Optimizing Semantic Clustering of Cultural Heritage Question-Answering Corpora Using Sentence-BERT Embeddings and PCA-Enhanced K-Means

Nala Widyadhana[1], Nur Cahyo Wibowo[2], Tri Lathif Mardi Suryanto[3]
[1,2,3] Department of Information Systems, Faculty of Computer Science,
Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia

E-mail: trilathif.si@upnjatim.ac.id

Abstract. This study examines semantic text clustering using all-MiniLM-L6-v2 sentence embeddings and K-Means on a Dewi Durga question-answering corpus from Indian, Javanese, and Balinese cultural contexts. The dataset contains 1,620 Context-Question-Answer entries extracted from Chapters 1-22. Text preprocessing included structural checking, missing-value inspection, duplicate detection, case folding, non-alphanumeric character removal, and whitespace normalization. Each context was transformed into a 384-dimensional dense embedding vector. The optimal cluster number was evaluated using Auto K across K values from 2 to 10 with Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, while Manual K = 5 was used as a comparative setting for more detailed thematic interpretation. Six embedding transformation scenarios were tested in both modes. The results show that Auto_K_S5, combining normalization and PCA with 50 components, achieved the strongest internal validation performance with a Silhouette Score of 0.098899, Davies-Bouldin Index of 2.912914, and Calinski-Harabasz Index of 186.476974. Manual_K5_S3 produced more granular themes related to ritual, mythology, history, archaeology, and religious narrative.

Keywords: K-Means; Sentence-BERT MiniLM; PCA; Sentence Clustering; Internal Validation Metrics.

1. Introduction

The rapid growth of textual data has increased the need for computational methods that can organize large-scale text collections into meaningful thematic structures. Text clustering is one of the main unsupervised learning approaches used to group documents based on similarity without requiring predefined labels. In natural language processing, clustering has been widely used for document organization, topic discovery, short-text analysis, and semantic exploration of textual corpora [1], [2]. Unlike supervised classification, clustering is particularly useful when the thematic structure of a corpus is not known in advance. Therefore, it is relevant for exploratory studies involving question-answering datasets and cultural text corpora.

Early document clustering approaches generally relied on sparse text representations, such as bag-of-words and term-weighting schemes. Although these methods are computationally efficient, they often have limitations in capturing contextual meaning, semantic similarity, and relationships between conceptually related expressions [3]. The development of distributed word representations, such as Word2Vec and GloVe, introduced vector-based semantic representation by mapping words into continuous vector spaces [4], [5]. However, word-level representations are still limited when the unit of analysis is a sentence, paragraph, or question-answer pair. This limitation has encouraged the use of sentence-level embedding models that can represent a text segment as a dense semantic vector.

Transformer-based language models have significantly advanced text representation by using attention mechanisms to capture contextual relationships between words [6]. Models such as BERT further improved contextual language understanding through deep bidirectional pre-training [7]. Building on this development, Sentence-BERT introduced a siamese network architecture to generate sentence embeddings suitable for semantic similarity and clustering tasks [8]. Recent studies have shown that transformer-based embeddings can improve short-text clustering and semantic grouping because

they encode contextual information more effectively than traditional sparse representations [9], [10], [11]. These developments indicate that sentence embeddings are suitable for clustering corpora that contain narrative, historical, and interpretive content.

Despite their advantages, transformer-based embeddings still require careful experimental design when used for clustering. One major issue is the selection of the number of clusters in algorithms such as K-Means. K-Means remains widely used because of its simplicity, scalability, and effectiveness in partition-based clustering, but its performance is strongly influenced by the selected value of K [12]. If K is too small, the resulting clusters may be overly general; if K is too large, the clusters may become fragmented and difficult to interpret. For this reason, internal validation metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index are commonly used to evaluate cluster quality [13], [14], [15]. However, the value of K that performs best numerically does not always produce the most interpretable thematic structure, especially in exploratory text analysis.

Another important factor in text clustering is the transformation of embedding vectors before clustering. Dense embeddings may contain redundant dimensions or distributional patterns that affect cluster separation. Dimensionality reduction techniques can be used to reduce noise and improve the structure of the embedding space. Previous studies on text embeddings and clustering have shown that embedding selection, representation quality, and post-processing strategies can influence clustering performance [16], [17], [18]. Principal Component Analysis (PCA) is frequently applied to reduce dimensionality while preserving major variance in the data. In addition, vector normalization may affect K-Means clustering because the algorithm is sensitive to distance and scale. Therefore, evaluating normalization and PCA-based transformation is necessary to determine which configuration provides better internal validation performance.

The use of natural language processing in cultural heritage and digital humanities has also become increasingly important. NLP methods can support the organization, enrichment, and interpretation of cultural heritage data, including artefact descriptions, historical texts, and chronological corpora [19], [20]. In the context of cultural studies, computational text analysis can help reveal patterns that may not be immediately visible through manual reading alone. The corpus used in this study is related to the study of Dewi Durga across India, Java, and Bali, which represents a culturally rich and semantically complex domain [21], [22]. Such a corpus requires a method that can identify semantic similarity while still allowing thematic interpretation.

This study investigates text clustering on a question-answering corpus related to Dewi Durga using transformer-based sentence embeddings and K-Means clustering. The dataset consists of 1,620 text entries constructed from chapters 1–22 and includes context, question, and answer fields. Each text entry is represented using the all-MiniLM-L6-v2 sentence embedding model with a 384-dimensional vector representation. The clustering process is conducted using two cluster-number strategies: an automatic K selection based on internal validation metrics and a manual K = 5 setting for more granular thematic interpretation. Six embedding transformation scenarios are evaluated under both strategies, resulting in twelve experimental configurations.

The main contribution of this study is the evaluation of transformer-based sentence embeddings for clustering a cultural question-answering corpus. Specifically, this study compares Auto K and Manual K = 5 strategies, evaluates the effect of normalization and PCA-based dimensionality reduction, and analyzes the resulting clusters using internal validation metrics and dominant thematic terms. This study also highlights the importance of balancing quantitative validation and interpretability in semantic clustering. The findings are expected to provide methodological insight into the use of sentence embeddings and K-Means clustering for exploratory analysis in cultural and digital humanities corpora.

2. Methodology

2.1 Research Design

This study employed an experimental text clustering approach to evaluate the performance of transformer-based sentence embeddings combined with K-Means clustering. The experiment was designed to compare two cluster-number strategies: an automatic K selection strategy and a manual K = 5 strategy. The automatic K strategy was used to identify the best number of clusters based on internal validation metrics, while the manual K = 5 strategy was used to obtain a more granular thematic

interpretation of the corpus. This design allows the study to examine both quantitative clustering quality and qualitative interpretability.

The overall workflow consisted of six main stages: dataset construction, text preprocessing, semantic embedding representation, cluster-number selection, clustering experiment, and internal validation. The clustering process used one transformer-based sentence embedding model, namely all-MiniLM-L6-v2. The model was selected because sentence-level embeddings are suitable for representing semantic similarity in text clustering tasks [8], [9], [10]. The clustering algorithm used in this study was K-Means due to its simplicity, scalability, and common use in partition-based clustering [12].

2.2 Research Method / Framework

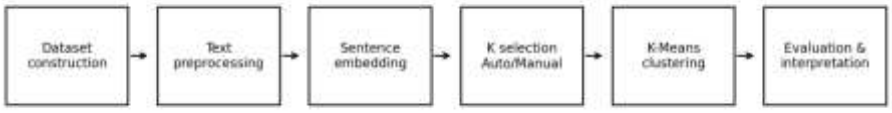


Figure 1. Research Workflow for Semantic Clustering of the Dewi Durga QA Corpus

Figure 1 presents the research framework used to transform the Dewi Durga QA corpus into semantic clusters. The procedure begins with dataset preparation and structural checking, followed by moderate text preprocessing, sentence embedding generation using all-MiniLM-L6-v2, embedding transformation through normalization and PCA, K-Means clustering, and internal validation using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

The workflow also applies a dual-track clustering strategy. Auto K evaluates K values from 2 to 10 to identify the strongest metric-based partition, while Manual K = 5 is retained to obtain a more detailed thematic interpretation. The six embedding scenarios are evaluated in both tracks so that the contribution of normalization and PCA can be compared systematically.

2.3 Dataset

Table 1. Example Structure of the Question Answering (QA) Dataset

Sample of the first 3 records:

No.	Context ID	Context	Question	Answer
0	Bab1_001	Perjalanan Dewi Durga Mahisasuramardini: India, Jawa, dan Bali.	How is the concept of Dewi Durga's journey understood in Indian, Javanese, and Balinese cultural studies?	The journey refers to the diffusion, transformation, and adaptation of Durga from India to Nusantara, including shifts in religious meaning, symbols, and cultural function.
1	Bab1_002	Durga in India is initially depicted as a beautiful warrior goddess with many arms holding divine weapons.	What is the initial depiction of Dewi Durga in India?	Durga is described as a warrior and protector goddess created to defeat Mahisasura, the buffalo-headed demon king.
2	Bab1_003	Durga is described in Devi Mahatmya, a section of the Markandeya Purana, and later known as Mahisasuramardini.	Where is Durga's form described and what does Mahisasuramardini mean?	The form appears in Devi Mahatmya; Mahisasuramardini means Durga who successfully defeats Mahisasura.

The dataset used in this study is a question-answering (QA) corpus related to the study of Dewi Durga across Indian, Javanese, and Balinese contexts. The dataset consists of 1,620 entries derived from chapters 1–22 and contains four main attributes: context_id, context, question, and answer. The context_id attribute functions as an identifier for each textual context, while the context attribute contains the main textual information. Meanwhile, the question and answer attributes represent paired questions and responses generated based on the available context.

This corpus was constructed to represent cultural, historical, mythological, and interpretive information related to Dewi Durga. Therefore, the dataset is relevant for semantic clustering because it contains narrative and thematic information that cannot be fully captured through keyword matching alone. In the context of cultural heritage and digital humanities, natural language processing can support the enrichment, exploration, and organization of textual knowledge [19], [20]. The corpus is also

associated with previous work on Durga-related cultural heritage question-answering data [21] and textual sources discussing Dewi Durga in Indian, Javanese, and Balinese traditions [22].

In this study, the context field was used as the main semantic unit for clustering because it contains the most complete textual representation of each topic. The text was processed to generate sentence embeddings, allowing the clustering model to group entries based on semantic similarity rather than surface-level word overlap. This approach is consistent with recent text clustering studies that utilize transformer-based embeddings and semantic representation to improve clustering quality in short and domain-specific texts [9], [10], [11], [12].

2.4 Text Preprocessing

Before embedding generation, the dataset was inspected to ensure structural consistency. The preprocessing stage included checking the number of rows and columns, identifying missing values, detecting duplicated records, and examining text length distribution. The text was then standardized through case folding, removal of non-alphanumeric characters, and whitespace normalization, which are commonly applied in text mining and text clustering workflows [1], [2], [15]. Aggressive stemming was not applied because the corpus contains cultural, historical, and religious terms whose original forms may carry semantic meaning, particularly in cultural heritage and digital humanities datasets [19], [21], [22], [20].

This preprocessing strategy was intended to reduce textual noise while preserving important domain-specific expressions. Since the clustering task relied on semantic embeddings rather than sparse frequency-based representation, the preprocessing process was kept moderate. Semantic embedding models are designed to capture contextual and sentence-level meaning, so excessive normalization may reduce meaningful lexical and contextual variation [23], [24], [25], [26], [27]. In addition, embedding-based topic modeling and text clustering approaches emphasize the preservation of semantic structure in textual data, making moderate preprocessing more suitable than aggressive stemming for domain-specific corpora [28], [29], [30], [31].

2.5 Sentence Embedding Representation

Each text entry was represented using the all-MiniLM-L6-v2 sentence embedding model. This model produces dense vector representations with 384 dimensions. Transformer-based sentence embeddings were used because they can encode contextual and semantic relationships more effectively than traditional sparse representations such as bag-of-words or TF-IDF. The development of transformer architecture, BERT, Sentence-BERT, and other sentence embedding models has shown that contextual representations are effective for capturing semantic information at the sentence level [6], [7], [8], [23], [24]. Recent studies on text embeddings and transformer-based clustering also show that sentence-level representations can improve semantic similarity measurement and short-text clustering performance [9], [10], [11], [32], [32], [33], [34], [16].

The embedding vectors were used as the numerical input for the clustering process. Since dense embeddings may contain redundant dimensions or distributional patterns that influence clustering performance, several embedding transformation scenarios were evaluated. These transformations included vector normalization and dimensionality reduction using Principal Component Analysis (PCA). Previous studies have shown that embedding selection, representation quality, and post-processing strategies can affect clustering performance and semantic representation quality [17], [18], [28], [25], [26], [27].

In addition, recent developments in dense representation and transformer-based architectures show that contextual embeddings are widely used in semantic representation and text retrieval tasks. Dense passage retrieval and unsupervised dense retrieval demonstrate the effectiveness of vector-based semantic representation for matching textual meaning beyond surface-level keyword overlap [35], [36]. Meanwhile, long-sequence transformer models such as LongT5, BigBird, and Longformer were developed to address the limitations of standard transformer architectures in processing longer textual inputs [37], [38], [39]. These studies support the use of embedding-based representation as a suitable approach for processing narrative and domain-specific text data before clustering.

2.6 Cluster-Number Selection

The optimal number of clusters in this framework was algorithmically evaluated using an automatic K selection procedure across a continuous constraint interval of $K \in [2], [10]$. This evaluation was conducted because centroid-based clustering methods such as K-Means require the number of clusters to be determined before the clustering process is executed [12], [2], [3]. For each iteration, the partitioning output generated by the clustering process was evaluated using internal validation metrics, including the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) [13], [14], [15]. Additionally, inertia values were observed as part of the traditional Elbow Method analysis to monitor the reduction of within-cluster distortion [12], [15].

Mathematically, the Silhouette Score measures how well each data point fits within its assigned cluster compared to neighboring clusters [13], whereas the Davies-Bouldin Index evaluates the ratio between within-cluster dispersion and between-cluster separation [14]. Meanwhile, the Calinski-Harabasz Index assesses the ratio between inter-cluster dispersion and intra-cluster dispersion. Based on the automated tracking results executed over the dense vector space, $K = 2$ obtained the strongest numerical performance according to the highest Silhouette Score and the optimized Calinski-Harabasz Index. Therefore, $K = 2$ was selected as the automated operational anchor, referred to as Auto K.

Although $K = 2$ produced the best internal validation performance, this study also established a constrained Manual $K = 5$ configuration as a comparative baseline. This design choice is methodologically important because the mathematically optimal number of clusters does not always align with text interpretability, especially in complex historical, cultural, and digital humanities corpora [19], [21], [22], [20]. In addition to centroid-based clustering methods such as K-Means, density-based clustering approaches can also be used to identify cluster structures based on data distribution. HDBSCAN is a hierarchical density-based clustering method that can detect clusters with varying densities and identify noise in the data [40]. However, this study maintained K-Means as the main clustering algorithm because the research design required a controlled comparison between Auto $K = 2$ and Manual $K = 5$ configurations.

In neural clustering and semantic topic modeling literature, alternative approaches such as topic modeling in embedding spaces, Top2Vec, and BERTopic are often used to identify latent topic structures from textual data [29], [11], [31]. However, because transformer-based embeddings, instruction-finetuned representations, and SBERT optimization techniques can produce dense semantic vector spaces, implementing a manual $K = 5$ configuration provides a controlled taxonomic comparison [10], [32], [8], [26]. This allows the study to capture more granular micro-thematic layers, such as localized Javanese adaptations and specific architectural classifications, that may be compressed into broader macro-thematic groups under the Auto $K = 2$ configuration.

2.7 Experimental Scenarios

To evaluate the parametric impact of vector space modifications, six embedding transformation scenarios were systematically designed and cross-examined under both the algorithmic Auto K and constrained Manual $K = 5$ operational tracks [10], [12]. Consequently, the experimental matrix comprised twelve distinct pipeline configurations designed to monitor partitioning volatility across varying dense mathematical structures [1]. These structured setups included the raw, un-optimized embedding representation (S1), length-normalized embeddings (S2), a 50-dimensional Principal Component Analysis (PCA) projection (S3), and a 100-dimensional PCA projection (S4) [8], [15]. To evaluate cascading linear adjustments, the framework further integrated combinations of normalized embeddings followed by a 50-component PCA mapping (S5), and normalized embeddings followed by a 100-component PCA compression (S6) [22], [17].

The experimental scenarios are summarized as follows:

Table 2. Embedding Transformation Scenarios

Scenario	Embedding Transformation	Normalization	PCA Dimension
S1	Original embedding	No	No PCA
S2	Normalized embedding	Yes	No PCA
S3	PCA transformation	No	50

Scenario	Embedding Transformation	Normalization	PCA Dimension
S4	PCA transformation	No	100
S5	Normalization + PCA	Yes	50
S6	Normalization + PCA	Yes	100

Each scenario was evaluated under two cluster-number settings: Auto K = 2 and Manual K = 5. Therefore, the total number of experiments was calculated as follows:

Table 3. Experimental Configuration Summary

Mode	Number of Scenarios	K Value	Total Experiments
Auto K	6	2	6
Manual K = 5	6	5	6
Total	12	2 and 5	12

2.8 Clustering Algorithm

K-Means clustering was applied to the embedding vectors generated in each experimental scenario. The algorithm partitions data into K clusters by minimizing the distance between data points and their corresponding cluster centroids. K-Means was chosen because it is widely used for document and text clustering tasks and remains effective for evaluating partition-based grouping in embedding spaces [3], [41]. The implementation was conducted using the Scikit-learn library in Python [28], and the experiment was executed in Google Colaboratory [42].

For each experimental configuration, K-Means produced a cluster label for every text entry. The resulting labels were then used for internal validation, cluster distribution analysis, dominant-word extraction, and thematic interpretation. The same evaluation procedure was applied to all configurations to ensure comparability between Auto K and Manual K = 5.

2.9 Evaluation Metrics

The clustering results were evaluated using three internal validation metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. These metrics were selected because the dataset does not contain predefined class labels, making external validation unsuitable. Internal validation is appropriate for unsupervised clustering because it evaluates compactness and separation based on the structure of the data itself [43], [15].

A higher Silhouette Score indicates better separation between clusters. A lower Davies-Bouldin Index indicates better cluster compactness and separation. A higher Calinski-Harabasz Index indicates a better ratio between inter-cluster and intra-cluster dispersion. The best configuration was selected by considering the overall performance across these metrics, while also taking into account the interpretability of the resulting clusters.

2.10 Cluster Interpretation

After the best configurations were identified, the clusters were interpreted using distribution analysis and dominant-word extraction. The number of data points in each cluster was calculated to examine the balance of cluster membership. Dominant words were extracted from each cluster to identify the main semantic tendency of the grouped texts, while representative examples were selected to support thematic interpretation [2], [3].

The Auto K = 2 result was interpreted as the configuration that provides the strongest internal validation performance, while the Manual K = 5 result was interpreted as a more granular configuration for thematic analysis. This dual interpretation was used to balance quantitative validation and qualitative readability, especially in cultural and digital humanities corpora where meaningful interpretation is not always fully reflected by numerical clustering metrics alone [20].

3. Result and Discussion

3.1 Dataset Description / Exploratory Data Analysis

Table 4. Results of Missing Value and Data Duplication Checks

Missing Value per Kolom			
No	**Kolom**	**Jumlah Missing**	**Persentase Missing**
0	context_id	0	0.0
1	context	0	0.0
2	question	0	0.0
3	answer	0	0.0

Based on the exploratory data analysis (EDA), the dataset utilized in this study exhibits no missing values across all structural attributes, specifically context_id, context, question, and answer. Furthermore, comprehensive data deduplication checks confirmed the absolute absence of duplicate records, thereby establishing high data integrity and cleanliness for subsequent vector representation and semantic clustering workflows [21], [15]. These preprocessing assessment results demonstrate that the corpus maintains optimal structural consistency [21]. Consequently, ad-hoc data imputation strategies or custom filtering operations for duplicate removal were deemed unnecessary, allowing the raw text structures to be directly fed into the subsequent embedding pipeline [15].

Table 5. Text Length Statistics of the Dataset

Text Length Statistics		
	Character Count	**Word Count**
count	1620.000000	1620.000000
mean	551.298148	73.310494
std	231.581014	30.332203
min	101.000000	15.000000
25%	335.000000	45.000000
50%	585.000000	78.000000
75%	738.250000	98.000000
max	1126.000000	149.000000

Statistical analysis of the text length distribution reveals that the dataset possesses an average character count of 551 characters and an average word count of 73 words per entry. The text length ranges from a minimum of 15 words to a maximum of 149 words, corresponding to a character-level boundary spanning from 101 to 1,126 characters. This pronounced variance in text length indicates a high degree of textual complexity and semantic heterogeneity across the corpus. In the context of short-text mining, such structural variability often presents a challenge for traditional lexical matching [9], [1]. Consequently, utilizing advanced Sentence Transformer architectures is imperative to map these diverse text lengths into a unified, high-dimensional vector space, thereby capturing robust semantic nuances and deep contextual representations that conventional keyword-based clustering techniques fail to capture [16] [8].

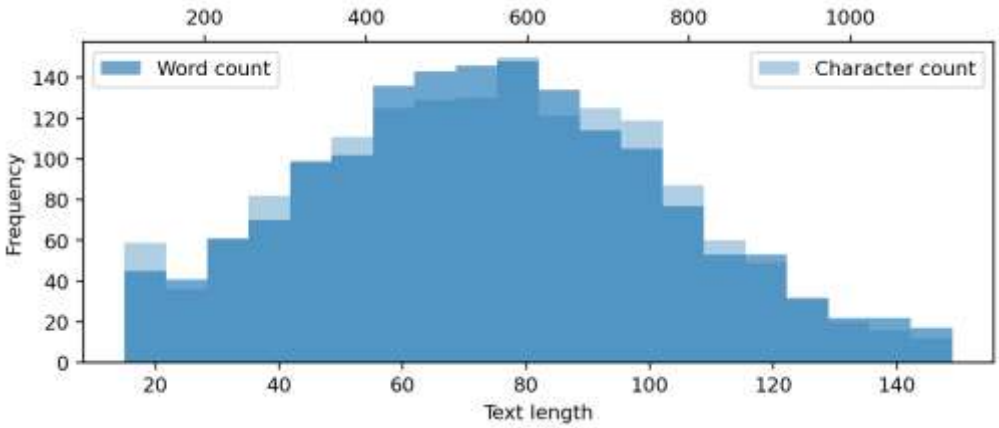


Figure 2. Histogram of Word Count and Character Count Distributions

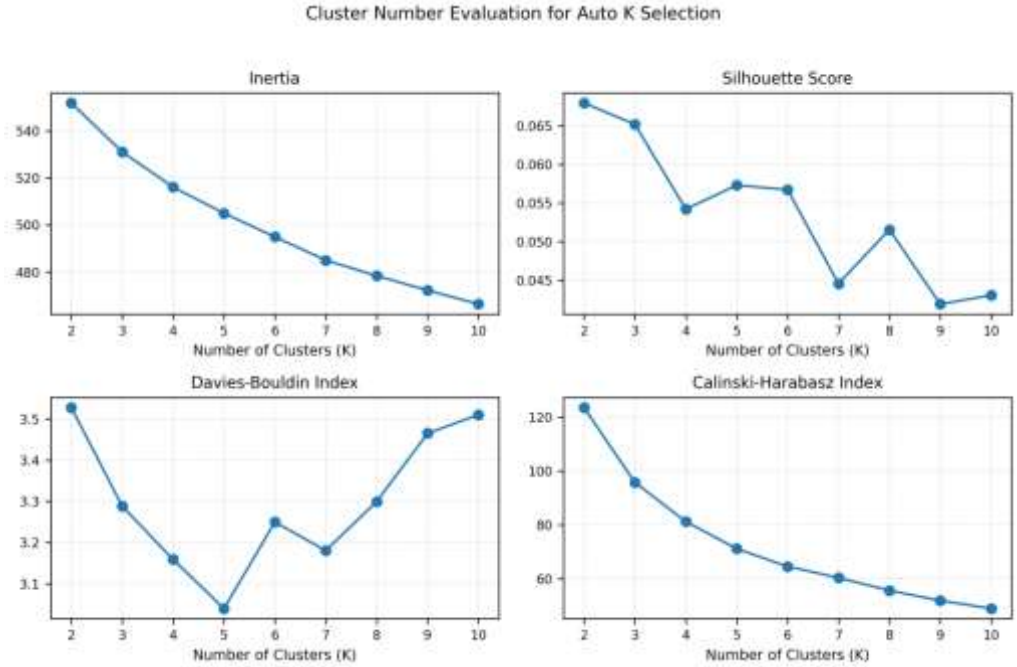
Based on the histograms illustrating the distributions of word and character counts, it is evident that the majority of the data falls within the mid-range text length interval, although a minor tail of longer text entries persists. This empirical distribution indicates that the corpus exhibits a moderately skewed, non-homogeneous text length distribution. Such structural variation and lack of absolute homogeneity strongly justify the application of dense text embedding and clustering pipelines [11], [2]. By leveraging these representation methods, the model can effectively capture fine-grained semantic similarities and map the non-uniform text structures into a continuous vector space where relative semantic proximity governs the data grouping process [2], [42].

3.2 Cluster-Number Selection Result

Table 6. Evaluation Table for Cluster Number Determination

Evaluation of Cluster Number (K)					
	Cluster_Count	**inertia**	**silhouette_score**	**davies_bouldin_index**	**calinski_harabasz_index**
0	2	551.843201	0.067935	3.527839	123.557472
1	3	531.101746	0.065190	3.288378	95.727280
2	4	516.143127	0.054209	3.159223	81.238258
3	5	505.017792	0.057323	3.040013	71.126869
4	6	494.988023	0.056724	3.249949	64.561035
5	7	485.032288	0.044594	3.180515	60.387852
6	8	478.438385	0.051562	3.299882	55.615524
7	9	472.250854	0.041938	3.465075	51.909012
8	10	466.394104	0.043080	3.510129	48.938168

Description

Description	Value
Optimal Auto K based on Silhouette 	2
Compared Manual K	5

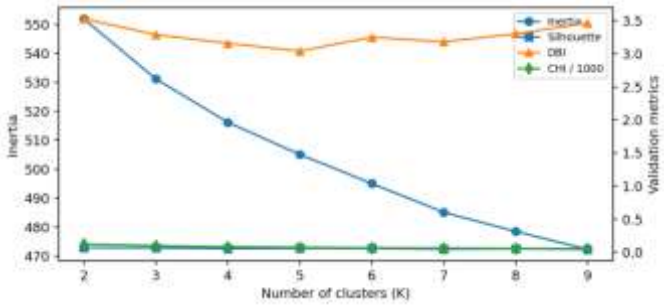


Figure 3. Comparison of Cluster Number (K) Determination Methods

Based on the Silhouette Score analysis, the highest metric value is achieved at K = 2 with a score of 0.067935. This mathematical peak indicates that the cluster separation at K = 2 yields the most optimal mathematical partition quality compared to other configurations [13]. This finding is further validated by the Calinski-Harabasz Index (CHI), which also reaches its maximum value at K = 2 (123.557472), implying strong inter-cluster segregation and relatively minimal intra-cluster variance.

Although these internal evaluation metrics conclusively identify K = 2 as the best automated partition configuration (Auto K), this study introduces a manual configuration of K = 5 as an empirical baseline comparison. Setting K = 5 is intended to extract more granular, micro-thematic partitions that offer higher domain interpretability, allowing a deeper qualitative extraction of the semantic features inherent within the corpus [44], [43]. Consequently, this research formalizes a dual-track strategy for cluster size determination: an algorithmic tracking via Auto K to maximize objective spatial validation, and a Manual K = 5 configuration to fulfill the analytical requirements of interpretive qualitative analysis [44], [34].

3.3 Evaluation and Validation Results

Table 7. Performance Ranking for All Clustering Experiments

Ranking	Experiment ID	Mode K	Skenario	Model	n_clusters	Normalize	PCA Dim	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
1	Auto_K_S5	Auto K	S5	all-MiniLM-L6-v2	2	True	50	0.098899	2.912914	186.476974
2	Auto_K_S3	Auto K	S3	all-MiniLM-L6-v2	2	False	50	0.093703	2.964195	173.714447
3	Manual_K5_S3	Manual K = 5	S3	all-MiniLM-L6-v2	5	False	50	0.085375	2.564335	103.546387
4	Auto_K_S6	Auto K	S6	all-MiniLM-L6-v2	2	True	100	0.083093	3.205876	153.601868
5	Auto_K_S4	Auto K	S4	all-MiniLM-L6-v2	2	False	100	0.078206	3.272435	143.185471
6	Manual_K5_S5	Manual K = 5	S5	all-MiniLM-L6-v2	5	True	50	0.076219	2.978697	104.785767
7	Manual_K5_S4	Manual K = 5	S4	all-MiniLM-L6-v2	5	False	100	0.071544	2.799714	83.526405
8	Auto_K_S2	Auto K	S2	all-MiniLM-L6-v2	2	True	No PCA	0.067935	3.527839	123.557472
9	Auto_K_S1	Auto K	S1	all-MiniLM-L6-v2	2	False	No PCA	0.067935	3.527839	123.557472
10	Manual_K5_S6	Manual K = 5	S6	all-MiniLM-L6-v2	5	True	100	0.061009	3.349425	85.218124
11	Manual_K5_S1	Manual K = 5	S1	all-MiniLM-L6-v2	5	False	No PCA	0.057323	3.040013	71.126869
12	Manual_K5_S2	Manual K = 5	S2	all-MiniLM-L6-v2	5	True	No PCA	0.057323	3.040013	71.126869

Based on the comprehensive performance ranking of all experimental scenarios, this study evaluates the interaction of multiple preprocessing and dimensionality reduction pipelines using the all-MiniLM-L6-v2 Sentence Transformer model under both Auto K and Manual $K = 5$approaches [10], [16]. The partitioning quality was rigorously validated using three internal clustering metrics: the Silhouette Score [13], the Davies-Bouldin Index (DBI) [14], and the Calinski-Harabasz Index (CHI). The empirical rankings demonstrate that the optimal global performance is achieved by the Auto_K_S5 configuration, which pairs data normalization with a 50-dimensional Principal Component Analysis (PCA) reduction. This specific setup yielded a Silhouette Score of 0.098899, a Davies-Bouldin Index of 2.912914, and a Calinski-Harabasz Index of 186.476974. These statistical outputs indicate that the Auto_K_S5 scenario generates the most structurally sound internal validation performance among all evaluated combinations, particularly showcasing dominant margins in cluster cohesion and separation as monitored by the Silhouette Score and CHI [10], [15].

In summary, Auto_K_S5 is the best configuration for metric-based validation, whereas Manual_K5_S3 is the strongest Manual K = 5 configuration for interpretive analysis. This confirms that Auto K and Manual K = 5 serve complementary purposes: the first prioritizes compact and separated clusters, while the second supports richer thematic granularity.

3.4 Clustering Visualization

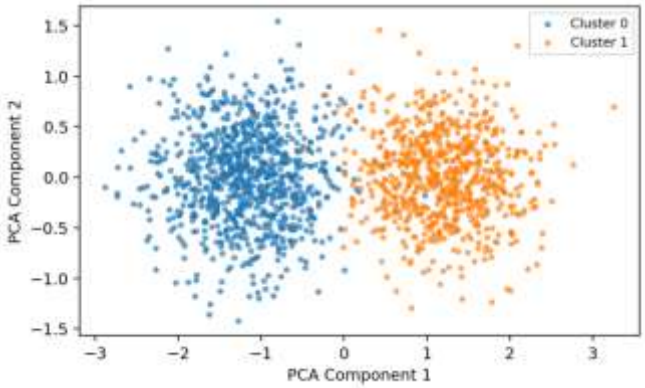


Figure 4. Clustering Visualization of the Auto K Approach

Based on the spatial visualization of the clustering configurations, the 2D Principal Component Analysis (PCA) projection reveals that the Auto K and Manual K = 5 approaches generate distinctly different distribution patterns. The high-dimensional sentence embeddings are reduced via PCA to map the semantic vectors into a continuous two-dimensional subspace, thereby facilitating a comprehensive visual analysis of cluster cohesion [45], [42]. Each discrete color coordinate on the scatter plots represents an isolated cluster partition mapped by the K-Means algorithm [2], [15].

In the Auto K visualization pipeline, where internal validation metrics optimize the partition size at K = 2, the distribution exhibits two primary macro-clusters separated by a distinct margin along the first principal axis (PCA Component 1). The vast majority of the data entries assigned to the first cluster are tightly localized on the left hemisphere of the plot, whereas the second cluster populates the right hemisphere. This structural pattern demonstrates that the embedding pipeline paired with K-Means effectively discriminates the corpus based on underlying semantic proximity, yielding two distinct primary groups with highly divergent contextual characteristics [2], [10].

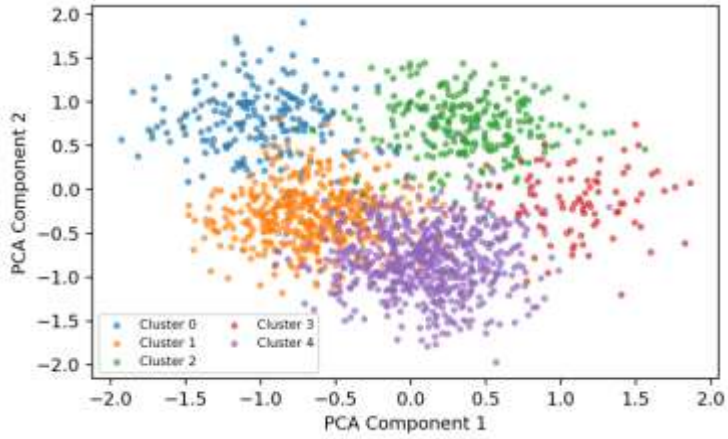


Figure 5. Clustering Visualization of the Manual K = 5 Approach

Meanwhile, in the Manual K = 5 visualization pipeline, the corpus is partitioned into five discrete clusters, yielding a more granular distribution topology compared to the automated Auto configuration. Several cluster boundaries appear adjacent, and partial semantic overlaps are observed, particularly localized within the central core of the PCA scatter plot. This spatial layout indicates that while increasing the partition density provides higher resolution, the inter-cluster separation boundaries become less distinct compared to the clean segregation achieved at K = 2 [12], [2]. Nonetheless, establishing K = 5 offers a significant advantage for qualitative cluster interpretation, as it uncovers niche contextual characteristics and micro-thematic patterns unique to each sub-group [44], [11].

Broadly speaking, the visual data structures demonstrate that the Auto approach provides superior spatial cluster segregation—consistent with the objective internal validation metrics—whereas the Manual K = 5 configuration delivers a more detailed and interpretatively useful semantic taxonomy [2], [13]. Consequently, both analytical paths are formally integrated into this study to comprehensively

benchmark empirical space partitioning against qualitative text organization during the subsequent semantic extraction phase [19], [20].

3.5 Cluster Characteristics

Table 8. Summary Table of Cluster Characteristics

Approach	Cluster	Data Count	Percentage (%)	Dominant Keywords	Thematic Interpretation	Representative Sample
Auto K	0	902	55.68	dewi, durga, spiritual, bahwa, memiliki	Transformasi citra Durga dan pengalaman spiritual	Penggambaran Durga Mahisasuramardini di Nusantara, khususnya di Jawa dan Bali, mengalami evolusi perubahan radikal..
Auto K	1	718	44.32	durga, dewi, kekuatan, jawa, pemujaan	Genealogi mitologis dan penyebaran pemujaan Durga	Istilah "perjalanan" Dewi Durga merujuk pada proses penyebaran, transformasi, dan adaptasi figur Durga dari India h..
Manual K = 5	0	197	12.16	ritual, durga, spiritual, dewi, kekuatan	Ritual, yatra, dan pengalaman spiritual	Tujuan utama melakukan jatra dan yatra adalah untuk meminta izin dan anugerah dari Durga. Hal ini dilakukan supaya..
Manual K = 5	1	458	28.27	durga, dewi, kekuatan, jawa, dewa	Mitos asal-usul dan fungsi ilahiah Durga	Istilah "perjalanan" Dewi Durga merujuk pada proses penyebaran, transformasi, dan adaptasi figur Durga dari India h..
Manual K = 5	2	251	15.49	durga, dewi, jawa, bali, kekuatan	Perkembangan historis Durga di Jawa dan Bali	Penggambaran Durga sebagai Mahisasuramardini menjadi pengaruh utama yang menyebar ke Indonesia. Penggambaran ini me..
Manual K = 5	3	89	5.49	candi, durga, arca, jawa, relief	Artefak, arca, relief, dan bukti arkeologis	Sama halnya dengan riset di India, penulis mengadakan survey ke museum, candi-candi di Jawa Tengah dan Jawa Timur,..
Manual K = 5	4	625	38.58	dewi, bahwa, memiliki, durga, spiritual	Narasi religius dan kekuatan supranatural	Penggambaran Durga Mahisasuramardini di Nusantara, khususnya di Jawa dan Bali, mengalami evolusi perubahan radikal..

Based on the cluster data volume distribution charts, the Auto K approach yields two primary clusters characterized by a relatively balanced allocation of data entries. Cluster 0 contains a slightly larger volume of data points compared to Cluster 1; however, this variance is statistically marginal, indicating that the embedding and partition workflow divides the vector space with optimal structural stability [10], [12]. These empirical results confirm that data separation under the automated Auto K approach tends to be more parsimonious and mathematically robust according to internal validation metrics [3], [13].

Conversely, under the Manual K = 5 approach, the distribution of data counts across the clusters exhibits higher volatility and variance. Cluster 4 accumulates the highest concentration of text records, whereas Cluster 3 registers the lowest sample size. This uneven distribution demonstrates that expanding the target partition scale generates a higher-resolution data taxonomy; consequently, specific sub-groups encapsulate narrower contextual attributes and highly specialized micro-thematic properties compared to others [9], [11].

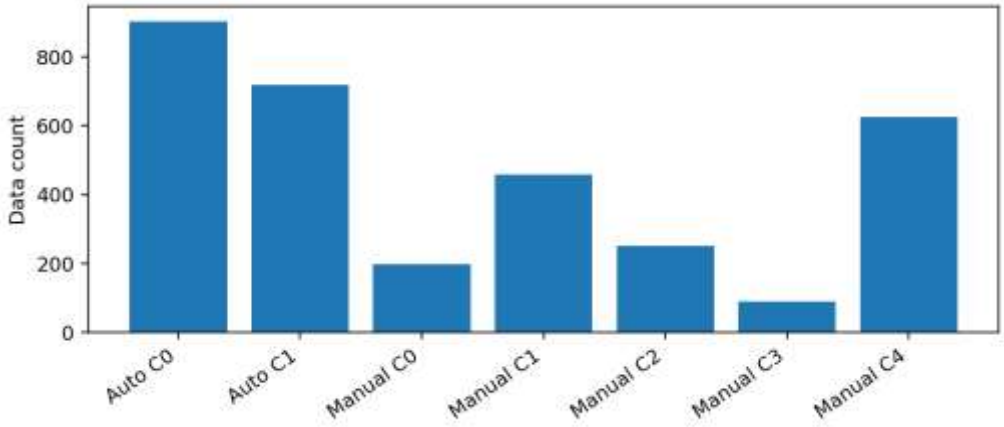


Figure 6. Data Volume Distribution per Cluster Across Auto and Manual K = 5 Approaches

Based on the cluster characteristic summary profiles, each partition exhibits distinct semantic properties as monitored by high-frequency keywords and representative textual samples. Under the Auto K approach, the clusters are predominantly defined by lexical tokens such as "dewi" (goddess), "durga",

"spiritual", "kekuatan" (power), and "pemujaan" (worship). This distribution indicates that the embedding space successfully partitions into two primary themes that capture broad cultural and theological discourses surrounding the spirituality of Goddess Durga Mahisasuramardini [21], [22].

Conversely, the Manual K = 5 configuration delivers a more localized and detailed semantic resolution. The resulting cluster topologies are governed by specialized keywords, including "ritual", "spiritual", "candi" (temple), "arca" (statue), "jawa" (Java), and "kekuatan". This shifting keyword pattern demonstrates that expanding the target partition size enables the extraction of highly granular, micro-thematic layers compared to the broader categorization of the Auto approach [11], [34].

Collectively, the experimental findings confirm that while the automated Auto pipeline yields a parsimonious, mathematically optimized, and structurally balanced partition space [12], [13], the Manual K = 5 configuration provides an enriched semantic taxonomy that enhances the deep qualitative analysis and textual interpretation necessary for domain-specific discoveries [19], [20].

3.6 Insight Analysis

Table 9. Cluster Insight Summary Based on Dominant Words

No.	Mode	Cluster	Top 3 Kata Dominan + Frekuensi
1	Auto K	0	dewi (473), durga (371), spiritual (365)
2	Auto K	1	durga (1422), dewi (868), kekuatan (289)
3	Manual K = 5	0	ritual (272), durga (136), spiritual (136)
4	Manual K = 5	1	durga (997), dewi (730), kekuatan (209)
5	Manual K = 5	2	durga (314), dewi (114), jawa (95)
6	Manual K = 5	3	candi (166), durga (100), arca (83)
7	Manual K = 5	4	dewi (326), bahwa (283), memiliki (258)

Based on the dominant keyword summary presented in Table 9, each cluster demonstrates a distinct thematic inclination. Under the Auto K approach, Cluster 0 is heavily dominated by the tokens "dewi" (473), "durga" (371), and "spiritual" (365), whereas Cluster 1 is governed by "durga" (1,422), "dewi" (868), and "kekuatan" (289). This frequency distribution implies that the automated Auto K pipeline partitions the corpus into two macro-thematic groups that still center tightly around the core subject of Goddess Durga, specifically encapsulating her spiritual dimensions and divine representations of power [21], [22].

Conversely, the Manual K = 5 configuration delivers a more localized and granular thematic taxonomy. Cluster 0 is characterized by "ritual", "durga", and "spiritual", mapping explicitly onto ritualistic and spiritual practices. Cluster 1 highlights "durga", "dewi", and "kekuatan", signifying discussions on the iconographical figure of Durga as a symbol of power. Cluster 2 distinguishes itself through the keyword "jawa", representing the historical and regional evolution of Durga within the Javanese context [14]. Cluster 3 is dominated by architectural descriptors such as "candi" (temple) and "arca" (statue), which isolates an archeological theme focusing on physical cultural heritage remains [19]. Lastly, Cluster 4 contains systemic narrative particles like "bahwa" and "memiliki", encapsulating a general descriptive narrative regarding the deity. Consequently, while the Auto K approach remains mathematically superior according to objective internal validation metrics [12], [13], the Manual K = 5 partition structure provides higher practical utility for micro-thematic text exploration and granular qualitative discovery [11], [20].

3.7 Ablation Study

Table 10. Ablation Study Results Across All Experimental Scenarios

No.	Mode K	Skenario	Konfigurasi	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Delta Silhouette	Delta DBI
1	Auto K	S1	S1 | Normalize=False | PCA=No PCA	0.067935	3.527839	123.557472	0.000000	0.000000e+00
2	Auto K	S2	S2 | Normalize=True | PCA=No PCA	0.067935	3.527839	123.557472	0.000000	-8.44160e-08
3	Auto K	S3	S3 | Normalize=False | PCA=50	0.093703	2.964195	173.714447	0.025767	-5.63642e-01

No.	Mode K	Skenario	Konfigurasi	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Delta Silhouette	Delta DBI
4	Auto K	S4	S4 | Normalize=False | PCA=100	0.078206	3.272435	143.185471	0.010271	-2.55432e-01
5	Auto K	S5	S5 | Normalize=True | PCA=50	0.098899	2.912914	186.476974	0.030963	-6.14943e-01
6	Auto K	S6	S6 | Normalize=True | PCA=100	0.083093	3.205876	153.601868	0.015158	-3.21967e-01
7	Manual K = 5	S1	S1 | Normalize=False | PCA=No PCA	0.057323	3.040013	71.126869	0.000000	0.000000e+00
8	Manual K = 5	S2	S2 | Normalize=True | PCA=No PCA	0.057323	3.040013	71.126869	0.000000	8.62368e-08
9	Manual K = 5	S3	S3 | Normalize=False | PCA=50	0.085375	2.564335	103.546387	0.028052	-4.75677e-01
10	Manual K = 5	S4	S4 | Normalize=False | PCA=100	0.071544	2.799714	83.526405	0.014222	-2.40299e-01
11	Manual K = 5	S5	S5 | Normalize=True | PCA=50	0.076219	2.978697	104.785767	0.018897	-6.13181e-02
12	Manual K = 5	S6	S6 | Normalize=True | PCA=100	0.061009	3.349425	85.218124	0.003687	3.09415e-01

Based on the ablation study framework, this study systematically analyzes the architectural impact of text preprocessing configurations and dimensionality reduction on semantic clustering quality, utilizing the baseline experimental scenario S1 as a comparative reference point [16], [2]. Each dimensional scenario is rigorously evaluated against empirical shifting patterns in the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) [13], [14]. The comparative synthesis demonstrates that the strategic implementation of Principal Component Analysis (PCA) exerts a highly statistically significant influence on boosting downstream partitioning performance across both the unsupervised Auto K configuration and the constrained Manual K = 5 routing taxonomy [32], [15].

Under the automated Auto K pipeline, scenario S5 yields the most substantial optimization gain relative to the un-optimized S1 baseline execution. This specific experimental scenario integrates a combination of vector normalization and a 50-dimensional PCA projection, culminating in a positive Delta Silhouette increment of 0.030963 alongside a profound Delta CHI structural leap of 62.919502 [10], [46]. Concurrently, the DBI validation metric exhibits a significant minimization shift of -0.614923, highlighting a denser, more distinct semantic cluster separation. These empirical trends confirm that combining length normalization with a calibrated 50-dimensional PCA compression yields a highly optimized latent representation space for short-text embedding architectures compared to standard unprocessed setups [9], [17].

Conversely, regarding the targeted Manual K = 5 clustering partition, the maximum structural performance acceleration is uniquely attained via scenario S3, which deploys a 50-dimensional PCA projection in the absence of vector normalization. This specific architectural routine triggers a positive Delta Silhouette gain of 0.028052 paired with an upward Delta CHI translation of 32.419518 over the original S1 baseline reference [12], [13]. Furthermore, the internal cluster distance profile is optimized as evidenced by a DBI drop of -0.475678, consolidating the overall validity of the semantic groups. This phenomenon indicates that while normalization is critical for unconstrained automated partitions, the application of a 50-dimensional PCA standalone mapping provides the most robust variance-preserving contribution toward maximizing clustering quality under constrained, pre-defined cluster constraints K = 5 [43], [3], [47], [48].

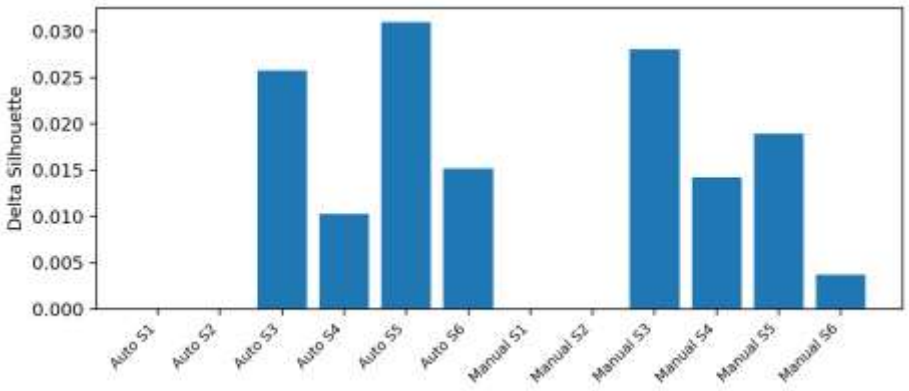


Figure 7. Comparison of Delta Silhouette Against the S1 Baseline

Based on the experimental curves illustrating the Delta Silhouette variations relative to the reference S1 baseline, scenarios configuration deploying a 50-component Principal Component Analysis (PCA) demonstrate a significantly more consistent performance acceleration compared to both the baseline options without PCA compression and those utilizing a 100-component projection [10], [16]. Conversely, expanding the dimensionality reduction mapping to 100 components yields a noticeably smaller optimization margin than the tighter 50-component constraint. This crucial empirical finding indicates that scaling up the number of preserved latent components within a transformer-based semantic pipeline does not monotonically scale cluster spatial quality, as excessive component retention often re-introduces geometric noise or latent multi-collinearity into dense vector topographies [12], [1], [49].

Broadly speaking, the comprehensive synthesis of the ablation study highlights that executing a calibrated feature reduction via a 50-component PCA setup delivers the most consistent positive contribution toward expanding downstream partitioning validity across both the unconstrained Auto K tracking and the pre-defined Manual K = 5 routing model [10], [15]. Concurrently, length-based embedding normalization injects an auxiliary performance boost under localized operational setups, exhibiting high efficacy when coupled with automated cluster size tracking [17]. Consequently, these trends confirm that constructing a well-calibrated combination of manifold transformation operations is imperative to optimize the geometric properties of dense semantic spaces, ultimately maximizing cluster cohesion and inter-group segregation margins [11], [43], [50].

3.8 Limitations of Clustering Performance

The relatively low Silhouette Score obtained in this study should be interpreted within the context of short-text semantic clustering using transformer-based sentence embeddings. The best score of 0.098899 indicates weak numerical separation according to conventional clustering validation standards [13], [50]. However, this value is still interpretable in semantic text clustering because question-answering corpora often contain overlapping meanings, shared domain vocabulary, and smooth semantic transitions between topics [9], [16]. In the Dewi Durga corpus, many entries discuss closely related cultural, spiritual, mythological, historical, and archaeological concepts, which makes hard cluster boundaries difficult to form [21]. Therefore, the clustering result should not be assessed solely from the absolute Silhouette Score, but also from the comparative improvement across scenarios, Davies-Bouldin Index reduction, Calinski-Harabasz Index improvement, and thematic interpretability of the resulting clusters [12], [14], [18], [50].

Another limitation is related to embedding representation bias. The all-MiniLM-L6-v2 model is derived from pretrained sentence-transformer architectures, meaning that the resulting vector space may reflect patterns learned from general-domain corpora rather than cultural heritage-specific knowledge [8], [43]. This condition may reduce the sharpness of semantic separation in a specialized corpus containing religious, historical, and regional cultural expressions. Consequently, although the proposed PCA-enhanced K-Means framework improves internal validation performance, the findings should be

understood as exploratory semantic grouping rather than definitive cultural classification. Future studies should include domain expert validation, compare alternative embedding models, and evaluate clustering quality using both quantitative metrics and qualitative cultural interpretation.

4. Conclusion

This study investigated semantic clustering of a 1,620-entry Dewi Durga question-answering corpus using all-MiniLM-L6-v2 sentence embeddings and K-Means clustering under Auto K and Manual K = 5 configurations. The results demonstrate that transformer-based sentence embeddings can effectively organize cultural and religious texts into meaningful semantic structures. Based on internal validation metrics, the Auto K procedure identified K = 2 as the optimal partition, while Manual K = 5 produced a more detailed thematic taxonomy encompassing ritual-spiritual practices, mythological narratives, historical development, archaeological artefacts, and general religious descriptions.

The ablation analysis revealed that PCA contributed positively to clustering quality, particularly at 50-dimensional projections. The best overall configuration was Auto_K_S5, while Manual_K5_S3 offered the most interpretable thematic separation. These findings indicate that cluster-number selection should balance internal validation performance with semantic interpretability, especially in cultural heritage and digital humanities corpora. Future research may extend this work by evaluating additional sentence embedding models, alternative clustering algorithms, and expert-driven validation of thematic cluster labels.

References

[1] C. Peersman, “A Survey of Relevant Text Mining Technology,” 2022.
[2] C. C. Aggarwal and C. Zhai, “Chapter 4 A SURVEY OF TEXT CLUSTERING ALGORITHMS”.
[3] M. Steinbach, “A Comparison of Document Clustering Techniques,” pp. 1–2.
[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” pp. 1– 12, 2013.
[5] J. Pennington, R. Socher, and C. D. Manning, “GloVe : Global Vectors for Word Representation,” pp. 1532–1543, 2014.
[6] A. Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems, 2017.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” no. Mlm, 2018.
[8] N. Reimers and I. Gurevych, “Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks,” pp. 3982–3992, 2019.
[9] K. Abdalgader, A. A. Matroud, and K. Hossin, “Experimental study on short-text clustering using transformer-based semantic similarity measure,” pp. 1–41, 2024, doi: 10.7717/peerj-cs.2078.
[10] Y. Ortakci, “Engineering Science and Technology , an International Journal Revolutionary text clustering : Investigating transfer learning capacity of SBERT models through pooling techniques,” Eng. Sci. Technol. an Int. J., vol. 55, no. May, p. 101730, 2024, doi: 10.1016/j.jestch.2024.101730.
[11] M. Abay, M. Gemeda, and J. Kalita, “Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms,” Procedia Comput. Sci., vol. 244, pp. 121–132, 2024, doi: 10.1016/j.procs.2024.10.185.
[12] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, and B. Abuhaija, “K-means clustering algorithms : A comprehensive review , variants analysis , and advances in the era of big data,” Inf. Sci. (Ny)., vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.
[13] P. J. Rousseeuw, “Silhouettes : a graphical aid to the interpretation and validation of cluster analysis,” vol. 20, pp. 53–65, 1987.
[14] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, 1979.
[15] F. Pedregosa, O. Grisel, M. Andreas, R. Weiss, A. Passos, and M. Brucher, “Scikit-learn : Machine Learning in Python,” vol. 12, pp. 2825–2830, 2011.
[16] R. Saha, “Influence of various text embeddings on clustering performance in NLP arXiv : 2305 . 03144v1 [cs . LG] 4 May 2023,” no. April, pp. 1–22, 2023.
[17] J. Mu and P. Viswanath, “ALL-BUT-THE-TOP: SIMPLE AND EFFECTIVE POST PROCESSING FOR WORD REPRESENTATIONS,” pp. 1–25, 2018.
[18] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, “MTEB : Massive Text Embedding Benchmark,” pp. 2014–2037, 2023.
[19] S. Ferro, R. Giovanelli, M. Leeson, M. De Bernardin, and A. Traviglia, “A novel NLP-driven approach for enriching artefact descriptions , provenance , and entities in cultural heritage,” Neural Comput. Appl., vol. 37, no. 25, pp. 21275–21296, 2025, doi: 10.1007/s00521-025-11449-2.
[20] A. Pawłowski, “NLP for Digital Humanities : Processing Chronological Text Corpora,” pp. 105–112, 2024.
[21] T. Lathif, M. Suryanto, and A. Prasetya, “Generated cultural heritage question – answer dataset : Durga in multi-dimensional perspectives,” vol. 65, 2026, doi: 10.1016/j.dib.2026.112495.
[22] I. G. A. P. Seno Joko Suyono, D. Butler, Membaca Durga (Bunga Rampai Tulisan Pemikiran Tentang Durga). Yogyakarta: Borobudur Writers and Cultural Society (BWCF), 2022.
[23] D. Cer, Y. Yang, S. Kong, N. Hua, and N. Limtiaco, “Universal Sentence Encoder”.

[24] Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," 2018.
[25] Z. Li et al., "Towards General Text Embeddings with Multi-stage Contrastive Learning," 2023.
[26] M. Ostendorf, W. Y. Noah, and M. Ai, "One Embedder, Any Task: Instruction-Finetuned Text Embeddings," 2022.
[27] L. Wang, N. Yang, X. Huang, and B. Jiao, "Text Embeddings by Weakly-Supervised," pp. 1–17.
[28] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," vol. 5, pp. 135–146, 2017.
[29] A. B. Dieng and D. M. Blei, "Topic Modeling in Embedding Spaces," vol. 8, pp. 439–453, 2020.
[30] D. Angelov, "T 2v : d r t," pp. 1–25.
[31] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2020.
[32] Y. Ortakci and B. Borhan, Optimizing SBERT for long text clustering : two novel approaches with empirical insights. Springer US, 2025. doi: 10.1007/s11227-025-07414-4.
[33] S. Yeasmin, N. Afrin, K. Saif, and M. Rezwanul, "Text Clustering Framework using Transformer based Embeddings".
[34] J. Tussupov, A. Kassymova, A. Mukhanova, and A. Bissengaliyeva, "Analysis of Short Texts Using Intelligent Clustering Methods," pp. 1–15, 2025.
[35] G. Izacard, "Unsupervised Dense Information Retrieval with Contrastive Learning," 2022.
[36] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," pp. 6769–6781, 2020.
[37] M. Guo, J. Ainslie, and D. Uthus, "LongT5: Efficient Text-To-Text Transformer for Long Sequences," 2021.
[38] M. Zaheer et al., "Big Bird : Transformers for Longer Sequences," no. NeurIPS, 2020.
[39] M. E. Peters and A. Cohan, "Longformer: The Long-Document Transformer".
[40] L. Mcinnes, J. Healy, and S. Astels, "hdbscan : Hierarchical density based clustering," vol. 2, no. 2017, pp. 11–12, doi: 10.21105/joss.00205.
[41] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation," pp. 4512–4525, 2020.
[42] L. Mcinnes, J. Healy, and J. Melville, "UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction," 2020.
[43] T. Gao, X. Yao, and D. Chen, "SimCSE : Simple Contrastive Learning of Sentence Embeddings," pp. 6894–6910, 2021.
[44] A. Petukhova, "Text Clustering with Large Language Model," pp. 1–25, 2024.
[45] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," vol. 9, pp. 2579–2605, 2008.
[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," no. Figure 1, 2019.
[47] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 1027–1035.
[48] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
[49] N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation," arXiv:2004.09813, 2020.
[50] S. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," Journal of Intelligent Information Systems, vol. 17, no. 2–3, pp. 107–145, 2001.