❏   641

K-Means Optimization Algorithm to Improve Cluster Quality on Sparse Data

Yully Sofyah Waode , Anang Kurnia , Yenni Angraini
IPB University, Bogor, Indonesia

Article Info	ABSTRACT
	The aim of this research is to cluster sparse data using various K-means optimization algorithms. Sparse data used in this research came from Citampi Stories game reviews on Google Play Store. This research method is Density-Based Spatial Clustering of Applications with Noise-Kmeans (DB-Kmeans), Particle Swarm Optimization-Kmeans (PSO-Kmeans), and Robust Sparse Kmeans Clustering (RSKC) are evaluated using the silhouette score. Clustering sparse data presented a challenge as it could complicate the analysis process, leading to suboptimal or non-representative results. To address this challenge, **the research employed an approach** that divided the data based on the number of terms in three different scenarios to reduce sparsity. **The results of this** research showed that DB-Kmeans had the potential to enhance clustering quality across most data scenarios. Additionally, this research found that dividing data based on the number of terms could effectively mitigate sparsity, significantly influencing the optimization of topic formation within each cluster. **The conclusion of this** research is that this approach is effective in enhancing the quality of clustering for sparse data, providing more diverse and easily interpretable information. The results of this research could be valuable for developers seeking to understand user prefer-ences and enhance game quality.

Corresponding Author:

Yully Sofyah Waode, +6287735323703,
Faculty of Mathematics and Natural Sciences, Statistics and Data Science,
IPB University, Bogor, Indonesia,
Email: yullysofyah.waode@apps.ipb.ac.id.

How to Cite:
Y. Waode, A. Kurnia, and Y. Angraini, "K-Means Optimization Algorithm to Improve Cluster Quality on Sparse Data", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 3, pp. 641-652, July, 2024.
This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

1. INTRODUCTION

The rapid development of technology and information today has an important role in people's lives, causing an expo-nential increase in the amount of data. One of them is the increase in digital text data available online. Various techniques have been developed to extract knowledge and analyze very large and messy data more simply [1–3]. Text mining is a technique that functions to handle unstructured data such as collections of emails, social media, websites, reviews, and others [4]. Text clustering is one of the tasks in text mining that is widely used on digital text data. Text clustering can facilitate un-derstanding of specific information based on the topics discussed in a text dataset by clustering [3]. Text clustering is a process of grouping texts with similar characteristics into the same cluster [5]. One of the most commonly used algorithms in text clus-tering is K-means clustering. However, this algorithm has several disadvantages, requiring users to determine the number of clusters to be formed in advance, but sometimes difficult to understand. Another disadvantage is the clustering results using this algorithm are very sensitive to the selection of the initial centroid [6]. Initial centroids that are determined randomly and inappropriately can result in local optimums and may even produce incorrect clustering results [7]. Several studies have pro-posed optimization algorithms to overcome the weaknesses of the K-Means algorithm, including DB-Kmeans [6] and PSO-Kmeans [8].

There are gaps that have not been resolved by previous research [6], namely the limited use of text data. Additionally, there is relatively little research discussing the DB-Kmeans algorithm. As a result, further studies are needed to compare DB-Kmeans with other K-Means optimization algorithms using different types of text data. This will help evaluate whether the optimization algorithm still maintains its superiority and effectiveness in clustering. This research implements the K-Means optimization algorithm for clustering digital text data, specifically a collection of game reviews on the Google Play Store. The results of sentiment clustering of game reviews are expected to make it easier to identify user responses, criticisms, and sugges-tions to improve game quality. A survey conducted by the Indonesian Institute of Sciences (LIPI) in 2020 recorded more than 70 game developers spread throughout Indonesia, and there may even be hundreds more [9]. One of the games developed in Indonesia is Citampi Stories, developed by Ikan Asin Production and downloaded more than 1 million times. Game user re-views, which are short text documents, are one of the focuses of this research.

The study of short-text documents has gained attention in recent years due to their presence in various fields [10]. How-ever, clustering short text documents presents unique challenges compared to plain text because it can produce noise and sparsity in the analysis process, which is known as sparse data [11]. This is because traditional similarity measures tend to produce very sparse vector representations, resulting in overlapping and ineffective clustering results [12]. Therefore, it is nec-essary to explore how to overcome these problems. This research divides the review data based on the number of terms into several scenarios to overcome the sparsity issue. This aims to maximize the identification of topics within each cluster and simplify the interpretation of clustering results on sparse data. Additionally, this research also uses a sparsity-based K-Means optimization algorithm as a comparison, namely Robust Sparse Kmeans Clustering (RSKC) [13].

Based on the previous description, this research focuses on clustering the sentiment of Citampi Stories game reviews us-ing DB-Kmeans, PSO-Kmeans, and RSKC algorithms. The difference between this research and previous studies is the use of sparse data in clustering and the comparison of performance among different optimization algorithms. The aim of this re-search is to evaluate various K-means optimization algorithms for clustering digital text data and to address the challenges associated with clustering sparse data. The results of this research are expected to offer a new perspective for improving the quality of clustering results on sparse data and addressing the sparsity problem in the analysis of short text documents. Addi-tionally, this research is expected to contribute to advancing K-Means optimization algorithms for more effective clustering and improving the understanding of user preferences, assisting developers in designing and improving game quality.

2. RESEARCH METHOD

The stages of this research are described using the flowchart shown in Figure 1. This research method utilizes K-Means optimization algorithms, specifically DB-Kmeans, PSO-Kmeans, and RSKC. The data is sourced from reviews data on the Google Play Store.
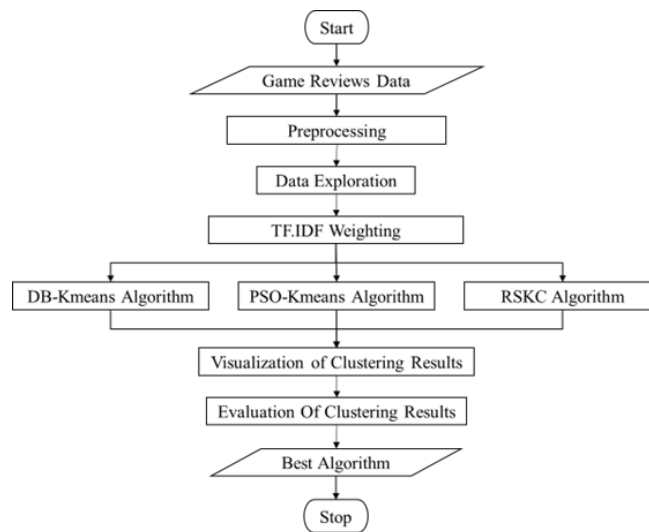
Figure 1. Research Flowchart

2.1. Game Reviews Data

The data used in this research is a collection of Citampi Stories game reviews on the Google Play Store, which were taken through scraping using Python. This game offers a similar experience to the game Harvest Moon: Back To Nature, where players embark on missions to inhabit a village, cultivate a farm, interact with villagers, and engage in various activities. The reviews analyzed are restricted to the Indonesian language. Data was collected over a period spanning from May 7, 2019, to August 9, 2023, amounting to a total of 52046 reviews.

2.2. Preprocessing

The collected data then undergoes the initial process of text mining, namely preprocessing. This process aims to trans-form the utilized data into a more structured format that is ready for processing. Preprocessing removes noise and extracts features for subsequent processing using machine learning algorithms. Noise removal from the text dataset includes tasks such as correcting spelling errors, reducing duplicate characters, and clarifying ambiguous acronyms. The stages of preprocessing are as follows [14, 15]. Cleaning the review data by removing duplicates and eliminating emojis, punctuation marks, num-bers, URLs, and hashtags. Case folding refers to converting all data into lowercase. Tokenizing is the process of identifying and separating sentences into the smallest parts known as tokens. Tokens are typically words or phrases. Normalization is the process of converting words into standard word forms, according to the Big Indonesian Dictionary (KBBI). Stopword removal is the process of removing words that are irrelevant in the analysis where the words are so common in the corpus that they may provide little information. Examples include in, and, at, and others. Stemming is the process of removing affixes on words to reduce them to their basic word forms.

2.3. Data Exploration

After the preprocessing stage, the data is explored by examining the number of terms in each review. This examination helps determine the number of terms formed and the number of reviews with the same number of terms. The review data is then divided into three scenarios based on the number of terms determined based on the exploration results.

2.4. TF-IDF Weighting

Before being used in the clustering process, the data is compiled into a matrix containing the terms' weights. This is done to determine the importance of each term [16]. The specific method used for term weighting in this research is known as Term Frequency - Inverse Document Frequency (TF-IDF).

2.5. K-Means

K-Means algorithm is an algorithm used for clustering analysis that leads to partitioning n observations into k clusters. In this research, data is grouped into two or more clusters based on the closest distance to the centroid. The stages performed in the K-Means algorithm are as follows [17]: (a). Determine the number of clusters (k) to be formed and randomly select k centroids; (b). Group each object in the cluster based on its closest distance to the centroid by using Equations (1) and (c). Calculate the centroid value of the k-th cluster by using Equation (2), and (d). Repeat step b until there is no change in the centroid value. with $D(x,y)$: distance between data x to data y, p : data dimension. with m_k : centroid value of the k-th cluster, n_k : the number of data in cluster k, x_j : j-th data in cluster k.

$$D(x, y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \tag{1}$$

$$m_k = \frac{1}{n_k}\sum_{j=1}^{\pi_k} x_j \tag{2}$$

2.6. DB-Kmeans

DB-Kmeans is an algorithm that combines Density-Based Spatial Clustering of Application with Noise (DBSCAN) and K-Means by overcoming the shortcomings of these algorithms and maximizing their advantages. This algorithm can over-come the problem of K-Means being sensitive to outliers and can overcome the slow clustering using the DBSCAN algorithm. The DB-Kmeans algorithm optimizes the initial centroid selection using DBSCAN to improve the clustering accuracy. The stages performed in the DB-Kmeans algorithm are as follows [6] : (a). The DBSCAN algorithm is used for pre-clustering by dividing all data based on density to obtain the number of clusters and centroid of each cluster; (b). The centroid and number of clusters obtained in the previous stage as the initial centroid and the number of clusters to be formed (k); (c) . The K-Means algorithm is then employed for subsequent clustering.

2.7. PSO-Kmeans

PSO-Kmeans is an algorithm that combines Particle Swarm Optimization (PSO) and K-means with PSO, which serves to optimize the initial centroid to obtain an increase in cluster quality. PSO is a population-based optimization algorithm inspired by the social behavior of bird flocks. In PSO, particle positions are updated iteratively to optimize a specific objective function [8]. The stages performed in the PSO-Kmeans algorithm are as follows [18]: (a). Determine the number of clusters to be formed; (b). Determine the initial centroid value randomly and group the data at the nearest centroid point; (c). Calculating the fitness value of the formed cluster using the silhouette score which is further explained in the next subheading; (d). Calcu-lating the velocity (v) using equation (3); (e). Update the centroid using Equation (4); (f). The iteration stops when the number of iterations has reached the specified maximum limit. In this research, 200 iterations are used; (g). The result of this PSO is the centroid point that will be used as the initial centroid in the K-Means algorithm.

$$v_i^t = wv_i^{t-1} + c_1 r_1(P_i - x_i^{t-1}) + c_2 r_2(G - x_i^{t-1}) \tag{3}$$

With t : iteration, v_i : velocity for the i-th particle, w : inertia value that serves to balance local and global search obtained using random numbers generated with uniform distribution [0.5, 1], x_i : position of the i-th particle, P_i : best position of each particle (personal best), G : best position for all particles, c_1 and c_2 : cognitive and social constants, respectively, r_1 and r_2 : random numbers generated with uniform distribution [0, 1].

$$x_i^t = x_i^{t-1} + v_i^t \tag{4}$$

2.8. Robust Sparse Kmeans Clustering (RSKC)

The Robust Sparse Kmeans Clustering (RSKC) algorithm is designed to be resistant to outliers. RSKC works by iteratively removing outliers from the cluster analysis, assigning remaining samples to clusters, and then reintroducing the outliers to the analysis by grouping them into nearest-neighbor clusters [13]. In this research, RSKC was operated using the RSKC package in R Studio.

2.9. Visualization of Clustering Results

The scatter plots visualization in this research uses tSNE dimension reduction results where different groups are repre-sented by different colors. However, clustering is not performed on the tSNE dimension reduction data. tSNE is used for visu-alization because it works well in projecting points from high-dimensional space into 2D [13].

2.10. Evaluation of Clustering Results

Evaluation of clustering results is a technique used to compare the clustering results of several algorithms and determine the best one. This research utilizes a clustering result evaluation measure called the Silhouette Score because it is more suitable for clustering tasks as it doesn't require any training data. The Silhouette Score calculates how well an object is placed within a cluster and how well an object is separated from other clusters. The score ranges from -1 to 1, with a value near 1 indicating that the object is well-clustered, and a value near -1 indicating that the object is incorrectly clustered. The silhouette score is calculated using equation (5) [19]. With $s(i)$: silhouette value for the i-th review in cluster C_i, $\alpha(i)$: average distance of the i-th review to all reviews in cluster C_i, $b(i)$: average distance of the i-th review to all reviews in other clusters.

$$s(i) = \frac{b(i) - \alpha(i)}{max(\alpha(i), b(i))} \tag{5}$$

3. RESULT AND ANALYSIS

3.1. Preprocessing

The collection of Citampi Stories game reviews was obtained and cleaned by removing duplicates, emojis, punctuation marks, numbers, URLs, and hashtags. This cleaning is done to obtain important information in the data. Furthermore, case folding, tok-enizing, normalization, stopword removal, and stemming are applied. An illustration of the preprocessing results is shown in Table 1. After preprocessing, the formed terms are reduced by combining several and changing acronyms that are not in the dictionary. Term reduction aims to maximize the meaning of each term and lessen sparsity in the analysis process. An illustration of the term reduction results is shown in Table 2.

Table 1. Illustration of Preprocessing Results

No	Before	After
1	Ini salah satu game buatan Indonesia terbaik sih, suka banget sama karakter yang disediakan dan lagunya, kalo mau top up juga gak mahal yaaa masih terjangkaulah, dev nya juga mendengarkan setiap keluhan/bug di game dan langsung di fix. Saranku karakter utamanya tambahin biar bisa ganti gender dan tambahin karakter cowok yang bisa di nikahin. Sukses terus deh∂Ÿ'	'game', 'buatan', 'indonesia', 'baik', 'suka', 'karakter', 'sedia', 'lagu', 'top', 'up', 'mahal', 'jangkau', 'dev', 'dengar', 'keluh', 'bug', 'game', 'saran', 'karakter', 'utama', tambah', 'tambah', 'karakter', 'laki', 'laki', 'nikah', 'sukses'
2	Min, masih ada bug layar hitamnya	'bug', 'layar', 'hitam'
3	GAME SERU banget nih harus download!!!!!!!!!!!!!!	'game', 'seru', 'install'

Table 2. Illustration of Term Reduction Results

No	Before	After
1	'game', 'buatan', 'indonesia', 'baik', 'suka', 'karakter', 'sedia', 'lagu', 'top', 'up', 'mahal', 'jangkau', 'dev', 'dengar', 'keluh', 'bug', 'game', 'saran', 'karakter', 'utama', tambah', 'tambah', 'karakter', 'laki', 'laki', 'nikah', 'sukses'	game', 'karya_anak_bangsa', 'baik', 'suka', 'karakter', 'sedia', 'lagu', 'top_up', 'mahal', 'jangkau', 'developer', 'dengar', 'keluh', 'bug', 'game', 'saran', 'tambah_versi_perempuan', 'tambah_karakter', 'laki_laki', 'nikah', 'sukses'
2	'bug', 'layar', 'hitam'	'bug', 'black_screen'
3	'game', 'seru', 'install'	'game', 'seru', 'install'

3.2. Data Exploration

The results of preprocessing and term reduction are then explored by looking at the distribution of the number of terms in each review. The distribution of the number of terms determines how many terms form in the reviews. It also shows how many reviews

have the same number of terms. The terms formed in the reviews data are 365 by removing terms appearing in less than 100 reviews. This aims to get terms that have a role in forming topics in the review data. The distribution of the num-ber of terms is shown in Figure 2.
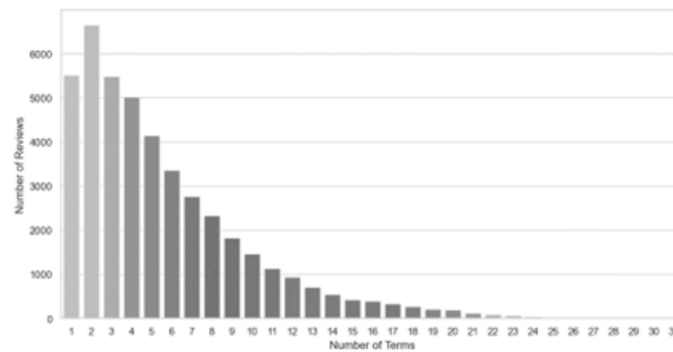

Figure 2. Distribution of The Number of Terms

Figure 2 shows that more than 35000 reviews have less than 10 terms. This can lead to creating a very sparse document-term matrix (DTM) when TF.IDF weighting is applied. Sparse data can result in ineffective clustering [12]. To overcome this problem, this research divides the review data based on the number of terms in the following scenarios. Scenario 1: all reviews data (Dataset 1), Scenario 2: reviews data divided into 2 datasets, namely data with 10 terms or fewer (Dataset 2) and data with more than 10 terms (Dataset 3), and Scenario 3: reviews data divided into 2 datasets, namely data with 15 terms or fewer (Dataset 4) and data with more than 15 terms (Dataset 5).

3.3. Scenario 1

Review data with Scenario 1, named Dataset 1, is converted into numerical form using TF-IDF weighting and then pre-sented to form DTM. The document-term-matrix (DTM) consists of 44163 reviews and 365 terms. Several experiments in this research found the best clustering results. These results were obtained when k = 8 for DB-Kmeans, k = 9 for PSO-Kmeans, and k = 8 for RSKC. Figure 3 shows the clustering results of Dataset 1 using scatterplots. Different colors represent different clus-ters.
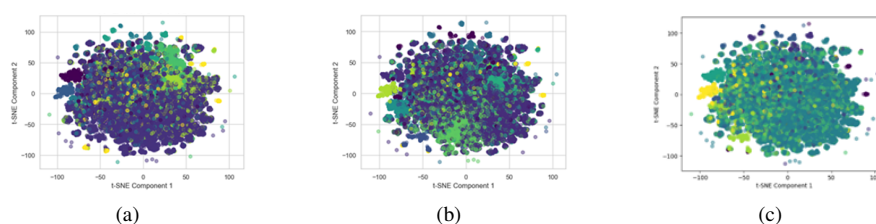

Figure 3. Scatterplot of Dataset 1 Clustering Results Using (a) DB-Kmeans; (b) PSO-Kmeans; (c) RSKC

3.4. Scenario 2

Review data with Scenario 2 consisting of Dataset 2 and Dataset 3 are converted into numerical form using TF-IDF weighting. In Dataset 2, the DTM formed has 37817 reviews and 365 terms. Meanwhile, in Dataset 3, the DTM formed has 5576 reviews and 365 terms. On Dataset 2, several experiments in this research found the best clustering results. These results were obtained when k = 9 for DB-Kmeans, k = 9 for PSO-Kmeans, and k = 8 for RSKC. Figure 4 shows the clustering results of Dataset 2 using scatterplots. Different colors represent different clusters. On Dataset 3, several experiments in this research found the best clustering results. These results were obtained when k = 9 for DB-Kmeans, k = 9 for PSO-Kmeans, and k = 8 for RSKC. Figure 5 shows the clustering results of Dataset 3 using scatterplots. Different colors represent different clusters.
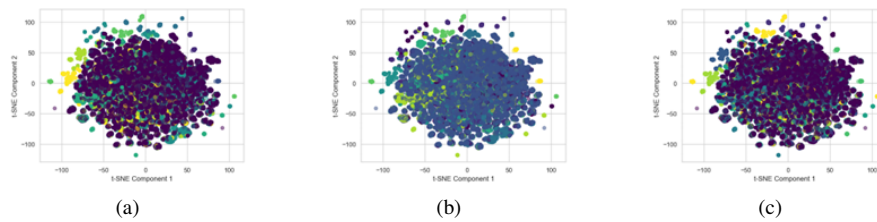
(a)   (b)   (c)

Figure 4. Scatterplot of Dataset 2 Clustering Results Using (a) DB-Kmeans; (b) PSO-Kmeans; (c) RSKC
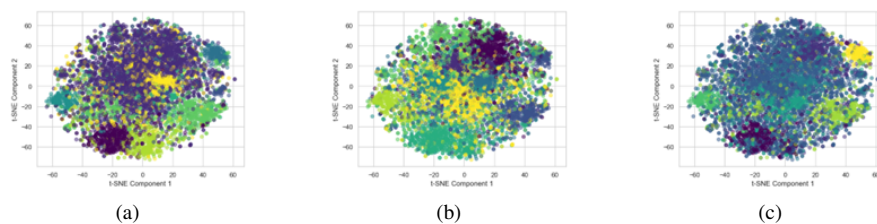

(a)   (b)   (c)

Figure 5. Scatterplot of Dataset 3 Clustering Results Using (a) DB-Kmeans; (b) PSO-Kmeans; (c) RSKC

3.5. Scenario 3

Review data with Scenario 3 consisting of Dataset 4 and Dataset 5 are converted into numerical form using TF-IDF weighting. In Dataset 4, the DTM formed has 41973 reviews and 365 terms. Meanwhile, in Dataset 5, the DTM formed has 1817 reviews and 365 terms. On Dataset 4, several experiments in this research found the best clustering results. These results were obtained when k = 9 for DB-Kmeans, k = 9 for PSO-Kmeans, and k = 8 for RSKC. Figure 6 shows the clustering results of Dataset 4 using scatterplots. Different colors represent different clusters. On Dataset 5, several experiments in this research found the best clustering results. These results were obtained when k = 2 for DB-Kmeans, k = 4 for PSO-Kmeans, and k = 9 for RSKC. Figure 7 shows the clustering results of Dataset 5 using scatterplots. Different colors represent different clusters.

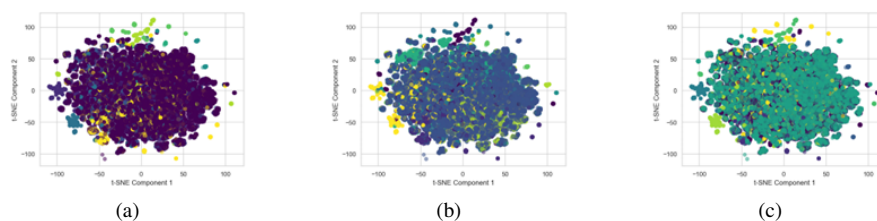(a)   (b)   (c)

Figure 6. Scatterplot of Dataset 4 Clustering Results Using (a) DB-Kmeans; (b) PSO-Kmeans; (c) RSKC
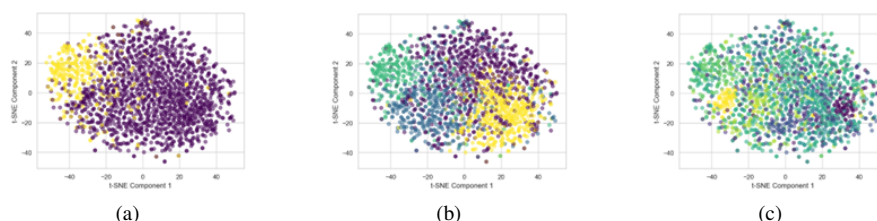

(a)   (b)   (c)

Figure 7. Scatterplot of Dataset 5 Clustering Results Using (a) DB-Kmeans; (b) PSO-Kmeans; (c) RSKC

3.6. Analysis

The clustering results obtained from Scenario 1, Scenario 2, and Scenario 3 are evaluated by calculating the silhouette score of the K-Means optimization algorithm utilized. This research employs other commonly used algorithms, namely K-Means and DBSCAN, for comparative analysis on the same dataset. The comparison of silhouette scores across Scenario 1, Scenario 2, and Scenario 3 is presented in Table 3. The DB-Kmeans algorithm yielded the highest silhouette score for Dataset 1, Dataset 2, and Dataset 5, while the PSO-Kmeans algorithm yielded the highest silhouette score for Dataset 3 and Dataset 4. The next step is to identify the topics in each scenario using the best-performing algorithm. Topics are determined by examin-ing terms with high frequency in the resulting term set. Table 4, Table 5, Table 6, Table 7, and Table 8 shows the topics formed in each scenario. Based on Table 4, the topics in Scenario 1 mostly show positive sentiments and the invitation to install the Citampi Stories game. In Scenario 1, other users' suggestions and complaints go undetected. On the other hand, the topics in Scenario 2 (Table 5 & Table 6) and Scenario 3 (Table 7 & Table 8) form a more diverse sentiment than Scenario 1. Clusters in scenarios 2 and 3 have positive responses, suggestions,

Table 3. Silhouette Score of Clustering Results

Scenario	Dataset	K-Means	DBSCAN	DB-Kmeans	PSO-Kmeans	RSKC
Scenario 1	Dataset 1	0.050765	0.012948	**0.054569**	0.049021	0.051323
Scenario 2	Dataset 2	0.061904	0.030913	**0.062623**	0.059008	0.057286
	Dataset 3	0.015077	-0.03971	0.013265	**0.015451**	0.012715
Scenario 3	Dataset 4	0.041686	0.022091	0.045681	**0.050773**	0.049194
	Dataset 5	0.013404	-0.01873	**0.014266**	0.012737	0.010982

Table 4. Topics on Clustering Results of Dataset 1

Cluster	Topic	Term
1	Positive response	Keren, Game, Bagus, Baru, Cerita, Tambah, Developer, Suka
2	Positive response	Game, Bagus, Baru, Tambah
3	Positive response	Mantap, Game, Bagus, Baru, Developer, Seru
4	Positive response	Seru, Game, Bagus, Baru, Main, Bosan
5	Positive response	Game, Suka, Bagus, Main, Baru, Seru
6	Invitation to install	Game, Install, Bagus, Nyesal, Seru, Wajib
7	Positive response	Main, Game, Bagus, Seru, Ulang, Baru, Save
8	Positive response	Game, Bagus, Baru, Developer, Bintang, Cerita

Table 5. Topics on Clustering Results of Dataset 2

Cluster	Topic	Term
1	Positive response	Bagus, Game, Tambah, Seru, Baru, Terima kasih, Developer, Anak
2	Add female characters	Bagus, Game, Bintang, Terima kasih, Semangat, Tambah, Perempuan, Karakter
3	Positive response	Bagus, Gemas, Sangat, Wibu, Singkat, Super, Debat
4	Positive response	Seru, Game, Bagus, Baru, Bosan, Main, Tambah, Pokok
5	Invitation to install	Main, Game, Bagus, Install, Seru, Terima kasih, Nyesal
6	Positive response	Baru, Game, Bagus, Developer, Tunggu, Lanjut, Seru
7	Positive response	Suka, Game, Bagus, Main, Baru, Seru, Cerita, Terima kasih
8	Positive response	Mantap, Game, Bagus, Baru, Developer, Suka, Karya anak bangsa, Main
9	Positive response	Keren, Game, Bagus, Baru, Main, Developer, Seru, Suka

Table 6. Topics on Clustering Results of Dataset 3

Cluster	Topic	Term
1	More children and wives	Game, Anak, Istri, Tambah, Baru, Bagus, Rumah, Jalan
2	Expand the map	Game, Bagus, Map, Tambah, Luas, Baru, Saran, Karakter
3	Difficult to find items	Game, Cari, Susah, Barang, Bagus, Baru
4	Add another character	Game, Perempuan, Karakter, Laki laki, Bagus, Tambah
5	Positive response	Game, Cerita, Bagus, Tambah, Baru, Main, Alur
6	Technical and advertising issues	Game, Save, Main, Bagus, Ulang, Iklan, Hilang, Data, Bug
7	Positive response	Game, Tambah, Bagus, Saran, Baru, Pakai, Rumah, Karakter
8	Mission to be accomplished	Game, Beli, Bagus, Rumah, Hutang, Uang, Bayar, Orangtua
9	Invitation to install	Game, Bagus, Main, Baru, Developer, Seru, Install

Table 7. Topics on Clustering Results of Dataset 4

Cluster	Topic	Term
1	Positive response	Suka, Game, Bagus, Seru, Main, Baru, Terima kasih, Keren
2	Positive response	Baru, Game, Tunggu, Bagus, Lanjut, Developer, Semangat
3	Positive response	Game, Bagus, Baru, Developer, Saran, Seru, Terima kasih
4	Add multi player mode	Seru, Game, Bagus, Baru, Main, Tambah mode multi player
5	Positive response	Main, Game, Bagus, Seru, Baru, Developer, Ulang, Suka
6	Invitation to install	Install, Game, Nyesal, Bagus, Seru, Wajib, Suka, Baru
7	Positive response	Bagus, Game, Developer, Sangat, Debat, Wow, Gemas
8	Positive response	Tambah, Game, Bagus, Baru, Saran, Developer, Area, Seru
9	Positive response	Mantap, Game, Keren, Bagus, Baru, Karya anak bangsa

Table 8. Topics on Clustering Results of Dataset 5

Cluster	Topic	Term
1	More children and wives	Game, Bagus, Tambah, Istri, Anak, Baru, Main, Saran
2	Technical and advertising issues	Game, Save, Main, Iklan, Baru, Bagus, Ulang, Data, Hilang

3.7. Discussion

The findings of this research show that DB-Kmeans performed exceptionally well across various data scenarios, con-sistently achieving the highest silhouette score in the majority of the datasets. This shows its ability to improve clustering quali-ty on a large amount of data. The results of this research are in line with previous research [6], which also showed the superior-ity of DB-Kmeans. Conversely, the Robust Sparse Kmeans Clustering (RSKC) did not yield optimal results in this research, garnering a lower silhouette score compared to DB-Kmeans and PSO-Kmeans in certain datasets. This outcome diverges from previous research [13], which showed the success of RSKC on sparse data. The disparity can be attributed to the sparse data utilized in this research not exhibiting high dimensions. Furthermore, the silhouette score of all algorithms in Table 3 is close to 0, indicating an overlap between clusters.

Sparse data presents a challenge in the clustering process [20]. As depicted in Figure 3, clustering sparse data leads to over-lapping and suboptimal results, which **aligns with findings** from previous studies [12]. The clusters formed in scenario 1 tend to have similar topics and do not effectively identify the issues encountered by different users. This makes it difficult to improve in subsequent versions of the game and highlights the necessity for further testing. The findings of this research show that clustering sparse data by dividing the data based on the number of terms significantly influences the optimization of top-ic formation within each cluster. This is evident from the diverse and unique topics observed in clusters formed in Scenario 2 and Scenario 3, compared to Scenario 1. As depicted in Figure 3, Figure 4 and Figure 6, the clustering results indicate subopti-mal cluster formationconversely, Figure 5 and Figure 7 display scatterplots depicting improved clustering outcomes.

The approach proposed in this research suggests that dividing the data based on the number of terms can provide more diverse and easily interpretable information. This makes understanding the sentiment in Citampi Stories game reviews easier without losing information. However, manually dividing the data based on the number of terms has several limitations, as it is not optimal and can cause bias in the topics formed. This research has not made an effort to divide the optimal data by con-sidering the goodness of the cluster results. The topics in Scenario 2 have more sentiment variations compared to Scenario 3. Table 5 and Table 6 show that Scenario 2 encompasses a wide range of complaints and constructive suggestions to improve the game's quality. Therefore,

for sentiment clustering of Citampi Stories game reviews, Scenario 2 is more recommended. The clustering results indicate that the majority of Citampi Stories game reviews are positive. Additionally, Citampi Stories game reviews also include invitations to download, mission descriptions, and some suggestions and complaints, such as the need for additional characters, map expansion, difficulty in finding items, technical problems, and advertising problems.

4.   CONCLUSION

The results of the evaluation using the silhouette score demonstrate that the DB-Kmeans algorithm outperforms other algorithms in most data scenarios, indicating its effectiveness in enhancing clustering quality for large datasets. The proposed approach of dividing the data based on the number of terms can be a solution to sparse data clustering, as it maximizes the formation of topics in each cluster. The topics obtained are easier to interpret and more informative. Visualization of cluster-ing results also shows that dividing data based on the number of terms results in better cluster formation. In the context of sentiment clustering for Citampi Stories game reviews, scenario 2 is preferred as it yields a wider variety of sentiments, includ-ing a diverse range of constructive feedback and suggestions, with the majority of the reviews being positive. However, it's important to note that this research did not conduct experiments to determine the optimal number of terms, which is a limita-tion when considering the quality of the cluster results for sparse data division. Therefore, future research should explore vary-ing data division methods to identify their impact on clustering results and compare them to other sparse data clustering tech-niques to achieve optimal clustering outcomes.

5.   ACKNOWLEDGEMENTS

6.   DECLARATIONS

AUTHOR CONTIBUTION

This research was conducted by three authors, each with a division of tasks: Yully Sofyah Waode handled data collection and analysis and wrote articles; Anang Kurnia checked the data results and provided suggestions and criticisms about the re-search; and Yenni Angraini checked the data results and provided suggestions and criticisms about the research.

FUNDING STATEMENT

COMPETING INTEREST

I have no declaration under financial, general, and institutional competing interests.

REFERENCES

[1] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research*, vol. 9, no. 1, pp. 381–386, 2020, https://doi.org/10.21275/ART20203995.

[2] I. C. Chang, T. K. Yu, Y. J. Chang, and T. Y. Yu, "Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals," *Sustainability (Switzerland)*, vol. 13, no. 19, pp. 1–20, 2021, https://doi.org/10.3390/su131910856.

[3] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *Journal of Big Data*, vol. 9, no. 1, pp. 1–21, 2022, https://doi.org/10.1186/s40537-022-00564-9.

[4] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, pp. 1–34, 2020, https://doi.org/10.3390/bdcc4010001.

[5] A. A. Amer and H. I. Abdalla, "A set theory based similarity measure for text clustering and classification," *Journal of Big Data*, vol. 7, no. 74, pp. 1–43, 2020, https://doi.org/10.1186/s40537-020-00344-3.

[6] X. Gao, X. Ding, T. Han, and Y. Kang, "Analysis of influencing factors on excellent teachers' professional growth based on DB-Kmeans method," *Eurasip Journal on Advances in Signal Processing*, vol. 117, no. 1, pp. 1–11, 2022, https://doi.org/10.1186/s13634-022-00948-2.

[7] S. He, D. Luo, and K. Guo, "Evaluation of mineral resources carrying capacity based on the particle swarm optimization clustering algorithm," *Journal of the Southern African Institute of Mining and Metallurgy*, vol. 120, no. 12, pp. 681–691, 2020, https://doi.org/10.17159/2411-9717/1139/2020.

[8] M. A. Hosen, S. H. Moz, S. S. Kabir, S. M. Galib, and M. N. Adnan, "Enhancing Thyroid Patient Dietary Management with an Optimized Recommender System based on PSO and K-means," in *Procedia Computer Science*, vol. 230, no. 3, 2023, pp. 688–697, https://doi.org/10.1016/j.procs.2023.12.124.

[9] M. B. Aulia and L. Kusdibyo, *Analisis Persepsi Konsumen Terhadap Desain Game Buatan Indonesia Dalam Konteks Teori Game Design*, Bandung, 2021, vol. 12, no. 12.

[10] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1427–1445, 2020, https://doi.org/10.1109/TKDE.2020.2992485.

[11] S. Yang, G. Huang, and B. Cai, "Discovering Topic Representative Terms for Short Text Clustering," *IEEE Access*, vol. 7, no. 7, pp. 92 037–92 047, 2020, https://doi.org/10.1109/ACCESS.2019.2927345.

[12] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, "A self-training approach for short text clustering," *Workshop on Representation Learning for NLP*, vol. 4, no. 8, pp. 194–199, 2020, https://doi.org/10.18653/v1/w19-4322.

[13] J. L. Balsor, K. Arbabi, D. Singh, R. Kwan, J. Zaslavsky, E. Jeyanesan, and K. M. Murphy, "A Practical Guide to Sparse k-Means Clustering for Studying Molecular Development of the Human Brain," *Frontiers in Neuroscience*, vol. 15, no. 11, pp. 1–28, 2021, https://doi.org/10.3389/fnins.2021.668293.

[14] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organizational Research Methods*, vol. 25, no. 1, pp. 114–146, 2022, https://doi.org/10.1177/1094428120971683.

[15] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," *Applied Sciences (Switzerland)*, vol. 12, no. 17, pp. 1–21, 2022, https://doi.org/10.3390/app12178765.

[16] R. G. García, B. B. Án, D. V. Nõ, C. Zepeda, and R. Martínez, "Comparison of Clustering Algorithms in Text Clustering Tasks," *Computacion y Sistemas*, vol. 24, no. 2, pp. 429–437, 2020, https://doi.org/10.13053/CyS-24-2-3369.

[17] N. Nurahman, A. Purwanto, and S. Mulyanto, "Klasterisasi Sekolah Menggunakan Algoritma K-Means berdasarkan Fasilitas, Pendidik, dan Tenaga Pendidik," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 337–350, 2022, https://doi.org/10.30812/matrik.v21i2.1411.

[18] I. G. M. S. S. Krisna, I. W. Supriana, I. D. M. B. A. Darmawan, A. Muliantara, N. A. S. ER, and L. G. Astuti, "Perbandingan Pengelompokan Metode PSO K-Means Dan Tanpa PSO Dalam Pengelompokan Data Alert," *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, vol. 11, no. 2, pp. 283–290, 2022, https://doi.org/10.24843/jlk.2022.v11.i02.p07.

[19] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021, https://doi.org/10.3390/e23060759.

[20] S. Cao and X. Li, "Research on Disease and Pest Prediction Model Based on Sparse Clustering Algorithm," in *Procedia Computer Science*, vol. 208, no. 7, 2022, pp. 263–270, https://doi.org/10.1016/j.procs.2022.10.038.

[This page intentionally left blank.]