International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 3, June 2023, pp. 3255~3265
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i3.pp3255-3265

k-means variations analysis for translation of English Tafseer Al-Quran text

Mohammed A. Ahmed[1,2], Hanif Baharin[1], Puteri Nor Ellyza Nohuddin[1,3]
[1]Institute of IR 4.0, Universiti Kebangsaan Malaysia, Bangi, Malaysia
[2]Network Engineering Department, College of Engineering, Al-Iraqia University, Baghdad, Iraq
[3]Faculty of Business, Higher Colleges of Technology, Sharjah, United Arab Emirates

Article Info

Article history:

Received Aug 1, 2022
Revised Aug 18, 2022
Accepted Oct 1, 2022

Keywords:

Chapter Al-Baqarah
Internal cluster validation metrics
k-means variations
No free lunch theorem
Tafseer datasets
Text mining

ABSTRACT

Text mining is a powerful modern technique used to obtain interesting information from huge datasets. Text clustering is used to distinguish between documents that have the same themes or topics. The absence of the datasets ground truth enforces the use of clustering (unsupervised learning) rather than others, such as classification (supervised learning). The “no free lunch” (NFL) theorem supposed that no algorithm outperformed the other in a variety of conditions (several datasets). This study aims to analyze the *k*-means cluster algorithm variations (three algorithms (*k*-means, mini-batch *k*-means, and *k*-medoids) at the clustering process stage. Six datasets were used/analyzed in chapter Al-Baqarah English translation (text) of 286 verses at the preprocessing stage. Moreover, feature selection used the term frequency–inverse document frequency (TF-IDF) to get the weighting term. At the final stage, five internal cluster validations metrics were implemented silhouette coefficient (SC), Calinski-Harabasz index (CHI), C-index (CI), Dunn’s indices (DI) and Davies Bouldin index (DBI) and regarding execution time (ET). The experiments proved that *k*-medoids outperformed the other two algorithms in terms of ET only. In contrast, no algorithm is superior to the other in terms of the clustering process for the six datasets, which confirms the NFL theorem assumption.

Corresponding Author:

Mohammed A. Ahmed
Institute of IR 4.0, Universiti Kebangsaan Malaysia
Bangi, Selangor, 43600, Malaysia
Email: p103761@siswa.ukm.edu.my

1. INTRODUCTION

Text mining is a popular approach for obtaining relevant information from documents; it may be applied with clustering, classification, regression, association rule, and frequent mining pattern tools. The unavailability of ground truths in experimental tests necessitates the use of clustering (unsupervised learning) rather than classification (supervised learning). Because this study’s datasets lacked ground truths (labels), clustering or text clustering approaches were employed. Cluster analysis is a way to organize a collection of data elements into manageable groups (observations). Each subset is a cluster, with members in the same cluster being similar to one another but different to those in other clusters. Cluster analysis is utilized in several disciplines, such as image pattern recognition, database management, online search, biology, and security using clustering algorithms [1]. However, according to Wolpert and Macready [2], the “no free lunch” (NFL) theorem demonstrated that there is no way to ensure that a particular algorithm and a clustering algorithm can operate better than the others in a variety of conditions.

The English translation (text) of Tafseer documents chapter Al-Baqarah is the input dataset for the experiments conducted in this study. The translator used the (Tafseer) for guidance in interpreting the verses.

Several Quran Tafseer books have been produced in numerous languages by classical to modern academics. Every Tafseer book has its unique style, topic matter, phrase context, and judgment. Some contend that the text of the Holy Quran cannot be formed in another language or format. In fact, the interpretation of the Holy Quran has always been complicated and difficult. Moreover, the words in the Holy Quran have contextual significance, making an exact translation much more challenging [3]. Figure 1 depicts the difficulties encountered by the reader in picking the best Tafseer for a certain Quran chapter. Consequently, it is essential to analyze many Qur'anic Tafseer translations for certain topics, such as. chapter Al-Baqarah in this study.

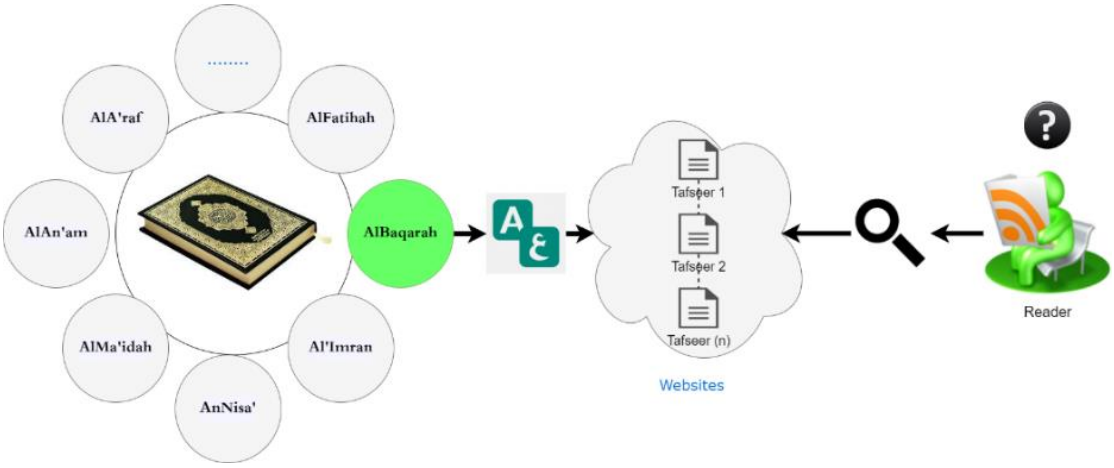


Figure 1. The research motivations

The following study aims to analyze the variations of the k-means cluster algorithm (mini batch *k*-means and *k*-medoids) plus the *k*-means algorithm in terms of five internal cluster validations metrics plus the execution time (ET) for six different translators of 286 verses chapter Al-Baqarah English Tafseer text regarding the NFL theorem. The remaining article is organised as follows: section 2 outlines some of the research related to this work. Section 3 demonstrated the methodology. In Section 4, findings and outcomes are addressed. Section 5 concludes with the report's findings.

2. RELATED WORK

The clustering experiment of [4] employs a mixture of *k*-means clustering variations, bisecting *k*-means and *k*-medoid, in addition to Jaccard and cosine similarity and correlation coefficients, to provide a variety of validation results. However, in the chapter Al-Baqarah clustering process experiments, the best cluster consists of 286 verses formed using cosine similarity with *k*-medoid. Pratama *et al*. [5] classified the Indonesian translations of Hadith text and compared the performance of the fuzzy c-means and *k*-means algorithms using a number of predetermined parameters and term frequency-inverse document frequency (TF-IDF). Silhouette coefficient (SC) and F-measure computations are utilized for clustering validation.

Ahmed *et al.* [6] determined which of the three cluster algorithms (*k*-means, density-based spatial clustering of applications with noise (DBSCAN), and ordering points to identify the clustering structure (OPTICS)) has outperformed others for clustering of chapter Al-Baqarah in terms of the SC and implementation time. DBSCAN has the optimal SC value but gets noise, whereas *k*-means has the fastest implementation time. Moreover, Ahmed *et al.* [6] also analyzed the 286 verses chapter Al-Baqarah clustering process using *k*-means and Mini Batch *k*-means cluster algorithms with TF-IDF. Mini Batch *k*-means algorithm has the highest execution time than the *k*-means algorithm.

Finally, Jansson *et al*. [7] used the *k*-means cluster algorithm and the principle component analysis (PCA) to a group and reduce the huge data of whole-rock, multivariate lithogeochemical. They examined an unsupervised, data-driven methodology for subdividing calcareous marble samples based on evaluating (64) distinct geochemical parameters in whole-rock lithogeochemical data and bright spectrometric data on (181) pieces of dolomitic marble from Sala inlier. PCA is employed to reduce the amount of data, and then *k*-means clustering is used to group samples. The results are then evaluated based on where the groups are located concerning mineral deposits and how well they can be understood using the knowledge of the geological domain.

3. METHOD

Figure 2 shows the sequence procedural for the research that consisted of three main stages (preprocessing, clustering process, and clustering validation). Figure 3 illustrates these stages in detail. The methodology sections consisted of the following.

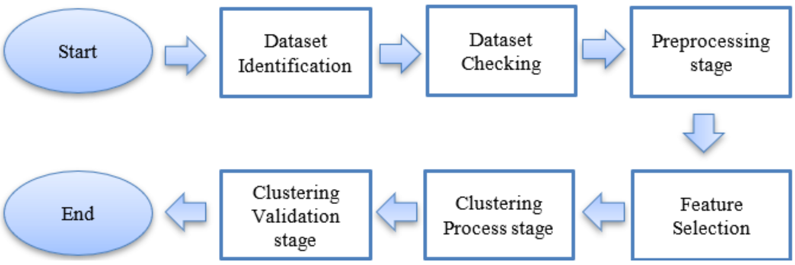


Figure 2. The research sequence operations

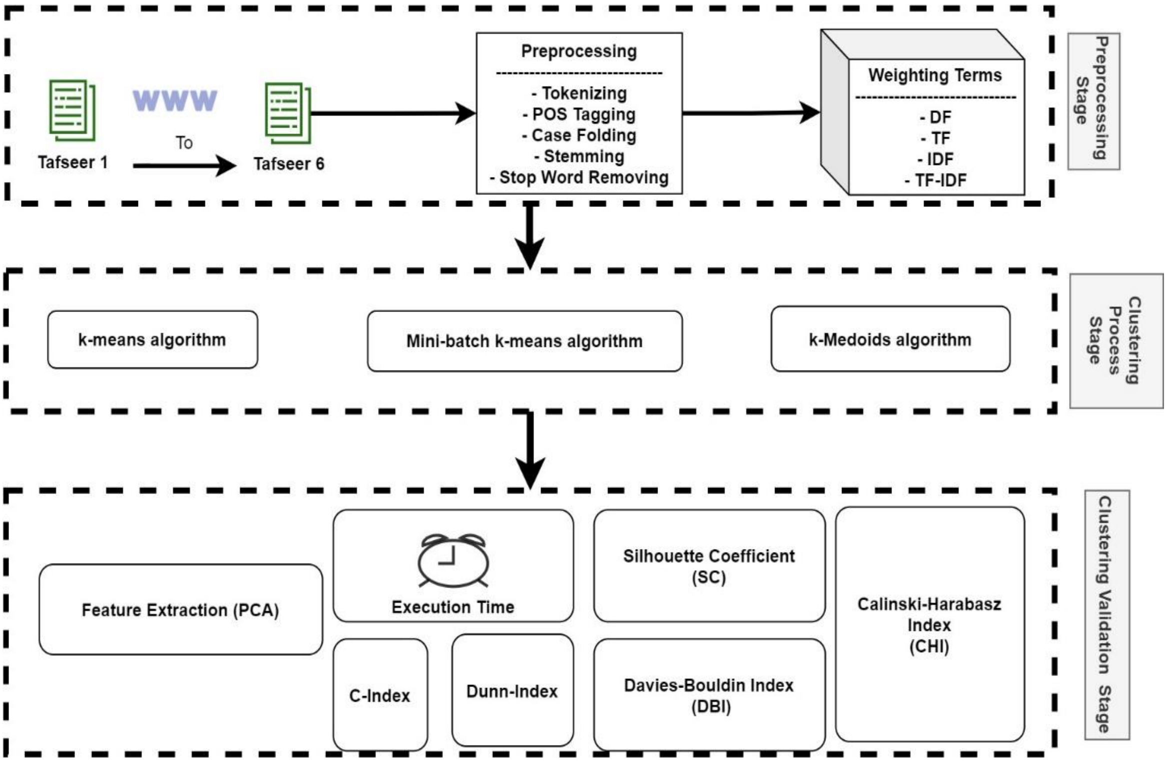


Figure 3. The research methodology

3.1. Dataset identification

The purpose of the study is to cluster the 286 verses in chapter Al-Baqarah English Tafseer text. The webpage (Tanzil) [8] was chosen for this reason. Generally, the following articles collected Qur'anic information text (Tafseer) from Tanzil independent of the languages (e.g., English, Malay, Arabic, and Indonesian) [6], [9]–[13]. Table 1 shows the translator's name and the Tafseer number prepared to use in the research experiments.

3.2. Dataset checking

Figure 4 illustrates the direction of our datasets (T1-T6) corresponding to the research experiments operations. The research datasets contained text data, which is not a standard dataset (the ground truth is unavailable). Therefore, the preprocessing and feature selection must be implemented before the clustering process. The external validation metrics cannot be used (only internal validation metrics can be used) [14].

Table 1. The translator's name, according to Tafseer's number

Tafseer No.	Translator name
T1	A. J. Arberry
T2	Abdul Majid Daryabadi
T3	Abul Ala Maududi
T4	Ali Quli Qarai
T5	Mohammad Habib Shakir
T6	Wahiduddin Khan

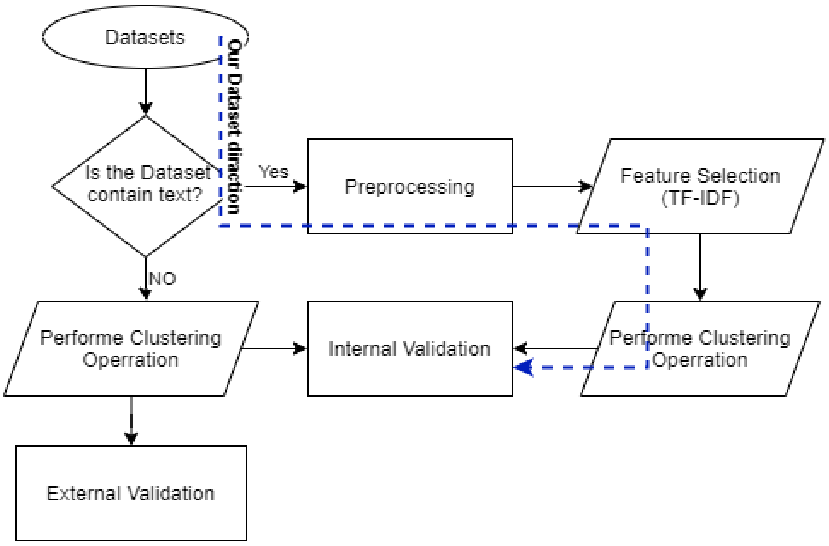

Datasets
Our Dataset direction
Is the Dataset contain text?
Yes
NO
Preprocessing
Feature Selection (TF-IDF)
Performe Clustering Operration
Internal Validation
Performe Clustering Operration
External Validation

Figure 4. The dataset conditions flowchart

3.3. Preprocessing stage

Text preprocessing, sometimes referred to as text cleansing, is the most important principle in text analysis. Usually, unprocessed text datasets comprise noisy and missing data (incomplete characteristics, inconsistent data, random mistakes, and unstructured information. It is essential to do repeated text cleaning since errors are usually detected on the initial attempt. Text data may lead to low data quality, decreasing the mining findings' precision [15]. Hence, text cleansing is a crucial stage since it improves the performance of the feature selection process (the following section) and yields more precise results. The five most popular text preprocessing are tokenization, normalization, stop word removal, part of speech (POS), and stemming [16].

3.3.1. Feature selection

Feature selection aims to transform textual input into a numeric value. The feature selection strategy (term weighting) seeks to omit a group of information that lowers redundancy and enhances target relevance (i.e., the label of the class). Numerous strategies for word weighting are reported in the literature. The vector space model (VSM) remains the most popular model for representing texts (corpus) and is used for these research experiments [5], [14], [15].

VSM calculates TF-IDF. Each document is seen as a vector, and the cell values are given weights based on (1),

$$d_i = \{F_{i,1}, F_{i,2}, \dots, F_{i,j}, \dots, F_{i,t}\} \tag{1}$$

where *t* is the number of features and F_{ij} denotes the weight of feature *j* in document *i*. The next weighting scheme has been used to determine the weight of the feature,

$$F_{i,j} = TF - IDF(i,j) = TF(i,j) \times \left(log \frac{d}{DF(j)}\right) \tag{2}$$

where $TF_{(i,j)}$ denotes the occurrence of feature *j* in document *i*, and *DF*(*j*) denotes all documents containing feature *j*. The VSM (corpus) is identified as a matrix of size *m*×*n* as (3).

$$VSM = \begin{pmatrix} F_{1,1} & F_{1,2} & F_{1,(t-1)} & F_{1,t} \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots \\ F_{(m-1),1} & \cdots & \cdots & F_{(m-1),t} \\ F_{m,1} & F_{m,2} & \cdots & F_{m,t} \end{pmatrix} \quad (3)$$

3.4. Clustering process stage

After the six corpora are created in the previous steps. These corpora are ready as input for the clustering process. The study experiments implemented the three cluster algorithms (partitioned based).

3.4.1. *k*-means

The most common method for partitional clustering is *k*-means clustering [16], [17]. *k*-means clustering is a greedy method and is also called NP-Hard [18]–[20]. It starts by picking K's initial center from a set of representative points. Then, using the chosen proximity metric, each point is given to the closest centroid. After clusters are made, the centers of each cluster are changed and kept updated. The technique then keeps repeating these two steps until the centers cannot expand or until some other relaxed convergence condition is met. The simple structure of the *k*-means algorithm makes it easy to change and construct more efficient strategies on top of it. Numerous ideas have been put forward for changing the *k*-means algorithm. These variations include i) choosing different reflective prototypes for the clusters (*k*-medoids), ii) choosing more precise centroid assumptions (mini-batch *k*-means), and iii) using some kind of feature transformation technique (weighted *k*-means).

The ideal number of clusters used to cluster chapter Al-Baqarah English Tafseer text is seven [4]–[6], [21]. Therefore, (*K*=7) is used for the three algorithms of the research experiments. The following two sub-sections describe two *k*-means variations used in this research experiment to achieve the research objectives.

3.4.2. Mini batch *k*-means

It is a better version of how the *k*-mean algorithm is used. Mini batches are often used to decrease the time it takes to process huge databases. It also tries to improve the results of clustering. In *k*-means optimization, the task is to find the set E of cluster centers $e \in R^m$ that minimizes this objective function using $|E| = k$ over a set XD of samples $xd \in R^m$.

$$Min \sum_{xd \in XD} \| f(E,\ xd) - xd \|) \quad (4)$$

The Euclidean distance between $e \in E$, the centre of the cluster, and *xd* is given by $f(E,\ xd)$.

3.4.3. *k*-medoid

k-medoid is a more reliable method for clustering analysis than *k*-means [22]. Like *k*-means, *k*-medoids try to find a way to cluster items together to reduce a certain objective function. Since the *k*-medoid technique employs the actual data points themselves as prototypes, it is more resistant to the noise in the data as well as any outliers. The *k*-medoids algorithm does not try to reduce the sum of squared errors (SSE) as much as it tries to reduce the absolute error criteria. Like the *k*-means algorithm, the *k*-medoids algorithm goes through a series of steps until the medoid of each cluster is discovered.

In literature, more *k*-means variations algorithms existed and were not implemented in this research. Such of these clustering algorithms are x-means [23], bisecting *k*-means [24], kernel *k*-means [25], and genetic *k*-means [26]. These algorithms can be considered for future work.

3.5. Clustering validation stage

Cluster evaluation often examines the feasibility and effectiveness of applying a clustering algorithm to a dataset and the accuracy of the clustering returns generated by the algorithm. Since the research datasets have no ground truth (labels), only internal validation metrics [27] could be applied. The execution time and five internal validation metrics used for the research experiments are as follows.

3.5.1. Execution time

The execution time of the clustering algorithms is related to the computer hardware. The hardware used for the experiment was an Intel Core i7-8550U CPU running at 1.80 GHz with 8 GB of RAM. The software used was Microsoft Windows 10 operating system and Python (3.7.7) platform.

3.5.2. Internal validation metrics

Since some datasets have no ground truth, researchers must rely on an internal validation approach to evaluate the clustering quality. Usually, intrinsic techniques assess clustering by analyzing the clusters' compactness and separation. Many validation metrics utilize a similarity metric to evaluate the similarity between objects and datasets. Internal validation contrasts the clustering assessment with the real outcome, which is the structure of the found clusters and their relationships to each other. Table 2 shows the five internal validation metrics used in this research. The benefit criterion type is referred to (the greater the value, the better) and the costly criterion type (the less the value, the better).

3.5.3. PCA of feature extraction

Using the PCA, the data from the original space is transformed into a lower-dimensional space unrelated to the original space's properties by using the PCA [17]. Thus, the research experiments utilized the TF-IDF through the preprocessing stage, whilst the PCA was only employed for clustering visualization (2D) purposes [28], [29].

Table 2. The five internal validation metrics used

Metrics	Criterion Type	Definition
Silhouette index (SC) [30]	Benefit	$\frac{1}{NC}\sum_{i}\left\{\frac{1}{n_i}\sum_{x\in C_i}\frac{b(x)-a(x)}{max[b(x),a(x)]}\right\}$ (4)
Calinski-Harabasz index (CHI) [31]	Benefit	$\frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x\in C_i} d^2(x,c_i)/(n-NC)}$ (5)
Dunn's indices (Dunn-index) (DI) [32]	Benefit	$\min_{i}\left\{\min_{j}\left(\frac{\min_{x\in C_i,y\in C_j} d(x,y)}{\max_{k}\left\{\max_{x,y\in C_k} d(x,y)\right\}}\right)\right\}$ (6)
C-Index (CI) [33]	Benefit	$\frac{\delta_w-min(\delta_w)}{max(\delta_w)-min(\delta_w)}$ (7) $\delta_w=(R(U),D)=\sum_{j=i+1}^{n}\sum_{i=1}^{n-1}r_{ij}(U)d_{ij}=\sum_{j=i+1}^{n}\sum_{i=1}^{n-1}\left[\sum_{k=1}^{c}u_{ki}.u_{kj}\right]d_{ij}$
Davies-Bouldin index (DBI) [34]	Costly	$\frac{1}{NC}\sum_{i}\max_{j,j\neq i}\left\{\left[\frac{1}{n_i}\sum_{x\in C_i}d(x,c_i)+\frac{1}{n_j}\sum_{x\in C_j}d(x,c_j)\right]/d(c_i,c_j)\right\}$ (8)

4. RESULTS AND DISCUSSION

This section shows the outputs of the methodology section. The section provides an analysis of the output results. It consisted of the following two subsections.

4.1. Feature statistics

The number of features that were presented both before and after the text cleansing process is detailed in Table 3, along with the top three most frequent features. Also, the table showed the common words and features ("Allah" or "God, "believe", "said", "shall, "into", "ye", and "people") that were used often. These features can be found (shared) in more than one Tafseer (T).

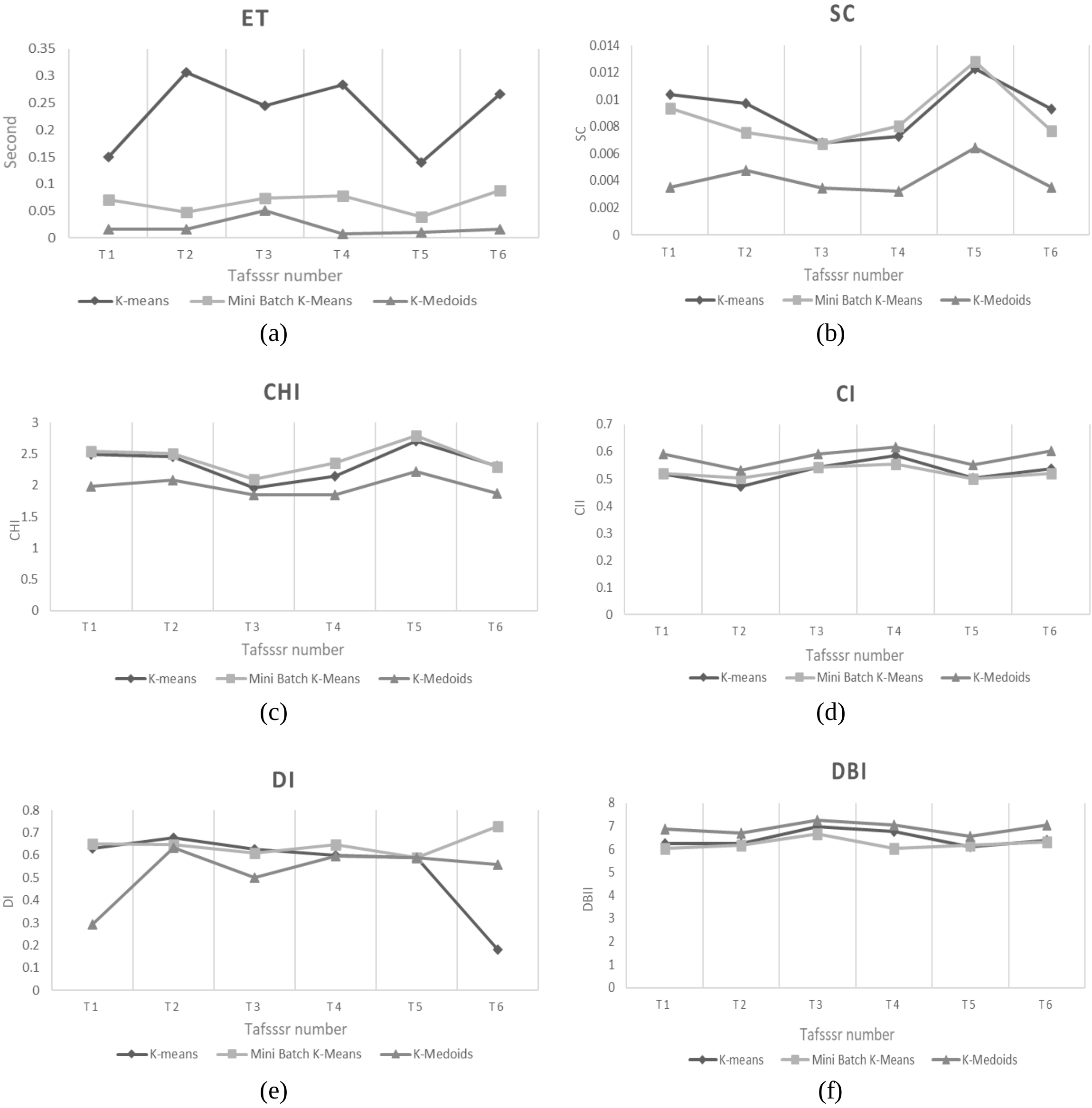
Table 3. The features statistics results

Tafseer No.	Total Features	Stop Words Removal	Stemming	First Three Frequent Features
T1	1466	1278	1033	God, shall, believe
T2	1591	1362	1198	Allah, into, ye
T3	2008	1775	1388	Allah, people, shall
T4	1591	1392	1254	Allah, said, say
T5	1563	1332	1060	Allah, shall, sure
T6	1705	1493	1183	God, shall, believe

4.2. Clustering validation

The result outputs of the three clustering algorithms are presented in Figure 5. In terms of the ET, Figure 5(a) proved that *k*-medoids are faster than Mini Batch *k*-means and the *k*-means. In contrast, Figure 5(b) proved the SC of *k*-means is better than Mini batch *k*-means and *k*-medoids. Figure 5(c) presented the CHI of mini-batch *k*-means performed better than *k*-means and *k*-medoids. Figure 5(d) showed *k*-medoids superior to the other two algorithms, while the other two algorithms performed approximately the same. Figure 5(e) demonstrates there is no one algorithm performed better than others. Finally, Figure 5(f) illustrates the CHI of mini batch *k*-means performed better than others. Tables 4 to 6 display the results of the clustering validation section for the three algorithms in detail. Moreover, Figure 6 displays the seven clusters allocated by each algorithm using PCA for Tafseer (T1-T3). Figure 7 shows these clusters for the reset Tafseer (T4-T6).

Hence, we can rank the performance of the three algorithms using ET only. However, the five cluster validation metrics proved there is no one algorithm outperformed the others for the six Tafseer datasets. Therefore, the research experiments confirm the NFL theorem assumption.



(a) (b)

(c) (d)

(e) (f)

Figure 5. The clustering validations diagrams for three algorithms according to (a) ET, (b) SC, (c) CHI, (d) CI (e) DI, and (f) DBI

Table 4. *k*-means

Tafsser No.	ET	SC	CHI	CI	DI	DBI
T1	0.15055	0.01037	2.49427	0.51684	0.63141	6.24313
T2	0.30618	0.00973	2.45939	0.47029	0.67927	6.253
T3	0.24534	0.00678	1.9668	0.54389	0.62583	6.97075
T4	0.28424	0.00729	2.1466	0.58458	0.5981	6.75686
T5	0.14057	0.01225	2.70327	0.50306	0.59065	6.09906
T6	0.26629	0.00927	2.30578	0.53672	0.18091	6.38556

Table 5. Mini batch *k*-means

Tafsser No.	ET	SC	CHI	CI	DI	DBI
T1	0.07076	0.00936	2.54523	0.52098	0.64932	6.02506
T2	0.0478	0.00754	2.50234	0.50207	0.64674	6.17125
T3	0.07373	0.00671	2.1036	0.54168	0.60871	6.68228
T4	0.07775	0.00803	2.35974	0.55301	0.64866	6.04956
T5	0.0399	0.0128	2.78878	0.49852	0.59065	6.18554
T6	0.08772	0.00767	2.29872	0.51938	0.72997	6.30662

Table 6. *k*-medoids

Tafsser No.	ET	SC	CHI	CI	DI	DBI
T1	0.01559	0.00352	1.99045	0.59095	0.29264	6.87424
T2	0.01558	0.00477	2.08027	0.53027	0.63353	6.69356
T3	0.05111	0.00344	1.84291	0.59053	0.50117	7.27606
T4	0.00794	0.00322	1.84866	0.61708	0.59652	7.06599
T5	0.01097	0.00644	2.22286	0.55116	0.59065	6.56997
T6	0.01562	0.00352	1.86907	0.60241	0.55833	7.06202

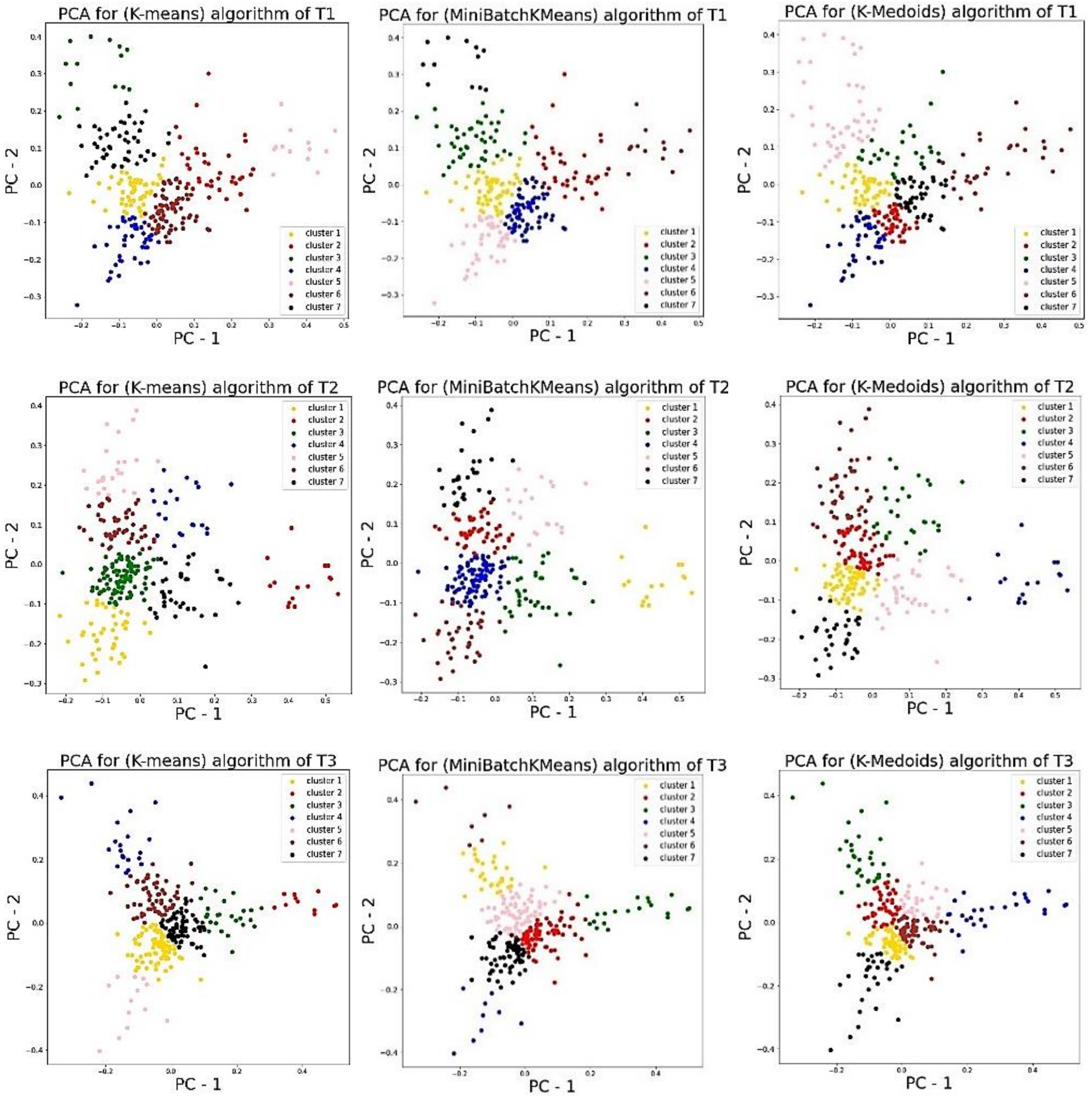


Figure 6. The seven clusters' diagrams of (T1-T3) for the three algorithms

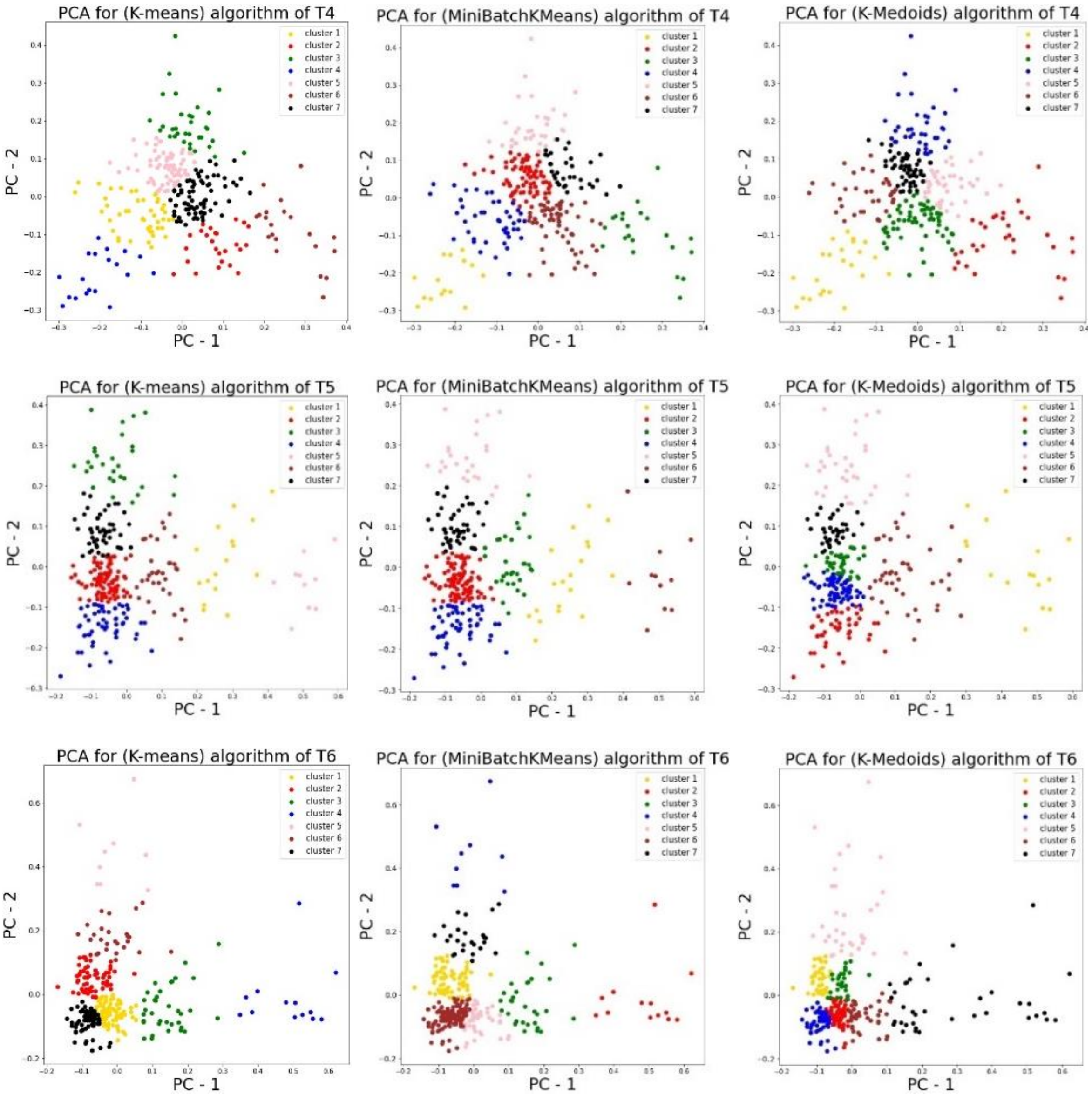


Figure 7. The seven clusters' diagrams of (T4-T6) for the three algorithms

5. CONCLUSION

Three stages were implemented in this study. At the first stage (the preprocessing), six of 286 verses of chapter Al-Baqarah English Tafseer text documents related to six translators represented the input datasets. Tokenization, normalization, stop-word removal, part of speech (POS), and stem are used at this stage. Moreover, TF-IDF is used to provide the weighting term. In the second stage (clustering process), three cluster algorithms (*k*-means) and two *k*-means variations (Mini Batch *k*-means and *k*-medoids) were executed. The number of clusters used was (K=7) obtained from the literature for all six datasets. In the third and final stage (clustering validation stage), five internal cluster validation metrics (SC, CHI, CI, DI and DBI) are used, and the ET is calculated. Moreover, at this stage, PCA was employed to present and visualize the seven clusters' output datasets/algorithms.

The research aim is to analyze the *k*-means cluster algorithms' variations behaviors according to the six input datasets that relate to the same topic or theme. The results output proved the *k*-medoids algorithm outperformed the two others in the ET. Moreover, the experiments demonstrated that none of the three algorithms outperformed the others in clustering validation. However, the research conclusion confirmed the NFL theorem.

In future works, the authors suggest that more *k*-means variations can be implemented. Expanding the dataset number to include more than six and analyzing the results. Expanding the number of internal cluster validation metrics.

ACKNOWLEDGEMENTS

The authors thank the sponsor and financial support of UKM and Al-Iraqia Universities.

REFERENCES

[1] C. Zong, R. Xia, and J. Zhang, *Text data mining*. Singapore: Springer Singapore, 2021, doi: 10.1007/978-981-16-0100-2.

[2] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, Apr. 1997, doi: 10.1109/4235.585893.

[3] M. Alkhatib and K. Shaalan, "The key challenges for Arabic machine translation," in *Intelligent Natural Language Processing: Trends and Applications*, Springer, 2018, pp. 139–156, doi: 10.1007/978-3-319-67056-0_8.

[4] A. F. Huda, M. R. Deyana, Q. U. Safitri, W. Darmalaksana, U. Rahmani, and Mahmud, "Analysis partition clustering and similarity measure on Al-Quran verses," in *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, Jul. 2019, pp. 1–5, doi: 10.1109/ICWT47785.2019.8978215.

[5] R. S. Pratama, A. F. Huda, A. Wahana, W. Darmalaksana, Q. U. Safitri, and A. Rahman, "Analysis of fuzzy C-Means algorithm on Indonesian translation of Hadits text," in *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, Jul. 2019, pp. 1–5, doi: 10.1109/ICWT47785.2019.8978264.

[6] M. A. Ahmed, H. Baharin, and P. N. E., "Analysis of k-means, DBSCAN and OPTICS cluster algorithms on Al-Quran verses," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 248–254, 2020, doi: 10.14569/IJACSA.2020.0110832.

[7] N. F. Jansson, R. L. Allen, G. Skogsmo, and S. Tavakoli, "Principal component analysis and k-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates, Sala area, Sweden," *Journal of Geochemical Exploration*, vol. 233, Feb. 2022, doi: 10.1016/j.gexplo.2021.106909.

[8] Tanzil, "Quran navigatorle," Tanzil, http://tanzil.net/trans/ (accessed Nov. 17, 2022).

[9] A. B. Muhammad, "Annotation of conceptual co-reference and text mining the Qur'an," PhD thesis, University of Leeds, 2012.

[10] C. Slamet, A. Rahman, M. A. Ramdhani, and W. Darmalaksana, "Clustering the verses of the holy Qur'an using k-means algorithm," *Asian Journal of Information Technology*, vol. 15, no. 24, pp. 5159–5162, 2016, doi: 10.3923/ajit.2016.5159.5162.

[11] S. K. Hamed and M. J. A. Aziz, "A question answering system on holy Quran translation based on question expansion technique and neural network classification," *Journal of Computer Science*, vol. 12, no. 3, pp. 169–177, Mar. 2016, doi: 10.3844/jcssp.2016.169.177.

[12] N. Suryana, F. S. Utomo, and M. S. Azmi, "Quran ontology: Review on recent development and open research issues," *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 3, pp. 568–581, 2018.

[13] M. I. Rahman, N. A. Samsudin, A. Mustapha, and A. Abdullahi, "Comparative analysis for topic classification in Juz Al-Baqarah," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, pp. 406–411, Oct. 2018, doi: 10.11591/ijeecs.v12.i1.pp406-411.

[14] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *International Journal of computers and communications*, vol. 5, no. 1, pp. 27–34, 2011.

[15] G. Forman and E. Kirshenbaum, "Extremely fast text feature extraction for classification and indexing," in *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, 2008, p. 1221, doi: 10.1145/1458082.1458243.

[16] Z. Zainol, P. N. E. Nohuddin, T. A. T. Mohd, and O. Zakaria, "Text analytics of unstructured textual data: A study on military peacekeeping document using R text mining package," in *Proceedings of the 6th International Conference on Computing and Informatics, ICOCI 2017*, Art. no. 019, 2017, pp. 1–7.

[17] C. C. Aggarwal and C. K. Reddy, *Data clustering*. Citeseer, 2014.

[18] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.

[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14, pp. 281–297.

[20] B. Bansal and S. Srivastava, "Hybrid attribute based sentiment classification of online reviews for consumer intelligence," *Applied Intelligence*, vol. 49, no. 1, pp. 137–149, Jan. 2019, doi: 10.1007/s10489-018-1299-7.

[21] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Mini-Batch k- means versus k- means to cluster English Tafseer text: view of Al-Baqarah chapter," *Journal of Quranic Sciences and Research*, vol. 2, no. 2, pp. 48–53, 2021.

[22] B. Mirkin, *Clustering for data mining: a data recovery approach*. Chapman and Hall/CRC, 2005.

[23] D. Pelleg, A. W. Moore, and others, "X-means: Extending k-means with efficient estimation of the number of clusters," in *ICML*, 2000, vol. 1, pp. 727–734.

[24] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," Technical Report, 2000. [Online], Available: https://www.stat.cmu.edu/~rnugent/PCMI2016/papers/DocClusterComparison.pdf

[25] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998, doi: 10.1162/089976698300017467.

[26] K. Krishna and M. Narasimha Murty, "Genetic k-means algorithm," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 3, pp. 433–439, Jun. 1999, doi: 10.1109/3477.764879.

[27] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 982–994, Jun. 2013, doi: 10.1109/TSMCB.2012.2220543.

[28] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, Sep. 1933, doi: 10.1037/h0071325.

[29] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[30] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[31] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[32] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, Jan. 1974, doi: 10.1080/01969727408546059.

[33] E. C. Dalrymple-Alford, "Measurement of clustering in free recall.," *Psychological Bulletin*, vol. 74, no. 1, pp. 32–34, Jul. 1970, doi: 10.1037/h0029393.

[34] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

BIOGRAPHIES OF AUTHORS

Mohammed A. Ahmed received a B.Eng. degree in computer and software from Al- Mustansiria University, Iraq, in 2003 and an M.Sc. degree in computer science from the University of Malaya, Malaysia, in 2012. Currently, he is a Ph.D. student at Universiti Kebangsaan Malaysia and works as a lecturer at the Network Engineering Department, College of Engineering, Al-Iraqia University, Baghdad, Iraq. His research interests include information security (applied cryptography and steganography), network security, information technology, text and voice recognition, data mining, text mining, multi-criteria decision-making, machine learning and clustering. He can be contacted at email: mohammed.abdalmunam@aliraqia.edu.iq.

Hanif Baharin has a Ph.D. in interaction design from The University of Queensland, Australia. He is currently a research fellow and a senior lecturer at the Institute of IR4.0, Universiti Kebangsaan Malaysia. His research interests include human-computer interaction, interaction design, and the intersections between computer science, AI, and the arts. He can be contacted at hbaharin@ukm.edu.my.

Puteri Nor Ellyza Nohuddin received her BSc. in Computer Science from the University of Missouri-Columbia, USA, and her MSc in IT from Universiti Teknologi MARA. In 2012, she was awarded her Ph.D. in Computer Science from the University of Liverpool, UK. She is a research fellow at the Institute of IR 4.0, Universiti Kebangsaan Malaysia. Currently, she is seconded to the Higher Colleges of Technology in Sharjah, UAE. Her primary research interests are in the field of big data, data mining, machine learning, software engineering and knowledge engineering. She can be contacted at puteri.ivi@ukm.edu.my.