDF with LDA for clustering applications (Istiqomah et al., 2023). This gap in the literature suggests a critical area for further exploration, as the combination of these methodologies could yield improved clustering results by optimizing the preprocessing of textual data.

Additionally, recent advancements in the field have demonstrated the importance of addressing redundancy and noise in textual data, which can significantly impact the performance of clustering algorithms. Rashid et al. conducted a comparative analysis of multiple kernel fuzzy topic modeling techniques against traditional LDA, highlighting the necessity of preprocessing steps like TF-IDF to mitigate redundancy issues in biomedical documents (Rashid et al., 2022). This indicates that employing TF-IDF as a preprocessing step could not only enhance the performance of LDA but also improve the overall clustering outcomes when subsequently applying K-Means.

Furthermore, the study by Nugroho et al. on aspect extraction in e-commerce illustrates the effectiveness of combining TF-IDF with LDA, revealing that this integration can lead to better hyperparameter tuning and feature selection, which are essential for effective clustering (Nugroho et al., 2022). This reinforces the notion that a systematic approach combining TF-IDF and LDA could optimize clustering tasks by leveraging the strengths of both methodologies. The exploration of this combination is timely, as it aligns with the growing interest in enhancing clustering methodologies to derive meaningful insights from large volumes of textual data.

In summary, the existing literature provides a foundation for understanding the individual contributions of TF-IDF and LDA in the context of clustering. However, the lack of research focusing on the integration of these techniques as a pre-processing step for k-means clustering in journal article datasets poses a significant gap. This study seeks to address this gap by systematically exploring the benefits of using TF-IDF before applying the LDA Technique to improve the performance of k-means clustering, thereby contributing to the broader field of data mining and machine learning.

METHOD

This section describes the research flow and the proposed method, as presented in the following flowchart Fig.1:



Fig. 1 Research Step

Data Collection

This study began by manually collecting a dataset of scientific articles from Google Scholar containing title data and abstract data. There were up to 427 data used. Here is an example of the data obtained as visualized in Table 1 below:

Table 1. The sample of data

No.	Title	Abstract
1	PERFORMANCE ANALYSIS OF TAGGED TEXT CLASSIFICATION ALGORITHM IN DETECTING HOAX NEWS	Technology that has developed certainly has positive and negative impacts. The internet, which is the result of technological development, does not escape the negative impact, namely the emergence of hoaxes.....
2	TRANSMISSION SIMULATION AND VOLTAGE ANALYSIS IN STEAM POWER PLANT USING ETAP 12.6.0 SOFTWARE	Power systems are generally located between generating centers and load centers that are far apart.....
3	Educational Transformation in the Digital Era: Big Data Analysis to Increase Teacher Management Efficiency in Vocational High Schools	Big Data analysis plays a role in optimizing the efficiency of teacher management in SMK during the transformation of education in the digital era. This research.....
…	…	…
4	APPLICATION OF MACHINE LEARNING IN DETERMINING THE CLASSIFICATION OF	Child nutrition is a health problem that must be solved by the government. Malnutrition has become....

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

No.	Title	Abstract
	CHILDREN'S NUTRITION WITH DECISION TREE	
5	Implementation of Chatbot as Virtual Assistant at Panca Marga University Probolinggo using TF-IDF Method	Panca Marga University has an information service called Siakad (campus academic information system), but this service is...

The dataset collected and processed is originally in Indonesian. Consequently, all processing steps, including stopword removal, utilize libraries designed for the Indonesian language, such as NLTK for Indonesian. However, for this manuscript, the presentation is adjusted to English to align with the template requirements.

Data Pre-processing

The next step is data preprocessing which is done to improve data quality. This process includes several stages, namely tokenization and stopword removal. First, the article text is converted into tokens through the tokenization process, which separates the words in the text into individual elements. Next, stopword removal is performed using the *nltk* library.

Feature Extraction

After the preprocessing step is complete, the next stage is feature extraction using the proposed method whcih combine TF-IDF and LDA. The TF-IDF method is used to calculate the weight of words in each document by considering the frequency of occurrence of the word in the document as well as how rarely the word appears in the entire document collection. This process is expected to produce more relevant words in a particular document to get a higher weight.

First, TF-IDF vectorization is performed, where each document in the dataset is converted into a numerical representation based on the weight of the words in the document. In this process, the TfidfVectorizer approach is used. In addition, the number of features is limited to 2000 features to reduce the dimensionality of the data and ensure that only the most informative words are used in the analysis process.

After the TF-IDF process, latent topic extraction is performed using the Latent Dirichlet Allocation (LDA) method. The LDA model works by identifying the probabilistic distribution of topics in each document based on the word weights that have been calculated using TF-IDF. In this study, the number of topics was set at five, which was chosen based on initial exploration to find the optimal number of topics in the dataset. The LDA model is then trained using the TF-IDF matrix to generate document representations based on the topics found. The result of this process is a feature matrix that reflects the probability distribution of the five topics in each document. This combination of TF-IDF and LDA results is used as input for the clustering process with K-Means algorithm.

Clustering Algorithm

Basically, clustering is a grouping method to find and group data that has similar characteristics of Algorithms (similarity*)* between one data and other data. K-Means clustering is a technique for classifying data by identifying features that capture similarities dan differences among data point. In other words, the primary goal of clustering is to ensure that data within the same cluster exhibit strong similarities, while maintaining clear distinctions between different clusters (Wang et al., 2023):

Moreover, clustering method is the method based on unsupervised learning (Suyal & Sharma, 2024). Unsupervised Learning (UL) is one of the three primary paradigms in Machine Learning (ML), alongside Supervised Learning and Reinforcement Learning (Rolf et al., 2024). Unlike supervised learning, which depends on labeled data for training, and reinforcement learning, which utilizes approximate feedback through a reward system, unsupervised learning functions without labeled outputs or direct feedback (Benjamin, Rolf, et.al, 2024).

K-Means falls under the partition-based category, where data points are divided into mutually exclusive groups, with each data point assigned to only one cluster. The K-Means algorithm operates by calculating the distance between each data point to the cluster centroid. The standard steps of the K-Means algorithm are as follows (Wang et al., 2023):

1. Determine the number of clusters *k* and randomly select *k* initial centroids;
2. Compute the distance between each data point and all centroids. To calculate this distance, it use the Euclidean distance theory which is formulated into the following equation (Wang et al., 2023) :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

where:
d(x,y) = Distance of *x-th* data to *centroid y*

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

x_i and y_i = The *i-th* sample point
n = The number of sample points contained by x or y

3. Assign each data point to the cluster whose centroid is nearest;
4. Update the centroids by calculating the mean of all points in each cluster;
5. Repeat the steps (2) and (3) and updating centroids until the centroids stabilize.

Based on the flowchart image Fig.1 the next step is to perform clustering using the K-Means algorithm. With the extracted features, the K-Means algorithm is applied to cluster scientific articles based on titles that have been combined with their abstracts. As for determining the optimal number of clusters k, this research uses the Silhouette Analysis method by calculating the silhouette score value. The following are the steps applied using the K-Means algorithm as visualized in Fig. 2 (Wardy et al., 2022):

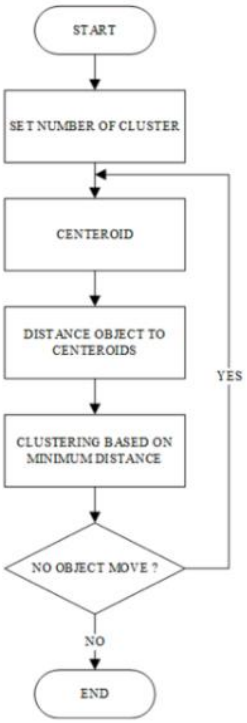


Fig. 2 Flowchart of K-Means Algorithm

The following steps are described as follows: 1) Start; 2) Determine the number of *clusters*, namely from the features that have been extracted with a combination of TF-IDF and LDA; 3) Initialize the initial value of the center point/centroid. In determining the center of the cluster, it is done randomly from the dataset; 4) Calculate the distance from each data point to the centroid using the Euclid distance formula; 5) Perform data clustering by assigning each data point in the dataset into the closest cluster based on the nearest distance/minimum distance; 6) After assigning the points, then the centroid is updated. If there is data that must move, then recalculate the position of each centroid. The process is completed until there is no change in the centroid and the centroid no longer move significantly; 7) Finish

Evaluation and Analysis

To assess the effectiveness of the method used, a performance evaluation is carried out with several key metrics, namely by using the Silhouette Score. Silhouette Score is used to measure how well each document is grouped in its cluster, with higher values indicating better cluster separation.

RESULT

Based on the results of the calculation of the optimal number of clusters applied, using Silhouette Score between the range of the number of clusters *k* is 2 to 16, then the following is the value of Silhouette Score obtained as presented in the following figure:

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

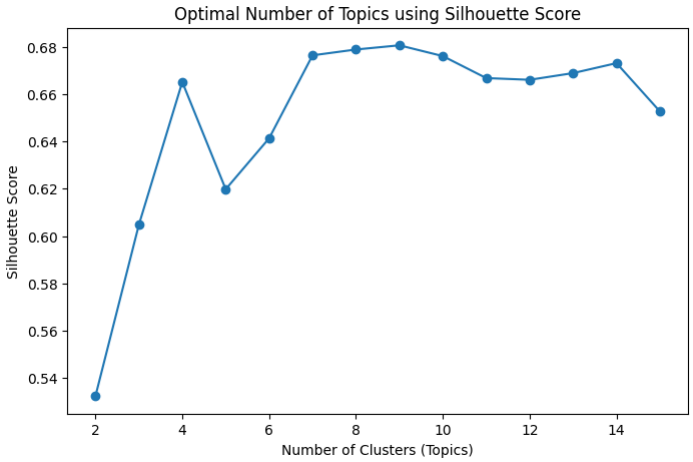


Fig. 3 Result of Silhouette Score

According to Figure 3, from the range of 2 to 16, the highest silhouette score is obtained at the number of clusters k = 9, which is **0.6806** as presented in the following Table 2:

Table 2. *The Result of Silhoutte Score in Each Number of Cluster*

No.	The number of clusters	Silhouette Score
1	2	0,5324
2	3	0,6048
3	4	0,6651
4	5	0,6198
5	6	0,6413
6	7	0,6764
7	8	0,6788
8	9	0,6806
9	10	0,6760
10	11	0,6668
11	12	0,6660
12	13	0,6689
13	14	0,6689
14	15	0,6731
15	16	0,6527

From the results obtained, the highest Silhouette Score value is achieved when the number of clusters is 9, with a score of 0.6806. This shows that at this number of clusters, the separation between clusters is quite clear and each document has a high correspondence with its respective cluster. In addition, as can be seen based on the table, increasing the number of clusters after 9 does not give a significant increase in the score, even experiencing a slight decrease, such as at 10 clusters (0.6760) and 11 clusters (0.6668).

Before reaching the optimal number of clusters, the Silhouette Score value increases gradually. For example, when the number of clusters is 2, the Silhouette Score value is still at 0.5324, which indicates that the documents in the cluster do not have optimal separation. However, as the number of clusters increases to 3 (0.6048) and 4 (0.6651), there is an improvement in the quality of separation. After reaching 7 to 9 clusters, the score continues to increase fairly steadily, with values of 0.6764 at 7 clusters, 0.6788 at 8 clusters, and finally peaking at 9 clusters. From this finding, it is concluded that dividing the documents into 9 clusters is the best choice in this study.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

DISCUSSION

According to the result, below is the word cloud visualization for each cluster:

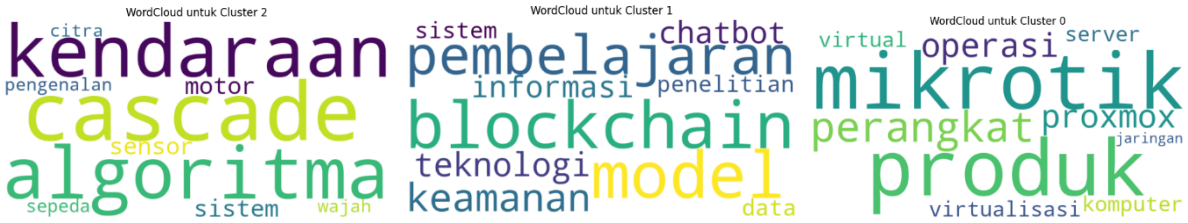


Fig. 4 WordCloud for Cluster 0, Cluster 1, and Cluster 2

Fig.4 visualized the WordCloud for Cluster 0, 1, and Cluster 2. Here is the interpretation of each cluster. WordCloud in Cluster 0 shows that the main topics in this cluster are related to computer networking and virtualization. Words such as *"mikrotik", "produk", "perangkat", "virtualisasi",* and *"proxmox"* indicate that the documents in this cluster focus on network devices, servers, and virtualization systems. In contrast to Cluster 0, Cluster 1 shows that the dominant topics are related to blockchain, security, and technology-based learning. Words such as *"blockchain", "pembelajaran", "model", "keamanan",* and "*chatbot*" indicate that the documents in this cluster discuss the application of blockchain technology, system security, and innovations in AI-based learning or chatbots,

Furthermore, the WordCloud analysis results for Cluster 2 show that the cluster focuses on vehicle recognition and image-based detection algorithms. Words such as *"kendaraan", "algoritma"*, and "cascade" indicate the use of Cascade Classifier or image processing methods for object identification, specifically vehicles. The presence of the terms "sensor", "motor", and "*sepeda*" indicate possible applications in traffic monitoring systems or vehicle security. In addition, the appearance of the word "face" leads to potential research in facial recognition for vehicle security systems. Overall, Cluster 2 reflects research related to computer vision, vehicle detection, and sensor-based systems, with potential applications in intelligent transportation and vehicle security.



Fig. 5 WordCloud for Cluster 4 to 8

Moving on, the WordCloud for Cluster 3 indicates that the documents in this group focus on data analysis, sentiment, and classification. Words such as *"data", "sentimen", "terbaik", "klasifikasi",* and *"kinerja"* indicate that the research in this cluster is related to data analysis, sentiment processing, and classification methods in various domains, such as performance evaluation or image processing.

Moreover, the terms in Cluster 4 to 8 are visualized using WordCloud as presented in Fig.5. WordCloud in Cluster 4 shows that this cluster is related to students, research, and academic systems. Words such as *"mahasiswa", "anak",* and *"hasil"* indicate that research in this cluster focuses on academic analysis, learning, or evaluation of student performance. The presence of the words *"bahasa", "basis",* and *"algoritma"* also indicates studies related to language processing or data-driven systems to support educational research. In Cluster 5, dominant terms such as *"server", "jaringan",* and *"komputer"* indicate a connection with information technology systems and data management. The words "hadith" and "water" indicate that there are studies that may be related to religious-based text processing or monitoring water resources with computer-based systems.

WordCloud for Cluster 6 shows the relationship between classification systems and sterilization processes. Words such as *"klasifikasi", "sterilisasi",* and *"sistem"* indicate the study of automating the sterilization process or grouping data in a particular system. The presence of the terms "firewall" and "*komputer*" also points to the application of information technology, possibly in the field of network security or sterilization process control.

In Cluster 7, words such as *"arsip", "kinerja",* and *"penilaian"* indicate that the research in this cluster focuses on

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

document management, performance assessment systems, and character analysis in digital archives. The presence of the term "SVM" indicates that the Support Vector Machine method is used in data classification or evaluation. Then, for Cluster 8, the appearance of the terms "*nilai*", "stmik", and "exam" leads to the analysis of academic grades and evaluation of the education system. In addition, the presence of the words "Twitter", "ChatGPT", and "*pelanggan*" indicate the possibility of research on user interaction on social media and its influence on digital learning or business.

CONCLUSION

This study demonstrates that integrating TF-IDF and LDA as feature extraction techniques significantly enhances the performance of K-Means clustering for journal article datasets. By applying a structured topic modeling approach, the proposed method effectively captures meaningful latent patterns, leading to improved clustering quality, as evidenced by the highest Silhouette Score of 0.6806 at k=9. The findings suggest that preprocessing with TF-IDF before applying LDA provides a more structured and informative text representation, ultimately optimizing clustering performance. The practical benefits of this research include improved accuracy in academic document clustering, with potential applications in journal recommendation systems, digital library indexing, and automated research trend analysis. The method can assist researchers and institutions in organizing large-scale textual data efficiently, reducing manual effort in categorization and information retrieval.

Despite its promising results, this study has several limitations. First, the dataset consists only of article abstracts, which may limit the richness of extracted features compared to full-text analysis. Second, the clustering approach relies solely on K-Means, which assumes spherical clusters and may not be optimal for complex text structures. For future research, it is recommended to explore alternative clustering techniques, such as hierarchical clustering or deep learning-based clustering methods, to assess their effectiveness in handling high-dimensional text data. Additionally, expanding the dataset to include full-text articles and incorporating semantic embeddings (e.g., Word2Vec, BERT) could further enhance feature extraction and clustering accuracy. By addressing these limitations, future studies can contribute to the ongoing development of text mining and document clustering methodologies.

REFERENCES

Aggarwal, A. G. (2024). *Online Reviews Based Ranking of Hotels Integrating Topic Modelling and Picture Fuzzy Approach*. https://doi.org/10.21203/rs.3.rs-3876413/v1

Al-Obaydy, W. N. I., Hashim, H. A., Najm, Y. A., & Jalal, A. A. (2022). Document Classification Using Term Frequency-Inverse Document Frequency and K-Means Clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(3), 1517. https://doi.org/10.11591/ijeecs.v27.i3.pp1517-1524

Chen, F., Yang, Y., Xu, L., Zhang, T., & Zhang, Y. (2019). *Big-Data Clustering: K-Means or K-Indicators?* https://doi.org/10.48550/arxiv.1906.00938

Dewald, C. L. A., Balandis, A., Becker, L. S., Hinrichs, J. B., Falck, C. v., Wacker, F., Laser, H., Gerbel, S., Winther, H. B., & Apfel, J. (2023). Automated Classification of Free-Text Radiology Reports: Using Different Feature Extraction Methods to Identify Fractures of the Distal Fibula. *Röfo - Fortschritte Auf Dem Gebiet Der Röntgenstrahlen Und Der Bildgebenden Verfahren*, *195*(08), 713–719. https://doi.org/10.1055/a-2061-6562

George, L., & Sumathy, P. (2023). An Integrated Clustering and BERT Framework for Improved Topic Modeling. *International Journal of Information Technology*, *15*(4), 2187–2195. https://doi.org/10.1007/s41870-023-01268-w

Ghosh, S. (2021). *Using Transformer Based Ensemble Learning to Classify Scientific Articles*. https://doi.org/10.48550/arxiv.2102.09991

Goh, C.-H., Wong, K. K., Tan, M. P., Ng, S.-C., Chuah, Y. D., & Kwan, B.-H. (2022). Development of an effective clustering algorithm for older fallers. *PLOS ONE*, *17*(11), e0277966. https://doi.org/10.1371/journal.pone.0277966

Im, Y., Park, J., Kim, M., & Park, K. (2019). Comparative Study on Perceived Trust of Topic Modeling Based on Affective Level of Educational Text. *Applied Sciences*, *9*(21), 4565. https://doi.org/10.3390/app9214565

Istiqomah, K. N., Widodo, I. D., Mufid, N. F., & Qurtubi, Q. (2023). Identifying Improvement Strategic From User Application Reviews Group Using K-Means Clustering and TF-IDF Weighting. *International Journal of Artificial Intelligence Research*, *7*(2), 152. https://doi.org/10.29099/ijair.v7i2.1062

Karthik, J., Tamizhazhagan, Dr. V., & Narayana, Dr. S. (2019). Data Leak Identification in Social Networks Using K Means Clustering &Amp; Tabu K Means Clustering. *International Journal of Innovative Technology and Exploring Engineering*, *9*(2), 2777–2783. https://doi.org/10.35940/ijitee.b6635.129219

Liu, B., Zhang, T., Li, Y., Liu, Z., & Zhang, Z. (2021). Kernel Probabilistic K-Means Clustering. *Sensors*, *21*(5), 1892. https://doi.org/10.3390/s21051892

Liu, H., Zhi-wang, C., Tang, J., Zhou, Y., & Liu, S. (2020). Mapping the Technology Evolution Path: A Novel Model for Dynamic Topic Detection and Tracking. *Scientometrics*, *125*(3), 2043–2090. https://doi.org/10.1007/s11192-020-03700-5

Muhajir, M., Rosadi, D., & Danardono, D. (2024). Improving the Term Weighting Log Entropy of Latent Dirichlet

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 4, **Number** 2, November 2024
https://doi.org/10.47709/brilliance.v4i2.5547

Allocation. *Indonesian Journal of Electrical Engineering and Computer Science*, *34*(1), 455. https://doi.org/10.11591/ijeecs.v34.i1.pp455-462

Nugroho, S. A., Bachtiar, F. A., & Wihandika, R. C. (2022). Aspect Extraction in E-Commerce Using Latent Dirichlet Allocation (Lda) With Term Frequency-Inverse Document Frequency (Tf-Idf). *Jurnal Ilmiah Kursor*, *11*(2), 53. https://doi.org/10.21107/kursor.v11i2.247

Pokharel, M., Bhatta, J., & Paudel, N. (2021). Comparative Analysis of K-Means and Enhanced K-Means Algorithms for Clustering. *Nuta Journal*, *8*(1–2), 79–87. https://doi.org/10.3126/nutaj.v8i1-2.44044

Rashid, J., Kim, J., Hussain, A., Naseem, U., & Juneja, S. (2022). A Novel Multiple Kernel Fuzzy Topic Modeling Technique for Biomedical Data. *BMC Bioinformatics*, *23*(1). https://doi.org/10.1186/s12859-022-04780-1

Rolf, B., Beier, A., Jackson, I., Müller, M., Reggelin, T., Stuckenschmidt, H., & Lang, S. (2024). A review on unsupervised learning algorithms and applications in supply chain management. *International Journal of Production Research*, 1–51. https://doi.org/10.1080/00207543.2024.2390968

Simoes, S., Deepak, P., & MacCárthaigh, M. (2022). *Exploring Rawlsian Fairness For K-Means Clustering*. 47–59. https://doi.org/10.1007/978-981-19-4453-6_3

Suyal, M., & Sharma, S. (2024). A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning. *Journal of Artificial Intelligence and Systems*, *6*(1), 85–95. https://doi.org/10.33969/AIS.2024060106

Wang, X., Shao, Z., Shen, Y., & He, Y. (2023). Research on fast marking method for indicator diagram of pumping well based on K-means clustering. *Heliyon*, *9*(10), e20468. https://doi.org/10.1016/j.heliyon.2023.e20468

Wardy, D. K., Putra, I. K. G. D., & Rusjayanthi, N. K. D. (2022). Clustering Artikel pada Portal Berita Online Menggunakan Metode K-Means. *JITTER : Jurnal Ilmiah Teknologi dan Komputer*, *3*(1), 985. https://doi.org/10.24843/JTRTI.2022.v03.i01.p34

Zhao, Y. (2022). *Exploring Redditors' Topics With Natural Language Processing*. https://doi.org/10.4995/carma2022.2022.15022

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.