Vol. 7 No. 2, 2026, pp. 144-160 **Jurnal Pepadun**
© 2026 Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0)

Recommendations for Branding Online Lectures in Indonesia Through Web Scraping with the Implementation of TF-IDF Vectorizer Techniques and K-Means Clustering Algorithm

[*1]Riad Sahara, [2]Syahid Abdullah, [3]Catur Nugroho, and [4]Ambros Magnus Rudolf Mekeng

[1,2,3,4]PJJ Informatika, Universitas Siber Asia, RM Harsono, Jakarta, Indonesia
e-mail: [*1]riadsahara@lecturer.unsia.ac.id, [2]syahidabdullah@lecturer.unsia.ac.id, [3]caturnugroho@lecturer.unsia.ac.id, [4]ambrosmekeng@lecturer.unsia.ac.id

Abstract *- The development of online lectures in Indonesia has encouraged universities to develop more targeted branding strategies, particularly in increasing brand awareness and digital engagement. Public perceptions reflected in open digital sources, particularly online news portals, are important data sources for understanding how online lectures are perceived by the public. This study aims to analyze public opinion on online lectures in Indonesia using web scraping, text mining, TF-IDF Vectorizer, and the K-Means Clustering algorithm. Initial data was obtained through web scraping using several Indonesian keywords related to online lectures and resulted in 108 text documents. After duplicate removal, noise filtering, and text preprocessing, the number of documents used in the analysis was reduced to 60 documents. The analysis stages included text cleaning, normalization, feature extraction using TF-IDF, determining the optimal number of clusters using the Elbow Method and Silhouette Score, and clustering using K-Means. The results showed that the optimal number of clusters was 4 clusters. The formed clusters represent the themes of online lecture academic activities, social and cultural dimensions, distance education and policy, as well as economic, regulatory, and technological aspects. The TF-IDF matrix yielded 1669 features, while the final evaluation yielded a Silhouette Score of 0.0345. PCA visualization showed overlap between clusters, indicating that public perception of online lectures is multidimensional. The conclusion of this study shows that the TF-IDF and K-Means approaches can be used to map public opinion and develop more contextual, data-driven, and relevant online lecture branding recommendations for universities.*

Keywords**:** *Brand Positioning; K-Means Clustering; Online Course; TF-IDF Vectorizer; Web Scraping*.

1. INTRODUCTION

The development of information technology has driven significant transformations in various sectors, including higher education [1]. This is especially true since the COVID-19 pandemic, which has accelerated the adoption of online lectures. Although some educational institutions have returned to face-to-face learning after the pandemic, online lectures remain a strategic choice, both for efficiency, expanded access, and learning flexibility. The phenomenon of online lectures has become increasingly relevant, especially after the pandemic, as people have begun to realize the flexibility and accessibility offered by online learning systems [2]. On the other hand, the public is now increasingly critical in choosing educational institutions, including in assessing the quality and credibility of online lecture programs. This has encouraged higher education institutions to strengthen their branding strategies, not only through conventional promotions, but also through data-driven approaches and a deeper understanding of public perception in the digital space. However, the challenge of building brand awareness and a positive image for online lecture programs in Indonesia remains a problem. Many higher education institutions face difficulties in positioning themselves amidst increasing competition and varying public perceptions of online education.

In this context, the use of data technologies such as web scraping, text mining, and machine learning opens opportunities to collect and analyze public opinion from various digital sources [3][4][5][6][7]. The development of digital technology and social media has generated large amounts of text data originating from public opinion, user reviews, and online discussions, necessitating text mining techniques to extract information and knowledge from this unstructured data [8]. One relevant approach is text feature extraction using the TF-IDF (Term Frequency - Inverse Document Frequency) method and opinion or topic grouping

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

using the K-Means clustering algorithm [9]. The combination of these two techniques can provide insight into perceptions and key themes developing in the community regarding online lectures, thus serving as the basis for developing a more targeted branding strategy [10].

Several previous studies have demonstrated the successful implementation of this approach. Research by Purniawan et al. clustered news results from web crawling from the Detik.com site using TF-IDF and K-Means, which proved effective in identifying dominant topics [10]. Fonda et al. utilized a similar method in analyzing public opinion on Twitter to support a travel agency's marketing strategy [11]. Meanwhile, Saputri et al. used TF-IDF and K-Means based on Cosine Similarity to cluster student internship titles [12]. In the field of disasters, Alfian applied the K-Means method in classifying disaster information from online text data [13]. Aryanti and Mahendra also integrated TF-IDF with classification methods for analyzing public opinion in Indonesian on social media [14]. Several studies have shown that document clustering techniques can be used effectively to analyze public opinion on social media and online platforms, including Twitter and Reddit [15].

However, there is still very little research that specifically examines the brand positioning of higher education institutions based on public opinion data related to online lectures through open digital data, such as social media and discussion forums using unsupervised learning approaches like K-Means. In fact, these platforms are a real reflection of public narratives, hopes, and criticisms that can be a source of valuable insights for educational institutions. The availability of large amounts of public opinion data in the digital space requires a systematic and objective analytical approach. Recent research trends also indicate the increasing need for adaptive and automated clustering methods, including the use of AutoML and more systematic cluster evaluation [16], [17]. This study seeks to fill this gap by offering a web-scraping-based approach to open digital sources, particularly online news portals, to explore public perceptions of online lectures in Indonesia. The obtained data will be processed using the TF-IDF method and then clustered with the K-Means algorithm to identify perception clusters that can be used as the basis for branding recommendations. In the context of branding, analysis of online opinion texts has been used to understand brand image and brand position in the minds of consumers, so that it can be the basis for developing more targeted digital communication strategies [18].

The use of text mining and clustering to formulate branding strategies for online lectures in Indonesia is still relatively limited, especially those based on digital public opinion data [19]. Therefore, the use of web scraping, text mining, and clustering techniques is important to extract patterns of public perception hidden in text data. Through this research, it is hoped that a thematic mapping of public opinion on online lectures in Indonesia will be obtained, which can be used as a strategic reference for Higher Education Institutions in developing more targeted and data-driven branding and communication approaches. These findings are also relevant for Distance Learning (PJJ) and Cyber Universities in Indonesia.

2. LITERATURE REVIEW

2.1. Literature Review

Branding online lectures in Indonesia has become increasingly important in the digital era, especially since the increased adoption of online learning post-pandemic. In this regard, data-driven approaches such as text mining and clustering have great potential to understand public perceptions of online-based higher education institutions. Text mining itself is an analytical approach that aims to extract information, patterns, and knowledge from large amounts of unstructured text data available [8]. One relevant approach is the use of the TF-IDF method and the K-Means algorithm in clustering public opinion from data obtained through web scraping. TF-IDF is used to represent text documents into numeric vectors by considering the relative importance of a word in the document and the entire corpus [20][21]. Furthermore, K-Means is one of the most widely used unsupervised learning algorithms in text clustering due to its simplicity and efficiency in handling large-scale data [22].

Purniawan et al. demonstrated the successful application of TF-IDF and K-Means in clustering news from detik.com into 27 groups based on content similarity [10]. This technique is highly relevant for clustering

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

public opinion from various online news sources related to educational institutions. Fonda et al. strengthened the relevance of this method in the business sector by extracting public opinion from Twitter as the basis for a travel agency's marketing policy [11]. This approach can be adapted for online college branding strategies, namely by analyzing how institutions are perceived by the public on social media. The combination of TF-IDF and K-Means has been applied in various studies to cluster text documents and social media data based on topic similarity [23][24]. Furthermore, cluster quality evaluation can be done using metrics such as Silhouette Score and visual analysis, as applied in social media document clustering studies [15][16].

Furthermore, Ompo and Pakereng studied how tweets related to Zalora Indonesia can be grouped to identify the type of content that attracts the most engagement [3]. This finding is important for the context of digital branding, because it can be applied in measuring the effectiveness of online campus communication strategies on platforms such as Twitter. Meanwhile, Simanjuntak et al. uses a Word Embedding approach to improve the quality of Indonesian language news clustering, which can enrich semantic analysis in mapping perceptions of online lectures [4]. In the realm of opinion analysis, Adhiatma and Qoiriah utilized the TF-IDF method and deep neural network for sentiment classification from customer reviews [9]. This technique is very applicable in evaluating prospective students' perceptions of online education services, which are recorded in digital reviews or comments. Analysis of online review texts and opinions has been used to evaluate brand image and brand positioning in various industrial sectors [18].

From these studies, it appears that the integration between web scraping, TF-IDF, and K-Means (as well as additional techniques such as Word Embedding and DNN) is very relevant and adaptive in the context of public opinion analysis. Therefore, applying similar techniques in online college branding studies in Indonesia can provide strategic insight for making promotional decisions and strengthening the identity of digital-based educational institutions. In the educational context, analysis of social media discussions can provide strategic insights to increase brand awareness and digital engagement of educational institutions [19][25].

2.2. Research Novelty

This research offers a new approach in analyzing public perceptions of online lectures in Indonesia using web scraping techniques for open digital sources, particularly online news portals, which are then processed using a combination of the TF-IDF Vectorizer and the K-Means Clustering algorithm. The novelty lies in the integration of these processes to form an online lecture branding strategy recommendation system based on public data, by using the dimensions of brand awareness and digital engagement as interpretive dimensions for strengthening the image of higher education institutions in the digital era.

2.3. Linkage to Literature Review

Research on text analysis and clustering has been extensively conducted in various contexts, such as news clustering [10], social media trend analysis [11], and digital marketing strategies [3]. The common approach used is a combination of TF-IDF as a feature extraction technique and K-Means as a clustering method to find patterns in large text data. The data sources used generally come from online news platforms or social media, and the data collection is done using crawling or web scraping. Several studies have shown the success of this approach in simplifying the classification of unlabeled documents, for example in news from detik.com [10], Twitter data in the tourism sector [11], and online business promotions such as Zalora [3]. In addition, advanced approaches have also been developed through a combination of TF-IDF with neural network methods for sentiment classification [9], as well as the use of word embedding to maintain semantic context [4]. However, the literature review has not found any research that specifically:

1. Integrates web scraping of digital content related to online lectures in Indonesia, particularly from open online news sources.
2. Uses clustering results as the basis for higher education branding strategies.
3. Uses brand awareness and digital engagement as interpretive dimensions to support educational institution branding recommendations.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

In contrast to previous research which focused on comparing clustering methods and topic modeling [26], this research focuses on the use of clustering results as a basis for recommendations for online lecture branding strategies. Therefore, this research occupies a unique position by combining proven techniques in text mining and clustering but applied specifically to the branding needs of online lectures in Indonesia. This approach is expected to not only provide a general overview of public perception but also provide a strong basis for developing higher education institutions' digital communication and marketing strategies in a more targeted manner.

3. RESEARCH METHODOLOGY

3.1. Research Design

This study uses an exploratory quantitative approach utilizing text mining and clustering techniques to analyze public opinion regarding online lectures in Indonesia. The web scraping method is used in the data collection stage to obtain public opinion data from open digital sources. The data collection and data analysis processes are carried out as two separate stages, with the web scraping dataset used as input for analysis in Google Colab. The collected dataset is then analyzed using Google Colab through several stages, namely text data preprocessing to clean and normalize the data, feature extraction using TF-IDF Vectorizer, and data clustering using the K-Means Clustering algorithm based on theme similarity. The TF-IDF and K-Means-based clustering approaches were chosen because they have been widely used in large-scale text document analysis and have proven effective in grouping documents based on content similarity [22][27]. The clustering results are analyzed and interpreted to produce recommendations for online lecture branding strategies based on public opinion data. This methodological approach was chosen because it can systematically process large amounts of unstructured text data and provide a thematic overview relevant to the needs of higher education branding.

In this study, public perception is operationally defined as the thematic representation of public opinions, narratives, or information regarding online lectures that appear in open digital sources and can be analyzed as text documents. Public perception is not measured as the individual attitudes of respondents; instead, it is mapped through document content similarity based on text features derived from TF-IDF. Consequently, this study does not involve sentiment analysis, nor does it aim to classify opinions as positive, negative, or neutral. Furthermore, the study eschews supervised classification approaches in favor of K-Means-based document clustering to group documents according to thematic proximity. Brand awareness and digital engagement are not included as numerical variables in the analysis model; rather, they serve as an interpretive framework for translating clustering results into branding strategy recommendations. Thus, the relationship between the four clusters and the dimensions of brand awareness and digital engagement is interpretive based on the dominant themes within each cluster rather than a causal relationship or a direct statistical measurement.

3.2. Research Flow

The research flow is exploratory with several evaluative and iterative stages, particularly in the data validation process and determining the number of clusters. The branching and iterations in the research flow reflect the decision-making process in obtaining representative clustering results. The research flow is visually shown in Figure 1 and will be explained further.

3.2.1. Problem Identification and Research Objectives

The research began with identifying problems related to the need for a branding strategy for online lectures in Indonesia and formulating research objectives.

3.2.2. Data Collection

a. Data Sources
The research data consisted of public opinion documents discussing online lectures in Indonesia. Data were collected from open digital sources, particularly online news portals, containing public information or narratives related to online lectures.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

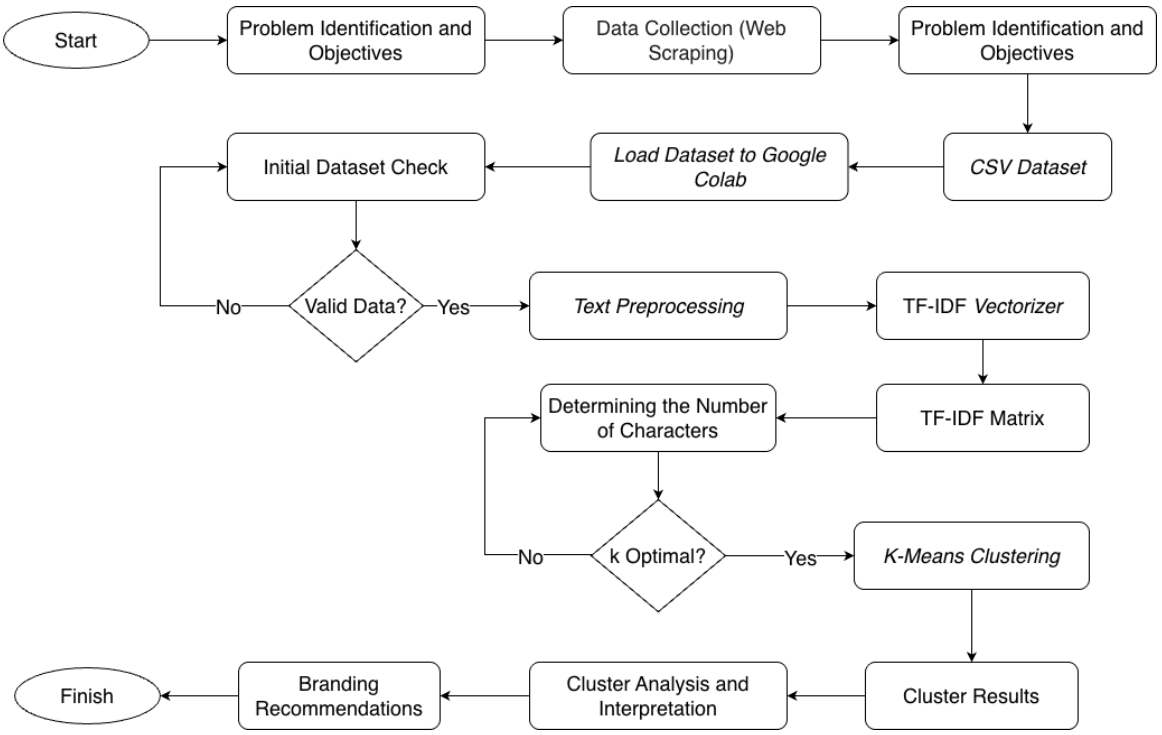


Figure 1. Research flowchart.

b. Data Collection Techniques
Data collection was conducted using web scraping techniques on public content relevant to the research topic. Web scraping-based digital data collection techniques have been widely used in public opinion analysis, document clustering, and social media mining [15], [28]. The data collection process utilized Python libraries appropriate to the development environment. The web scraped dataset was then saved in CSV format and used as input for subsequent analysis.

c. Dataset Description and Selection Criteria
The dataset used in this study was obtained through web scraping from open digital sources containing public information, opinions, or narratives related to online lectures in Indonesia. The scraping process was designed to collect data from several sources, including online news portals and online discussion forums. Based on the scraping scripts, the targeted sources included Detik, Kompas, CNN Indonesia, and Kaskus. However, the raw dataset that was successfully collected and used in the final analysis consisted of documents from Detik and Kompas. Each document contains four main attributes: date, source, keyword, and text. The date column represents the data collection date, the source column indicates the origin of the document, the keyword column shows the search keyword used during scraping, and the text column contains the main textual content extracted from the source.

Data collection was conducted on December 25, 2025, using five search keywords: “kuliah online”, “kuliah daring”, “pembelajaran jarak jauh”, “kampus online”, and “universitas online”. The initial scraping process produced 108 text documents, consisting of 58 documents from Detik and 50 documents from Kompas. Based on keyword distribution, the initial dataset consisted of 22 documents for “kuliah online”, 22 documents for “kuliah daring”, 22 documents for “pembelajaran jarak jauh”, 21 documents for “kampus online”, and 21 documents for “universitas online”.

The inclusion criteria in this study were as follows: documents had to be obtained from open digital sources, contain content relevant to online lectures or distance learning in the context of higher education, include processable textual content, and be retrieved using the predefined search keywords. Meanwhile, the exclusion criteria included duplicate documents, empty texts, very short texts, website

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

navigation fragments, advertisements, system-generated text, and textual content that did not substantively represent public information or public opinion related to online lectures.

During the initial data inspection, no missing values were found in the raw dataset. However, 48 duplicate documents were identified based on identical content in the text column. These duplicate documents were removed to prevent repeated information from influencing the clustering results. After duplicate removal, the dataset was reduced from 108 documents to 60 documents. Therefore, 48 documents, or approximately 44.44% of the initial dataset, were removed, while 60 documents, or approximately 55.56%, were retained as the clean dataset.

Text preprocessing was then conducted through several stages, including case folding, removal of non-alphabetic characters, tokenization, Indonesian stopword removal, and stemming. After preprocessing, an additional filtering step was applied based on the length of the clean_text column, with a threshold of more than 20 characters. This threshold was used to ensure that each document contained sufficient textual information to be represented using TF-IDF. The inspection results showed that all 60 cleaned documents had a clean_text length greater than 20 characters; therefore, no additional documents were removed at this stage.

3.2.3. Text Data Preprocessing

Data preprocessing was carried out to clean and normalize text data before further analysis. The preprocessing steps include:

a. Case folding (converting text to lowercase);
b. Cleaning of non-alphabetical characters (URLs, punctuation, and symbols);
c. Tokenization;
d. Removal of Indonesian stopwords; and
e. Stemming using the Indonesian stemming algorithm.

This stage is implemented using the Python programming language in the Google Colab environment.

3.2.4. Feature Extraction with TF-IDF

The preprocessed text data is then represented numerically using the TF-IDF Vectorizer. This method is used to measure the importance of a word in a document relative to the entire document collection. Mathematically, the TF-IDF weight of a term in a document can be written as follows:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

with:

$$IDF(t) = \log\left(\frac{1+N}{1+DF(t)}\right) + 1 \quad (2)$$

where $TF(t,d)$ indicates the frequency of occurrence of term t in document d, N indicates the total number of documents, and $DF(t)$ indicates the number of documents containing term t. This equation indicates that words that frequently appear in a single document but are less common across all documents will have a higher weight.

3.2.5. Clustering with K-Means

The K-Means Clustering algorithm is used to group public opinion data based on TF-IDF feature similarities. The K-Means algorithm is used because of its ability to efficiently group text data after feature extraction [22], [29]. The number of clusters is determined through exploratory analysis to obtain representative clustering results. The main goal of the K-Means algorithm is to minimize the distance between each data point and the nearest cluster center. The objective function of K-Means can be written as follows:

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - \mu_j \|^2 \quad (3)$$

where J is the total distance within the cluster, k is the number of clusters, x_i is the TF-IDF representation of the document vector, C_j is the j-th cluster, and μ_j is the centroid of the j-th cluster. The smaller the J value, the closer the document is to its cluster center. The centroid of each cluster is calculated based on the average of all data within that cluster. The centroid calculation can be written as follows:

$$\mu_j = \frac{1}{| C_j |} \sum_{x_i \in C_j} x_i \quad (4)$$

where μ_j is the centroid of the j-th, $| C_j |$ is the number of data points in the j-th, and x_i is the vector of documents belonging to that cluster. This equation is used to update the centroid position at each K-Means iteration.

3.2.6. Analysis and Interpretation of Results

The clustering results were analyzed by identifying dominant keywords in each cluster and interpreting the resulting public perception themes. Dimensionality reduction and cluster visualization were used to aid the interpretation of the clustering results, as recommended in a document cluster evaluation study [16]. The interpretation of the clustering results served as the basis for developing recommendations for branding online lectures in Indonesia.

3.2.7. Implementation Environment

The entire data analysis process, from preprocessing to clustering and evaluation, was implemented using Google Colab with the Python programming language.

4. RESULTS AND DISCUSSION

4.1. Description of Initial Data from Web Scraping Results

Research data was obtained through web scraping from online digital media sources using several Indonesian keywords related to online lectures. The initial data collection yielded 108 text documents, consisting of the titles and content of news articles/public opinion related to online lectures in Indonesia, as shown in Figure 2.

```
Initial data count: 108
```

	date	source	keyword	text
0	2025-12-25	Detik	kuliah online	\r\n ADVERTISEMENT\r\n Kementerian Pendidik...
1	2025-12-25	Detik	kuliah online	Universitas Sumatera Utara mengumumkan pihakny...
2	2025-12-25	Detik	kuliah online	Anda menyukai artikel ini Artikel disimpan Saa...
3	2025-12-25	Detik	kuliah online	Mahasiswi di Magelang ini rela tiap hari duduk...
4	2025-12-25	Detik	kuliah online	\r\n ADVERTISEMENT\r\n Pemerintah Finlandia...

Figure 2. Example of web scraping data from online lecture opinions (initial data).

Each document has date, source, keyword, and text attributes. At this stage, the data still contains various forms of noise, such as advertisement text, media system notifications, and snippets of text that do not substantively represent public opinion. Table 1 presents the distribution of documents based on the data sources used in this study. Although the scraping process was designed to target several open digital sources, including online news portals and online discussion forums, the final raw dataset that was successfully collected and processed consisted of documents from Detik and Kompas. Table 1 shows the number of documents obtained from each source before and after the data selection and duplicate removal process.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

Table 1. Distribution of documents by data source.

Data Source	Source Category	Initial Documents	Documents After Selection
Detik	Online news portal	58	50
Kompas	Online news portal	50	10
CNN Indonesia	Online news portal	0	0
Kaskus	Online discussion forum	0	0
Total		**108**	**60**

In addition to the source-based distribution, the dataset was also examined based on the search keywords used during the scraping process. Table 2 shows the distribution of documents for each keyword in the initial dataset and after the data selection process. This information is important to clarify how each keyword contributed to the final dataset used in the TF-IDF and K-Means clustering analysis.

Table 2. Distribution of documents by search keyword.

Search Keyword	Initial Documents	Documents After Selection
kuliah online	22	22
kuliah daring	22	8
pembelajaran jarak jauh	22	12
kampus online	21	9
universitas online	21	9
Total	**108**	**60**

4.2. Noise Filtering and Clean Dataset Formation

After the initial inspection of 108 documents, duplicate removal was performed based on identical textual content in the text column. A total of 48 duplicate documents were identified and removed, reducing the dataset to 60 documents. This means that 44.44% of the initial dataset was removed, while 55.56% was retained for further analysis. Subsequently, filtering based on the clean_text length was applied using a threshold of more than 20 characters. This threshold was used to remove texts that were too short to provide meaningful thematic information for TF-IDF representation. However, all 60 cleaned documents met this threshold, so no additional documents were removed during the length-based filtering stage. An example of the dataset after the filtering and text preprocessing processes is shown in Figure 3.

```
df = df[df["clean_text"].str.len() > 20].reset_index(drop=True)
print("Number of data after noise filtering:", len(df))
```

```
Number of data after noise filtering: 60
```

Figure 3. Dataset after noise filtering and text preprocessing.

	date	source	keyword	text	clean_text
0	2025-12-25	Detik	kuliah online	\r\n ADVERTISEMENT\r\n Kementerian Pendidik...	advertisement menteri didik sains teknologi ke...
1	2025-12-25	Detik	kuliah online	Universitas Sumatera Utara mengumumkan pihakny...	universitas sumatera utara umum kuliah serang ...
2	2025-12-25	Detik	kuliah online	Anda menyukai artikel ini Artikel disimpan Saa...	suka artikel artikel simpan media sosial pisah...
3	2025-12-25	Detik	kuliah online	Mahasiswi di Magelang ini rela tiap hari duduk...	mahasiswi magelang rela duduk pinggir jalan ku...

Figure 4. Filtered dataset.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

The filtered dataset was then used as a clean dataset for all subsequent analysis stages. The clean dataset, as seen in Figure 4, has the following column structure: date; source; keyword; text; clean_text. The clean_text column represents the preprocessed text, which has undergone case folding, non-alphabetic character cleaning, Indonesian stopword removal, and stemming. Qualitatively, the preprocessed text displays more concise words and focuses on the context of online lectures, such as lecture, online, student, campus, and university.

4.3. Feature Extraction Using TF-IDF Vectorizer

The clean dataset was then transformed into a numerical representation using the TF-IDF Vectorizer. The TF-IDF configuration used included the use of unigrams and bigrams (ngram_range = (1,2)), as well as word occurrence restrictions using min_df and max_df, as shown in Figure 5.

```
tfidf_vectorizer = TfidfVectorizer(
    ngram_range=(1, 2),
    min_df=2,
    max_df=0.9,
    max_features=5000
)
```

Figure 5. TF-IDF configuration.

The feature extraction results produce a 60 × 1669 TF-IDF matrix, indicating that each document is represented by 1669 word features. This matrix serves as the primary basis for the clustering process using the K-Means algorithm.

4.4. Determining the Optimal Number of Clusters

4.4.1. Evaluation with the Elbow Method

The number of clusters was determined exploratively using the Elbow Method by observing the change in inertia values for k = 2 to k = 10. Inertia represents the within-cluster sum of squared distances, where a lower value indicates that documents are closer to their assigned cluster centroids. As shown in Table 3, inertia continuously decreased as the number of clusters increased. However, the reduction became more gradual after k = 4, indicating that adding more clusters produced a smaller marginal improvement in compactness. Therefore, k = 4 was considered a reasonable candidate based on the Elbow Method, especially because it provided a balance between cluster compactness and interpretability.

Table 3. Inertia and silhouette score for different numbers of clusters.

Number of Clusters (k)	Inertia	Silhouette Score
2	53.2544	0.0210
3	50.9774	0.0310
4	49.2695	0.0345
5	48.1270	0.0302
6	46.8219	0.0351
7	45.0072	0.0428
8	43.6214	0.0477
9	42.2191	0.0565
10	40.8757	0.0515

The Elbow graph, as shown in Figure 6, shows a relatively sharp decrease in inertia at small k values and begins to level off after k = 4. This finding indicates that using more than four clusters does not significantly improve the quality of data clustering.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

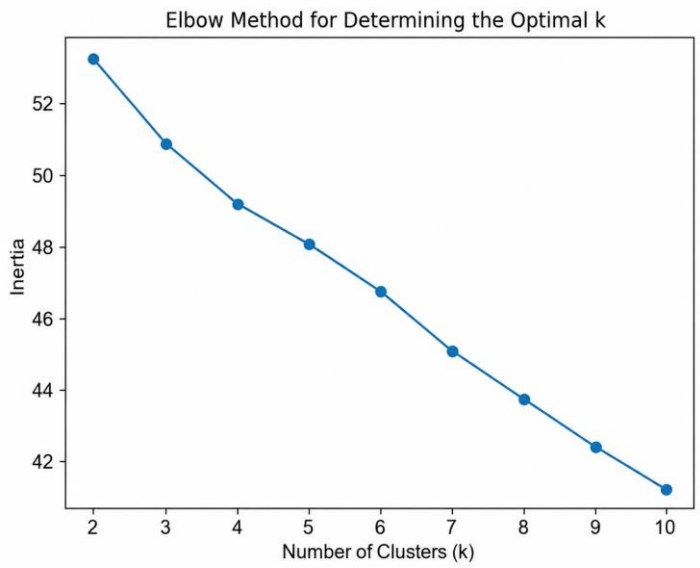


Figure 6. Elbow Method graph for determining the optimal number of clusters.

4.4.1. Evaluation with Silhouette Score (Multi-k)

In addition to the Elbow Method, an evaluation was also conducted using the Silhouette Score to measure the quality of cluster separation. The Silhouette Score is used to measure how close a data set is to its respective cluster compared to other nearby clusters. Mathematically, the silhouette value for a data set can be written as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

where $a(i)$ is the average distance of data i to all other data in the same cluster, while $b(i)$ is the average minimum distance of data i to the data in the closest other cluster. A higher Silhouette Score indicates that the data is more closely aligned with the cluster it belongs to.

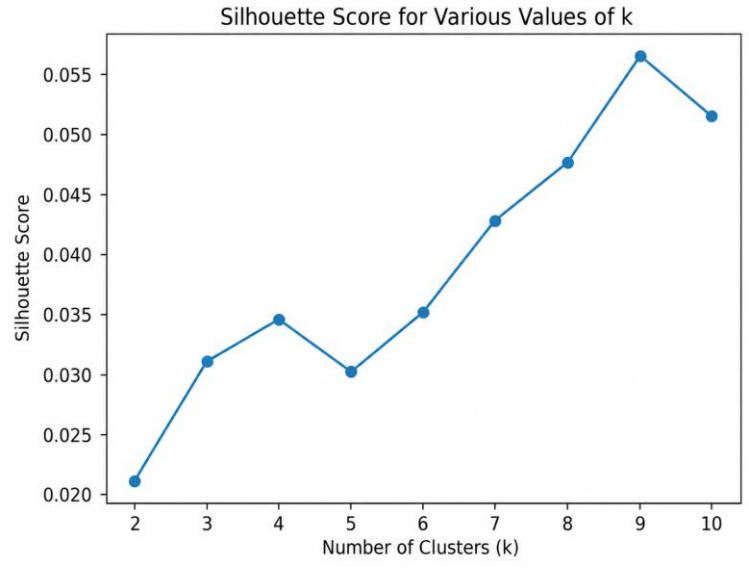


Figure 7. Silhouette score values for various numbers of clusters.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

In addition to the Elbow Method, Silhouette Score was used to evaluate the degree of separation between clusters. The Silhouette Score values for k = 2 to k = 10 are also presented in Table 3. The results show that the Silhouette Score values were generally low across all tested k values, indicating weak numerical separation among clusters. The highest Silhouette Score was obtained at k = 9, with a value of 0.0565. However, this value remained very low and produced too many clusters for a dataset consisting of only 60 documents. In such a small and heterogeneous text dataset, a higher k value tends to fragment the documents into smaller groups, which may reduce the practical interpretability of the resulting themes. Changes in the Silhouette Score value against variations in the number of clusters are shown in Figure 7.

Although k = 9 produced the highest Silhouette Score, this study selected k = 4 as the final number of clusters based on three considerations. First, the Elbow Method showed that the decrease in inertia began to become more gradual after k = 4. Second, k = 4 produced thematically interpretable clusters that could be mapped into meaningful branding dimensions, namely academic activities, social and cultural dimensions, distance education and policy, and economic, regulatory, and technological aspects. Third, the study aimed to generate branding recommendations rather than maximize predictive clustering performance. Therefore, k = 4 was selected as a compromise between numerical evaluation, thematic interpretability, and the practical objective of developing online lecture branding recommendations.

4.5. Clustering Results Using K-Means

The K-Means algorithm was implemented using a fixed random seed (random_state = 42) and 10 initializations (n_init = 10) to reduce the influence of random centroid initialization. The final clustering model with k = 4 produced a Silhouette Score of 0.0345 and an inertia value of 49.2695. The resulting cluster distribution was 17 documents in Cluster 0, 10 documents in Cluster 1, 8 documents in Cluster 2, and 25 documents in Cluster 3. A separate repeated-run stability test across multiple random seeds was not conducted in this study. Therefore, clustering stability was controlled through a fixed random seed and multiple initializations. The low Silhouette Score and overlapping PCA visualization indicate that the resulting cluster structure should be interpreted as exploratory thematic groupings rather than strongly separated or predictive clusters.

4.6. Top-Term Analysis and Cluster Theme Interpretation

Cluster theme analysis was conducted by identifying the words with the highest TF-IDF weights at each cluster centroid (top terms per cluster). The analysis results show the clusters as follow:

a. Cluster 0: *kuliah, daring, mahasiswa, ajar, kampus, universitas*
b. Cluster 1: *natal, wisuda, keluarga, agama, selamat*
c. Cluster 2: *sekolah, siswa, ajar jarak, dinas, pjj*
d. Cluster 3: *aplikasi, harga, rp, juta, riset, pasal*

The resulting cluster pattern shows similarities to the results of previous research on Twitter data that grouped opinions based on the main discussion themes [24][30], as shown in the following cluster thematic interpretation.

a. Cluster 0 – Academic Activities of Online Lectures in Higher Education
   Represents public perception of the core activities of online lectures, such as the learning process, the role of lecturers, and student experiences.
b. Cluster 1 – Social, Cultural, and Academic Moments
   Shows that online lecture discourse is also related to social and cultural values and important academic moments such as graduations and religious celebrations.
c. Cluster 2 – Distance Education and the Role of Government Institutions
   Depicts public perception, which still associates online lectures with national distance education policies and the role of government institutions.
d. Cluster 3 – Economic Aspects, Regulations, and Campus Technology Innovation
   Highlights aspects of education costs, the use of digital applications, and regulations and policies that influence the implementation of online lectures.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

4.7. Implications for Brand Awareness and Digital Engagement

The clustering results provide interpretive implications for online college branding strategies. The relationship between the clusters and branding dimensions in this study is based on the interpretation of dominant keywords and main themes in each cluster. Thus, brand awareness and digital engagement are positioned as interpretive outcomes of document clustering, not as variables directly calculated by the K-Means algorithm.

a. Brand Awareness. Clusters 0 and 2 demonstrate the importance of building the image of online colleges as a legitimate, structured, and high-quality higher education system.
b. Digital Engagement. Clusters 1 and 3 emphasize the importance of emotional engagement, social value, cost transparency, and ease of use of digital platforms to increase public engagement.

These cluster findings can be used as a basis for developing digital branding and communication strategies, in line with research that utilizes text analysis to understand brand perceptions and image [18], [19]. A cluster-based branding approach allows for the development of more contextual and targeted communication strategies, as shown in Table 4.

Table 4. Branding Strategy Recommendations Based on Public Opinion Clusters

Cluster	Main Theme	Branding Recommendations
Cluster 0	Academic Activities for Online Lectures	Emphasize the quality of the learning process, lecturer competence, and students' academic experience in online courses.
Cluster 1	Social and Cultural Dimensions	Integrate online course narratives with social values, family values, and academic moments such as graduation ceremonies.
Cluster 2	Distance Education and Policy	Clarify the differentiation between online university courses and distance learning at schools and affirm the institution's legitimacy.
Cluster 3	Economics, Regulation, and Technology	Display cost transparency, easy application access, and compliance with educational regulations.

4.8. Final Clustering Evaluation with Silhouette Score

The final clustering evaluation yielded a Silhouette Score of 0.0345. This value indicates that the separation between clusters is still relatively weak and there are overlapping themes between clusters. However, in the context of heterogeneous text-based public opinion data and a limited data set (60 documents), this value is still acceptable for exploratory and interpretative purposes. The focus of this research is on mapping themes and public perceptions, not on optimizing cluster performance predictively.

4.9. Dimensionality Reduction and Cluster Visualization Using PCA

To aid interpretation of the clustering results, dimensionality reduction was performed using Principal Component Analysis (PCA) into two main components. Simply put, the PCA process can be understood as the projection of high-dimensional data into a lower-dimensional space through principal components. This projection can be written as follows:

$$Z = XW \tag{6}$$

where X is the TF-IDF data matrix, W is the principal component matrix, and Z is the projection of the data into a lower-dimensional space. In this study, PCA was used to visualize the distribution of clusters into two principal components. The visualization of the clustering results in the two-dimensional PCA space is shown in Figure 8, which shows the distribution and overlap between clusters.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

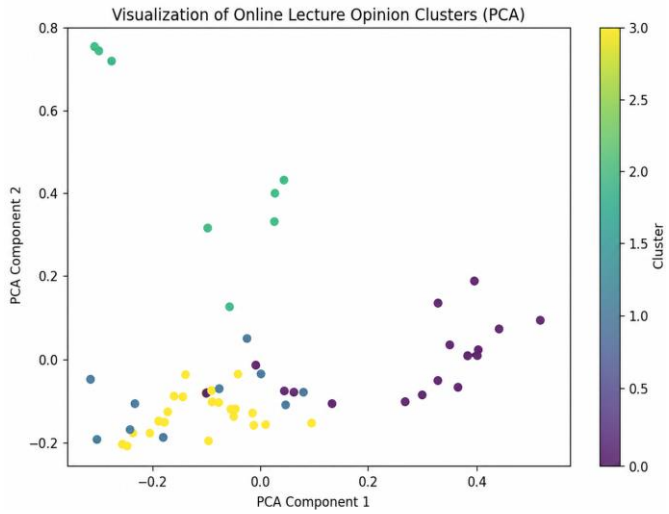


Figure 8. Visualization of online lecture opinion clusters using PCA.

PCA results show that the first component explains approximately 5.24% of the data variation, whereas the second component explains approximately 5.18% of the data variation. The total variation explained by the two principal components is approximately 10.4%, indicating high complexity in TF-IDF-based text data. Cluster visualization in the two-dimensional PCA space shows significant overlap between clusters. This finding is consistent with the low Silhouette Score and strengthens the conclusion that public perceptions of online lectures are multidimensional and overlapping.

4.10. Discussion

The results of this study show that TF-IDF Vectorizer and K-Means Clustering can be used to identify thematic structures in public discourse related to online lectures in Indonesia. The four clusters generated in this study represent academic activities of online lectures, social and cultural dimensions, distance education and policy, and economic, regulatory, and technological aspects. These findings indicate that public perception of online lectures is not formed by a single dominant issue, but by several interrelated themes. This result is in line with previous studies showing that TF-IDF and K-Means can be used to group text documents based on content similarity, such as online news clustering [10] and Twitter-based public opinion clustering for marketing policy [11]. However, unlike those studies, this research applies document clustering specifically to the context of online lecture branding in higher education.

The low Silhouette Score obtained in this study indicates that the clusters were not strongly separated numerically. This condition can be explained by several factors. First, the dataset used in this study was relatively small, consisting of only 60 documents after duplicate removal and preprocessing. A limited number of documents can make cluster boundaries less stable because each cluster is formed from a relatively small number of textual observations. Second, the TF-IDF matrix produced 1669 features from only 60 documents, indicating a high-dimensional and sparse representation. In text clustering, sparsity can reduce the ability of distance-based algorithms such as K-Means to clearly separate documents, because many documents may share only a small number of common terms.

Third, the heterogeneous nature of the scraped sources also contributed to the overlapping cluster structure. The dataset contained texts from online news portals with different writing styles, topics, and contextual focuses. As a result, several documents may contain overlapping terms related to education, technology, policy, and social issues at the same time. This explains why the PCA visualization showed overlapping clusters and why the Silhouette Score remained low. Similar challenges have also been discussed in document clustering and topic modeling studies on social media data, where short, noisy, and heterogeneous texts often produce overlapping thematic structures [15].

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

Compared with previous studies, the findings of this study confirm the usefulness of TF-IDF and K-Means as an initial exploratory approach for identifying dominant themes in textual data. Purniawan et al. [10] applied TF-IDF and K-Means to cluster online news articles, while Fonda et al. [11] used K-Means clustering on Twitter data to support marketing policy. In contrast, this study does not only group documents, but also interprets the resulting clusters as a basis for online lecture branding recommendations. Therefore, the main contribution of this study lies in translating document clustering results into strategic branding implications, particularly for strengthening brand awareness and digital engagement in higher education.

The results also show that the use of TF-IDF has limitations in capturing semantic similarity between documents. TF-IDF represents documents based on term frequency and inverse document frequency, but it does not fully capture contextual meaning or semantic relationships between words. Previous studies using word embeddings have shown that embedding-based representations can improve text clustering by capturing semantic relationships more effectively [23], [24]. Therefore, future studies may consider using word embeddings or large language model embeddings to obtain more semantically meaningful clusters, especially when dealing with heterogeneous public opinion texts.

In relation to document clustering and topic modelling, this study is closer to document clustering because it assigns each document to a cluster based on TF-IDF feature similarity. It does not perform sentiment analysis and does not classify opinions into positive, negative, or neutral categories. This distinction is important because previous research has compared document clustering and topic modelling as different approaches for discovering thematic structures in text data [26]. The present study contributes to this area by showing how document clustering can be adapted for a practical branding context, where the resulting clusters are not treated merely as computational outputs, but as interpretative foundations for developing more contextual online lecture branding strategies.

Overall, the low numerical cluster separation should not be interpreted as a failure of the study, but as an indication that public perception of online lectures is multidimensional and overlapping. The clusters provide exploratory thematic insights rather than predictive or strongly separated groupings. For practical branding purposes, these thematic insights are still useful because they help universities understand which issues are associated with online lectures in public discourse. Nevertheless, the clustering quality may be improved in future studies by expanding the data collection period, increasing the number of documents, incorporating more diverse digital sources, and using semantic text representations to reduce the limitations of sparse TF-IDF features.

5. CONCLUSIONS

Based on the research results, it can be concluded that web scraping, text mining, TF-IDF Vectorizer, and K-Means Clustering approaches can be used to analyze public perceptions of online lectures in Indonesia. Public opinion data was collected from open digital sources using several Indonesian keywords related to online lectures. The data was then stored in a dataset and analyzed using Google Colab. The initial data collection consisted of 108 text documents. After duplicate removal, noise filtering, and preprocessing, the data used in the analysis decreased to 60 documents. The preprocessing steps included text cleaning, normalization, removal of irrelevant characters, stopword removal, and stemming, thus preparing the text data for analysis.

Next, the cleaned text data was represented numerically using TF-IDF Vectorizer. The feature extraction results produced a 60 × 1669 TF-IDF matrix, indicating that each document is represented by 1669 feature words. This matrix was then used as input in the clustering process using the K-Means algorithm. Based on the evaluation using the Elbow Method and Silhouette Score, the number of clusters used in this study was determined to be four. The clustering results indicate that public perception of online lectures is not a single entity but rather is divided into several main themes: academic activities in online lectures, social and cultural dimensions, distance education and policy, and economic, regulatory, and technological aspects.

The final evaluation using the Silhouette Score obtained a value of 0.0345, indicating that the separation between clusters is relatively weak and there are overlapping themes between documents. This finding is

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

further reinforced by the PCA visualization, which shows that the distribution of clusters in two-dimensional space still overlaps. However, in the context of exploratory research based on heterogeneous and limited public opinion text data, these results still provide a useful initial overview for understanding public perception patterns toward online lectures in Indonesia.

Overall, this study demonstrates that the results of public opinion clustering can be used as a basis for developing more contextual and data-driven recommendations for online lecture branding. Clusters related to academic activities and distance learning policies can be used to strengthen brand awareness, particularly in building the image of online lectures as a legitimate, structured, and high-quality higher education system. Meanwhile, clusters related to social, cultural, economic, regulatory, and technological aspects can be utilized to increase digital engagement through communication narratives that are more emotional, transparent, easily accessible, and tailored to community needs. Thus, this research provides a practical contribution to universities in developing more targeted online lecture branding strategies, as well as providing academic contributions in the application of text mining and clustering techniques in the context of higher education in Indonesia. Future studies are recommended to expand the data collection period and include a more diverse range of digital sources to increase the number of documents analyzed. Such data expansion is expected to reduce sparsity in the TF-IDF matrix which comprises 1,669 features in this study and improve numerical cluster separation.

LITERATURE

[1] E. Wahyudinarti, P. Andini Rachmatika, dan R. Nurul Ain, "Meningkatkan Efektivitas Pembelajaran Mahasiswa dengan AI," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, hlm. 488–491, Des 2024, doi: 10.36040/jati.v9i1.12279.

[2] R. Sahara, S. Abdullah, dan M. I. Saputra, "Integration Model of Academic Information Systems and Learning Management Systems with REST Web Services Using External Databases," *Journal of Information and Organizational Sciences*, vol. 47, no. 2, hlm. 373–384, Des 2023, doi: 10.31341/jios.47.2.7.

[3] F. A. A. Ompo dan M. A. I. Pakereng, "Penerapan Text Mining Untuk Advertising Pada Data Tweets Zalora Indonesia Dengan Menggunakan Metode K-Means Clustering," *Progresif: Jurnal Ilmiah Komputer*, vol. 20, no. 1, hlm. 314, Feb 2024, doi: 10.35889/progresif.v20i1.1576.

[4] H. T. A. Simanjuntak, P. E. P. Silaban, J. K. S. Manurung, dan V. H. Sormin, "Klasterisasi Berita Bahasa Indonesia Dengan Menggunakan K-Means Dan Word Embedding," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 3, hlm. 641–652, Jul 2023, doi: 10.25126/jtiik.20231026468.

[5] M. T. Razaq, D. Nurjanah, dan H. Nurrahmi, "Analisis Sentimen Review Film Menggunakan Naive Bayes Classifier dengan Fitur TF-IDF," *eProceedings of Engineering*, vol. 10, no. 2, hlm. 1698–1712, Apr 2023, Diakses: 15 April 2025. [Daring]. Tersedia pada: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/19997

[6] A. Aziz dan Fauziah, "Analisis Sentimen Identifikasi Opini Terhadap Produk, Layanan dan Kebijakan Perusahaan Menggunakan Algoritma TF-IDF dan SentiStrength," *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 6, no. 1, hlm. 115–125, Mar 2022, Diakses: 15 April 2025. [Daring]. Tersedia pada: https://tunasbangsa.ac.id/ejurnal/index.php/jsakti/article/view/430

[7] A. Gondowastradmojo dan R. Kusumastuti, "Analisis Sentimen Ulasan di Tokopedia Dengan Metode Naive Bayes," dalam *Seminar Nasional AMIKOM Surakarta (SEMNASA) 2024*, Sukoharjo: STMIK Amikom Surakarta, Nov 2024, hlm. 369–382.

[8] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, dan M. R. Yeganegi, "Text Mining in Big Data Analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, hlm. 1, Jan 2020, doi: 10.3390/bdcc4010001.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

[9] F. D. Adhiatma dan A. Qoiriah, "Penerapan Metode TF-IDF dan Deep Neural Network untuk Analisa Sentimen pada Data Ulasan Hotel," *Journal of Informatics and Computer Science (JINACS)*, hlm. 183–193, Nov 2022, doi: 10.26740/jinacs.v4n02.p183-193.

[10] I. M. A. Purniawan, G. M. A. Sasmita, dan I. P. A. E. Pratama, "Clustering Berita menggunakan Algoritma TF-IDF dan K-Means dengan Memanfaatkan Sumber Data Crawling pada Situs Detik.com," *JITTER : Jurnal Ilmiah Teknologi dan Komputer*, vol. 3, no. 1, hlm. 821–830, Jan 2022, doi: 10.24843/JTRTI.2022.v03.i01.p18.

[11] H. fonda Fonda, B. Mustika, E. Sabna, dan Y. Irawan, "Implementasi Text Mining pada Twitter dengan Algoritma K-Means Clustering sebagai Dasar Kebijakan Marketing Biro Perjalanan Wisata," *Jurnal Ilmu Komputer*, vol. 9, no. 2, hlm. 134–147, Okt 2020, doi: 10.33060/jik/2020/vol9.iss2.180.

[12] I. K. Saputri, T. Setiadi, dan L. Zahrotun, "Penerapan Text Mining Pengelompokkan Judul Kerja Praktek Menggunakan Metode K-Means Clustering dengan Cosine Similarity," *Jurnal Sarjana Teknik Informatika*, vol. 6, no. 3, hlm. 87–95, Nov 2018, [Daring]. Tersedia pada: http://journal.uad.ac.id/index.php/JSTIF

[13] I. Alfian, "Penerapan Metode K-Means Dalam Melakukan Pengelompokan Bencana Alam di Indonesia Dilakukan dengan Memanfaatkan Teknik Text Mining," *Jurnal Algoritma*, vol. 20, no. 1, hlm. 139–147, Mei 2023, doi: 10.33364/algoritma/v.20-1.1275.

[14] P. A. N. Aryanti dan I. B. M. Mahendra, "Analisis Sentimen Opini Berbahasa Indonesia Pada Sosial Media Menggunakan TF-IDF dan Support Vector Machine," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 12, no. 1, hlm. 45–52, Jul 2023, Diakses: 15 April 2025. [Daring]. Tersedia pada: https://ojs.unud.ac.id/index.php/jlk/article/view/JLK.2023.v12.i01.p06/48126

[15] S. A. Curiskis, B. Drake, T. R. Osborn, dan P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Inf. Process. Manag.*, vol. 57, no. 2, hlm. 102034, Mar 2020, doi: 10.1016/j.ipm.2019.04.002.

[16] Y. Poulakis, C. Doulkeridis, dan D. Kyriazis, "A Survey on AutoML Methods and Systems for Clustering," *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 5, hlm. 1–30, Jun 2024, doi: 10.1145/3643564.

[17] V. Mehta, M. Agarwal, dan R. K. Kaliyar, "A comprehensive and analytical review of text clustering techniques," *Int. J. Data Sci. Anal.*, vol. 18, no. 3, hlm. 239–258, Sep 2024, doi: 10.1007/s41060-024-00540-x.

[18] M. Alzate, M. Arce-Urriza, dan J. Cebollada, "Mining the text of online consumer reviews to analyze brand image and brand positioning," *Journal of Retailing and Consumer Services*, vol. 67, hlm. 102989, Jul 2022, doi: 10.1016/j.jretconser.2022.102989.

[19] M. R. Borbon dan R. A. Ebardo, "Social media discussions on educators: Selecting and appraisal of recent research using TF-IDF," *Computers and Education Open*, vol. 9, hlm. 100293, Des 2025, doi: 10.1016/j.caeo.2025.100293.

[20] Q. Wan, X. Xu, dan J. Han, "A dimensionality reduction method for large-scale group decision-making using TF-IDF feature similarity and information loss entropy," *Appl. Soft Comput.*, vol. 150, hlm. 111039, Jan 2024, doi: 10.1016/j.asoc.2023.111039.

[21] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, dan T. Yang, "Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," *Heliyon*, vol. 10, no. 16, hlm. e35945, Agu 2024, doi: 10.1016/j.heliyon.2024.e35945.

[22] M. A. Mahdi, K. M. Hosny, dan I. Elhenawy, "Scalable Clustering Algorithms for Big Data: A Review," *IEEE Access*, vol. 9, hlm. 80015–80027, 2021, doi: 10.1109/ACCESS.2021.3084057.

[23] V. Mehta, S. Bawa, dan J. Singh, "WEClustering: word embeddings based text clustering technique for large datasets," *Complex & Intelligent Systems*, vol. 7, no. 6, hlm. 3211–3224, Des 2021, doi: 10.1007/s40747-021-00512-9.

© 2026 The Authors | doi: 10.23960/pepadun.v7i2.360

[24] J. Ravi dan S. Kulkarni, "Text embedding techniques for efficient clustering of twitter data," *Evol. Intell.*, vol. 16, no. 5, hlm. 1667–1677, Okt 2023, doi: 10.1007/s12065-023-00825-3.

[25] B. Amiri dan R. Karimianghadim, "A novel text clustering model based on topic modelling and social network analysis," *Chaos Solitons Fractals*, vol. 181, hlm. 114633, Apr 2024, doi: 10.1016/j.chaos.2024.114633.

[26] M. Yuan, P. Lin, dan J. Zobel, "Document Clustering vs Topic Models: A Case Study," dalam *Australasian Document Computing Symposium*, New York, NY, USA: ACM, Des 2021, hlm. 1–8. doi: 10.1145/3503516.3503527.

[27] G. Costa dan R. Ortale, "Document Clustering Meets Topic Modeling with Word Embeddings," dalam *Proceedings of the 2020 SIAM International Conference on Data Mining*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2020, hlm. 244–252. doi: 10.1137/1.9781611976236.28.

[28] H. Ito dan B. Chakraborty, "Social Media Mining with Dynamic Clustering: A Case Study by COVID-19 Tweets," dalam *2020 11th International Conference on Awareness Science and Technology (iCAST)*, IEEE, Des 2020, hlm. 1–6. doi: 10.1109/iCAST51195.2020.9319496.

[29] D. F. Surianto dan D. F. Surianto, "Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 2, hlm. 964–972, Mar 2025, doi: 10.47709/brilliance.v4i2.5547.

[30] J. R. Trillo, E. Herrera-Viedma, J. A. Morente-Molinera, dan F. J. Cabrerizo, "A large scale group decision making system based on sentiment analysis cluster," *Information Fusion*, vol. 91, hlm. 633–643, Mar 2023, doi: 10.1016/j.inffus.2022.11.009.