Prioritise Five Tafseer Translators Using Clustering Technique for Surah Al-Baqarah

Mohammed A. Ahmed[*], Shahad Mahgoob Nafl[**], Hanif Baharin[***], Puteri Nor Ellyza Nohuddin[****]

[*] Network Engineering Department, College of Engineering, Al-Iraqia University, 10053, Baghdad, Iraq
Email: mohammed.abdalmunam@aliraqia.edu.iq
https://orcid.org/0000-0002-3456-2533

[**] College of Medicine, University of Baghdad, Iraq
Email: shahad.m@comed.uobaghdad.edu.iq
https://orcid.org/0000-0001-8746-1587

[***] Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Malaysia
Email: hbaharin@ukm.edu.my
https://orcid.org/0000-0003-1474-0344

[****]Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Malaysia
Email: puteri.ivi@ukm.edu.my
https://orcid.org/0000-0003-0627-5630

Abstract

The English Tafseer Translation of the Holy Quran is essential for comprehending and interpreting Allah's words for non-Arabic Muslims. This research adopted five different English translators (TR1-TR5) of chapter (Surah) Al-Baqarah and invested the advantages of the text clustering process to rank (prioritise) between these input five datasets. The absence of dataset ground truth (not standard datasets) requires the use of unsupervised learning (clustering technique) instead of other techniques (e.g. classification (supervised learning)). This study expanded the assessment to include both partitioning-based and hierarchical-based clustering algorithms. In a cluster based on partitioning, k-means is utilized. While for the hierarchical-based, the Agglomerative has been implemented. This research's aim was achieved through a three-step procedure (stages). The first stage uses text cleansing to remove unnecessary words (Tokenisation, POS tagging, normalisation, stemming, and Stop-word removal). In addition, feature selection used VSM (Vector Space Model) and TF-IDF (Term Frequency-Inverse Document Frequency) to make the five corpora. The second stage implemented the clustering process. In the third stage, clustering validation was obtained using SC (Silhouette Coefficient) and DBI (Davies-Bouldin Index) metrics plus the execution time (ET). Principle Component Analysis (PCA) is used to visualise the clustering outputs. The results show, based on (ET, SC, and DBI) of the k-means algorithm, only ranks (1) and (3) demonstrate the same ranking for these five translators. In contrast, the Agglomerative algorithm shows the same five translators' positions; each (ET, SC, and DBI) has a distinct rank. However, to obtain the optimal union rank, it is crucial to use a modern approach technique such as MCDM (Multi-Criteria Decision-making Analysis) in future work.

Keywords- k-means algorithm, Agglomerative algorithm, Text clustering, TF-IDF, VSM, PCA, Al-Baqarah chapter.

I. INTRODUCTION

The Holy Quran is humanity's book of divine instruction and guidance. It is the primary religious text of Islam, around the word of God. Unlike many other world books, the Holy Quran stays preserved and original. Translating the Holy Quran requires a greater understanding to discover the majestic beauty of the Quran's message when translating it into English. However, there are many syntactic differences between Arabic and English languages. This study deals with digital documents of the English Holy Quran (Tafseer) [1].

The Al-Quran is considered to be a divine message from Allah, which was revealed and transmitted to Prophet Muhammad. It has been safeguarded by Allah since the time of the Prophet SAW till the present day. In the Holy Quran, God says that he will protect the Quran's book: "Surely We revealed the Book, and surely we are the protectors". It is taught and memorised by millions of Muslims worldwide in the same form as revealed [1], [2]. Surah Al-Baqarah is the Holy Quran's second Surah (chapter), the largest consisting of 286 verses, and is considered the Quran's longest chapter. The Holy Quran consists of many chapters, all of which are divine and origin miraculous. Some of the Holy Quran's chapters' have significance for someone else; for some reason, chapter Al-Baqarah is one of those special chapters [3]. Chapter Al-Baqarah includes 53 topics, where some topics remain on the same subject as others

(similar). The verses of the same issue are then grouped so that seven significant topics include (oneness and the power of Allah, warning to the Bani Israel, Kaaba as Qiblaa, Shari'ah, the messengers, wealthy, and three groups of people) [4]. Figure 1 shows the percentages of these seven topics.

There are several syntactic variations between Arabic and English when translating the Holy Quran into English. Verb tense is a syntactic challenge that translators typically face with the Holy Quran. Moreover, there is a transition to the imperfect verb from the past tense verb to achieve an effect that may cause problems and difficulties in translation [5]. Full comprehension of the Holy Quran and its translation for non-Arabic-speaking Muslims may be a challenge. However, today's digital documents are written by many translators, and Tafseer is available online through websites for each chapter of the Holy Quran, as shown in Figure 2. Therefore, it is essential to produce a method to automatically classify or cluster these digital documents by ranking them to select the best one using the most recent efficient techniques available for text mining [6]–[14].

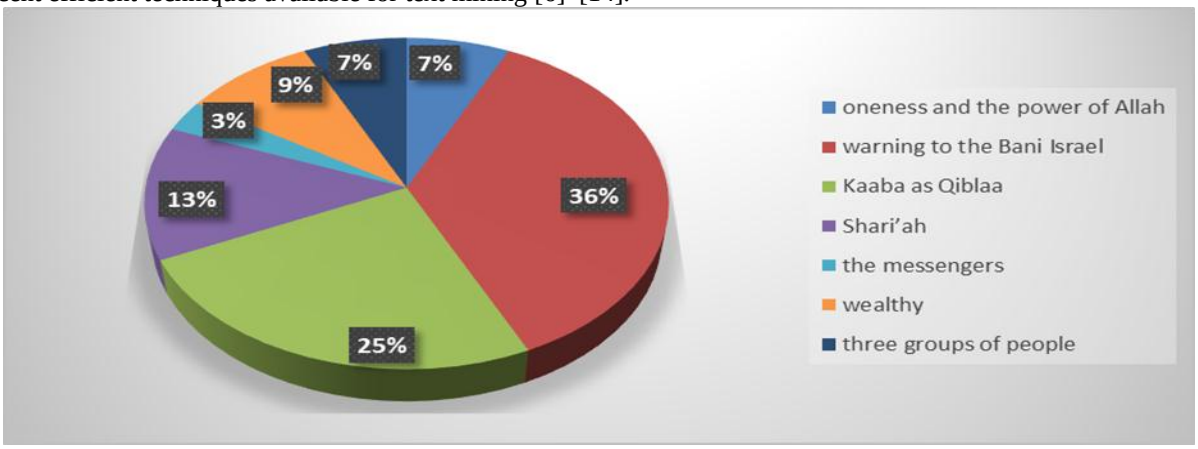


Figure 1. Chapter Al-Baqarah Seven themes.

Text mining is a common way of revealing meaningful knowledge from a collection of text; classification, clustering, regression, frequent mining patterns, and association rule techniques may be used. The absence of ground truths of the datasets in actual world experiments provides a challenge that requires the use of unsupervised learning (clustering technique) instead of other techniques (e.g. classification (supervised learning)) [15]. Therefore, this research adopted clustering techniques due to the lack or absence of the standard ground truths (labels) of this research dataset [16].

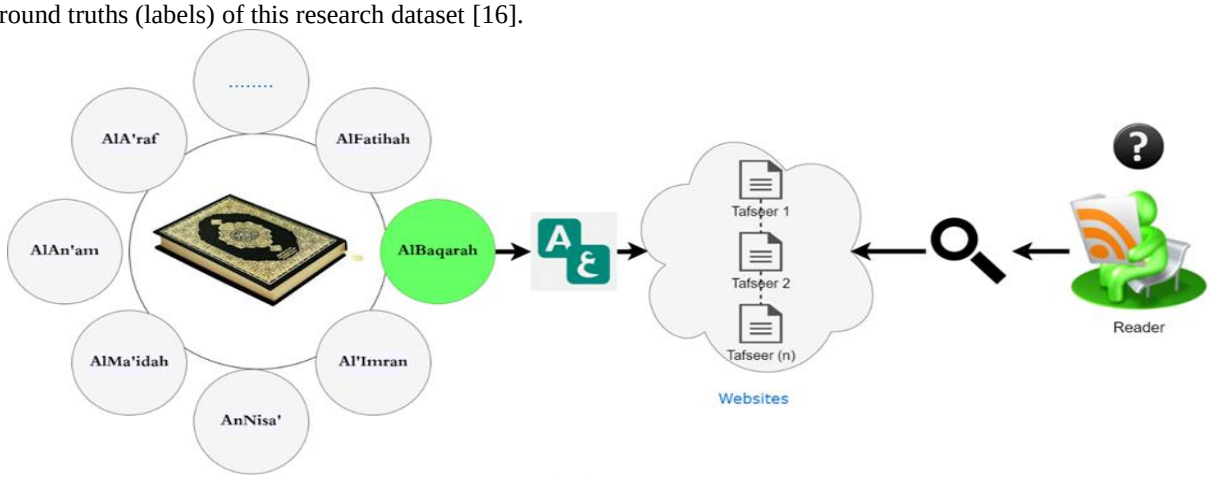


Figure 2. The Digital documents of Tafseer are available online (reader challenges).

External validation is a clustering assessment that compares the clustering outcome with a reference outcome known to be ground truth. Still, it has fundamental caution because, in most actual applications, the reference results or the ground truth are not given. Internal validation is the other clustering category assessment, where the clustering assessment is only compared to the results themselves. Furthermore, this is much more practical and effective in many real-world situations since it does not apply to any presumed external sources that can not always be provided. This article focused on internal clustering validation and studied its functionality for cluster algorithms' outputs due to previous reasons [17].

This research has extended the analysis to include both partitioning-based and hierarchical-based clustering algorithms [10]. In a partitioning-based cluster, the k-means [18] is used. While hierarchical-based, the Agglomerative [19] has been executed. Finally, the main aim of this research is to invest in the advantages of the clustering process to prioritise related datasets (the datasets have the same number of partitioning or clustering, e.g. (k=7)) of five different translators of Al-Baqarah chapter (ranking).

II. RELATED STUDIES

Table 1 shows the literature for many references concerning Quranic translation sources, the techniques or algorithms used to implement the clustering processes, and a brief description of each reference.

Table 1. The study literature related to the Quranic dataset and the cluster technique employed.

The Reference	The Quranic DataSet	What techniques were used to achieve the aim	Brief Description
[20]	The authors used two groups of datasets acquired from [1] of this paper. From 222 concepts defined from the ontology, the authors clustered them into 77, 24, and 6 concepts for Person, Location, and Time, respectively.	• TF-IDF • cosine similarity method	The authors created a QAS (Question-Answering System) for the Indonesian translation of the Quran that is based on semantics. In order to feed the semantic interpreter on the user query, researchers first propose three questions to the user and construct a weighted vector (TF-IDE) that includes each concept that belongs to the appropriate expected answering type (which is also referred to as named entity group).
[21]	Al-Qur'an in English translation data written by (Ahmed Ali), published by http://tanzil.net as a corpus	Partitioning method (k-means). Where ***k*=3.**	This study resulted in an initial practical step in which the structures of verses in the Holy Quran are clustered. The algorithm is utilized to cluster 6236 total verses, using unsteamed and steamed words that establish three clusters.
[22]	• Quranic Ontology from Quranic Arabic Corpus from (http://corpus.quran.com/) • (Muhammad Quraish Shihab) writes Tafseer in the Indonesian language	•TF-IDF	This paper presents work in generating a Weighted Vector for each concept in the Indonesian Translation of the Quran (ITQ) and implementing a semantic-based question-answering system (QAS) similar to the [20] reference. However, the author shows more details about TF-IDE results and has different work procedures.
[23]	http://tanzil.net/trans/ contains the translation of the Quran in various languages, including Indonesian. (Muhammad Quraish Shibab, and Jalal ad-Din al- Mahalli and Jalal ad-Din as-Suyuti) same as the Slamet et al. (2016)	No technique was listed except annotation is done in three layers: Syntax, Referential structure and Semantics annotations.	This paper presented a model to do semantic annotation on the Indonesian translation of the Quran corpus.
[24]	Collect Frequently Asked Questions (FAQs) about the Quran from: (i) Four web resources, (ii) questions and their answers were gathered from Islamic experts at the Holy Mosque in Mecca	Probabilistic clustering methods	Create an integrated Quran question and answer corpus, then cluster and visualize this corpus using WEKA free Java software. The probabilistic clustering method has clustered the corpus into four clusters.
[25]	Six short chapters (surah) of 130 keywords were selected from the Malay-translated Tafseer of Al-Quran for the experiment.	•TF-IDF • network analysis map (hierarchical methods)	This paper introduces a combination of text mining (TF-IDF) and network analysis (map) approaches to extract keywords and identify relationships between keywords and chapters of Tafseer. The proposed method is called the KCRA framework.
[26]	Surah **Al-Baqarah** of 286 documents as verses derived from the translation of the Qur'an	•TF-IDF •similarity measures (cosine, Jaccard, correlation	The clustering experiment uses a combination of K-means clustering techniques, bisecting K-means, and k-medoid, along with cosine similarity, Jaccard similarity and correlation coefficients to produce various validation values. However, the optimal cluster results in the

	in English	coefficient) •Clustering Validity (DBI) •k-means •Bisecting k-means •k-medoid k-means •Where ***k*=7**	Surah **Al-Baqarah** clustering process with 286 verses produced by k-medoid with cosine similarity.
[27]	Indonesian Translation of the Quran	•TF-IDE, •Cosine Similarity algorithm •Single Pass Clustering (SPC) Algorithm	This research aims to develop a web-based verse search system (information retrieval) for Al-Qur'an, which is integrated with a clustering algorithm (SPC) to help Muslims discover the relevant information in Quran verses by grouping the Quran verses into their own similar group.
[28]	Indonesian Translation Hadiths, The hadith chapter used for the experiment is chapter wudhu, salat, zakat, haji, and shaum. The amount of data used is 368 hadiths, with five documents	• TF-IDF •k-means •Fuzzy C-means	This research grouped the Indonesian translation of Hadith texts and compared the performance of k-means and Fuzzy C-means algorithms with some determined parameters and experiments. Silhouette Coefficient and F-measure calculations are used for clustering validation.
[29]	Surah **Al-Baqarah** of 286 verses written by (Ahmed Ali) derived from the translation of the Qur'an in English Tafseer collected from http://tanzil.net/trans/	•TF-IDF •SC •k-means •DBSCAN •OPTICS •Where ***k*=7**	The paper aims to find out which one of the three cluster algorithms has outperformed the others for clustering the Surah **Al-Baqarah** English Tafseer. The DBSCAN algorithm has a great value SC score compared to OPTICS and K-means algorithms. However, the DBSCAN algorithm got some clustering noise.
[30]	Surah **Al-Baqarah** of 286 verses written by (Muhammad Sarwar) derived from the translation of the Qur'an in English Tafseer collected from http://tanzil.net/trans/	•TF-IDF •k-means •Mini Batch k-means •Where ***k*=7**	The study proved the Mini Batch K-means cluster algorithm outperforms the standard k-means algorithm in terms of the execution time of clustering (faster) for Surah **Al-Baqarah** English Tafseer (where the number of clusters was seven).

III. METHOD

The method used to achieve the aim of this research consists of three main stages, as shown in Figure 3. These stages are as follows:

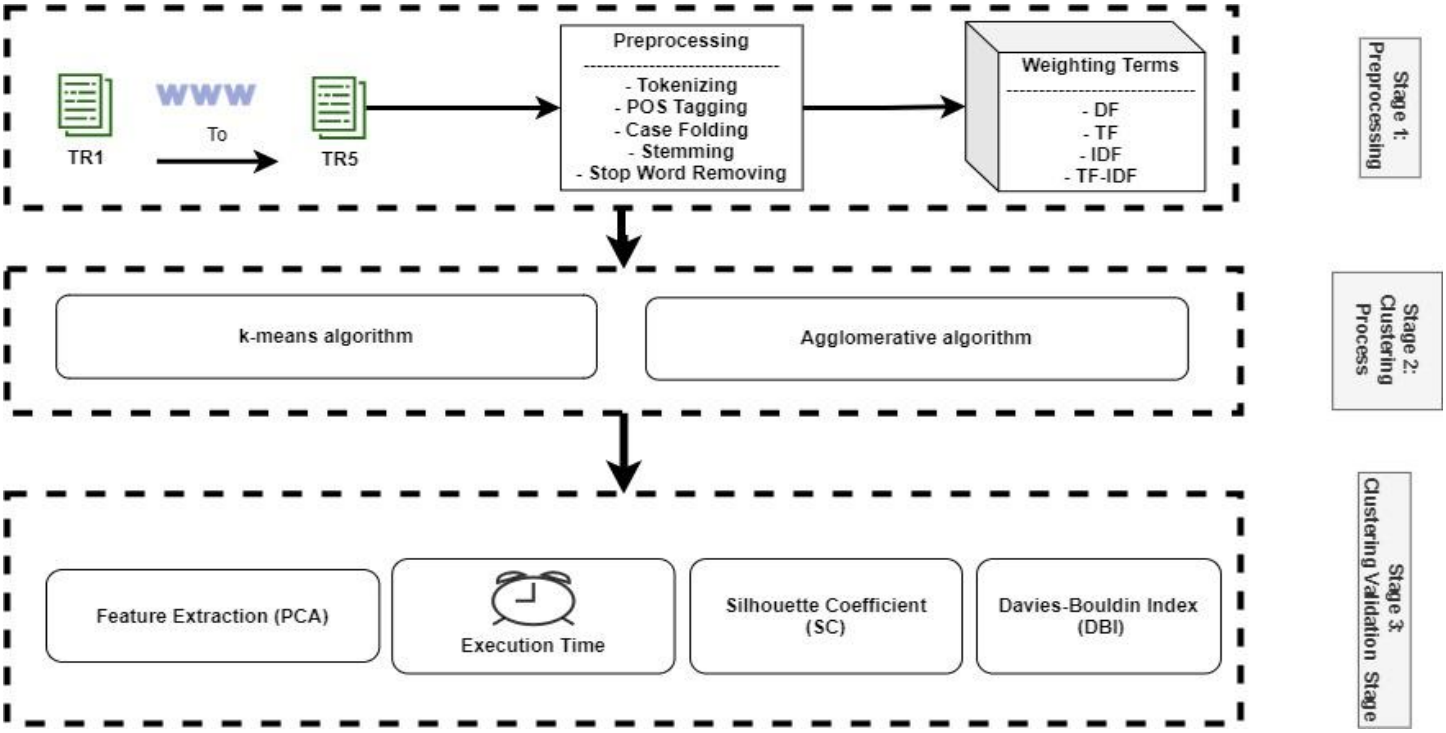


Figure 3. The methodology's three stages.

A. Datasets (Stage 1)

1. Dataset Collecting

The databases to be evaluated in this paper have been obtained from the database of (http://tanzil.net/trans/). The website includes a text explaining the translation of the Quran across different languages, including English, with several translators who translated from various Tafseer. Many researchers make use of the information gathered from this website, such as [25], [31] [32], [33]. Hence, the proposed method adapted the English text Tafseer chapter Al-Baqarah of five different translators, as shown in Table 2.

Table 2. The selected translator's name and TR number.

Translator No.	Translator name
TR1	Abul Ala Maududi
TR2	Ali Quli Qarai
TR3	Hasan al-Fatih Qaribullah and Ahmad Darwish
TR4	Muhammad Sarwar
TR5	Mohammad Habib Shakir

2. Dataset Identification

In the study proposed method, before implementing the clustering process, the input dataset should be checked to determine whether it is a corpus or not. For example, If the dataset is selected from [34], such as (e.g. Iris), Iris is a corpus already. Therefore, the implementation of the clustering process will be direct (the preprocessing operation is unnecessary), and both internal and external cluster validation metrics can be used.

However, if the dataset is not a corpus (e.g. the study dataset (TR1-TR5)), then the corpus must be created. The study datasets are a text where each Tafseer verse is represented as a row in the document. Hence, the text preprocessing operation and feature selection executions will be mandatory (e.g. Tokenisation and TF-IDF). Moreover, in this situation, the external cluster validation metrics cannot be used (because no ground truth is available). Figure 4 illustrates the study dataset's flowchart with conditions and the study input dataset's direction (TR1-TR5). Additionally, the study dataset website provides more information about the study dataset's reliability, authenticity and potential variations (transparency).

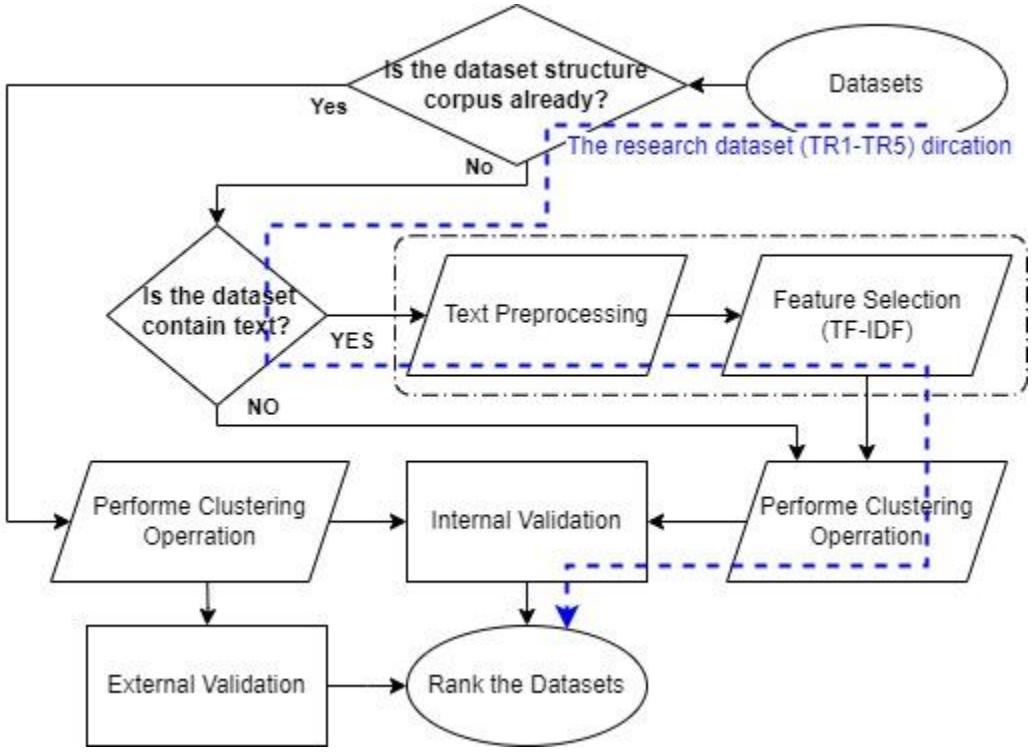


Figure 4. The study dataset preprocessing conditions flowchart.

B. Text Preprocessing (Stage 1)

Text preprocessing or text cleansing is essential in all types of text analysis. Typically, the set of raw text data is cluttered with errors, missing data, and noisy data (error-filled, incomplete data). The method of text cleaning is iterative because new errors will still be found before it is completed. According to [35], the text data may be in very bad conditions for text mining. The researchers recommend employing a text cleaning technique, as it helps to increase the text mining performance and delivers an accurate result. Until the clustering algorithms can be executed, the text documents must first be cleaned up. The basic method for text preparation is used for texts, which consists of the following steps: -

1) Tokenisation: Text data is split into an individual unit basic sequence.
2) Part-of-speech (POS) tagging is a linguistic technique employed to assign a specific grammatical category to each word in a given text.
3) Normalisation or case folding: Conversion to the same case of all text document characters, either the upper case or the lower.
4) Removal of the stop word: Remove those familiar words that most frequently happened, e.g. 'are', 'is', 'an', 'that', etc.
5) Stemming is a linguistic process that aims to regulate all of the variations of a given keyword by reducing them to a common root form. An instance of keyword grouping can be observed in the context of similar terms with varying spellings, such as 'training' or 'trains' and 'trained' being consolidated under the common term 'train'.

1. Feature selection (Weighting Terms)

Reducing dimensionality is one of the common noisy and redundant attribute removal methods (aka features). Techniques for the reduction of dimensionality can primarily be categorised in the extraction and selection of features. The technique of feature selection, namely term weighting, aims to exclude a particular collection of data that decreases redundancy and enhances the relevance of the target variable, which is the class label. The objective of term weighting is to convert textual data into a numerical representation. There are various methodologies for assigning weights to terms that have been documented in the academic literature. The vector space model (VSM) is frequently employed for the representation of written content [36] [37]. The calculation of the term frequency-inverse document frequency scheme (TF-IDF) is performed using the vector space model (VSM). By using a vector representation, each document is given the following cell weights:

$$d_x = \{FF_{x,1}, FF_{x,2}, \dots, FF_{x,y}, \dots, FF_{x,z}\} \qquad (1)$$

In this context, z represents the number of features, while F_{xy} represents the weight of the feature y in the document x. The feature's weighting was measured using the following weighting scheme.

$$FF_{x,y} = TF - IDF(x,y) = TF(x,y) \times \left(log \frac{d}{DF(y)}\right) \qquad (2)$$

where $TF_{(x,y)}$the feature y occurrence in document x, and $DF(y)$ is all the documents, including feature y. Matrix of the size $m \times n$ used to be the VSM as follows:

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

C. The Process of Clustering (Stage 2)

After the preprocessing stage has been completed, each Tafseer's corpus is ready as input to apply for the clustering process. The number of clusters used by the clustering algorithm should be entered manually (in this study, the clustering algorithm that determines the number of clusters automatically can not be used). Therefore, the selection of the study clustering algorithms was restricted by this condition. The value of clusters (k) must be decided before implementation, and it must be greater than one. Based on the literature, Chapter Al-Baqarah contains 53 themes, although some of them are duplicates. The corresponding verses are then organised into seven key themes or issues (see Figure 1). Therefore, the ideal value of k for clustering the Al-Baqarah chapter for the five cluster algorithms is seven and applied for all five TR [4], [26], [28]–[30], [38]. Hence, two algorithms are used and mentioned in the introduction before as follows:-

1.k-means

The k-means algorithm is the fundamental form of cluster partitioning algorithms. The k-means algorithm determines the centroid of a cluster by calculating the mean value of the data points within the cluster. It starts by choosing a centre or mean cluster at random from kc of Dcc items. The allocation of each remaining object to a cluster is determined by calculating the Euclidean distance between the object and the mean of each cluster and assigning the object to the cluster with the closest similarity. The within-cluster variance is then improved iteratively using the k-means algorithm. It calculates the new mean for every cluster by using the previous iteration object for each cluster. The new cluster centers would then reassign all objects using the updated means. The testing/iteration process continues until the assignment is consistent, which means that there are no major changes between the new and previous clusters formed. The procedure of k-means is summarised as follows[15]:-

Algorithm: k-means
Input: ***Dcc***: a dataset comprising ***m*** items,
kc: the number of clusters.
Output: A set of ***kc*** clusters.
Steps:
Step1: choose ***kc*** items arbitrarily from ***Dcc*** as the initial cluster centres;
Step2: **repeat**
Step3: re)assign each item to the cluster with which the item is most related to the basis that the items in the cluster has such a mean value;
Step4: Updating a clusters' means, requires computing the mean value of each cluster item;
Step5: **until** there is no change

2. Agglomerative hierarchical clustering

The following steps explain an overview of an agglomerative hierarchical clustering algorithm. The algorithm's pseudo-code steps are displayed below. In step 1, the dissimilarity matrix for all points in the dataset is computed. In steps 2–4, the dissimilarity matrix is updated by iteratively merging the closest pairs of clusters in a bottom-up manner. The rows and columns belonging to the previous clusters are deleted from the dissimilarity matrix and replaced with those belonging to the new cluster. Following that, merge procedures are performed using this updated dissimilarity matrix. Step 5 specifies the algorithm's termination condition[36].

Agglomerative Hierarchical Clustering is an algorithm that utilises the bottom-up approach, commencing by assigning each object as an individual cluster (referred to as an atomic cluster). Subsequently, these atomic clusters are iteratively merged into larger clusters until all objects have been integrated into a single cluster or until termination conditions are met. The term "Euclidean" refers to concepts, principles, or properties that are associated with distance and refers to the measure utilised to establish a connection between two groupings of objects. [15].

Algorithm : Agglomerative Hierarchical Clustering
Step1: Compute the dissimilarity matrix between all the data points.
Step2: **repeat**
Step3: Merge clusters as $C_{a \cup b} = C_a \cup C_b$. Set the new cluster's cardinality as $N_{a \cup b} = N_a + N_b$.
Step4: Insert a new row and column containing the distances between the new cluster $C_{a \cup b}$ and the remaining clusters.
Step5: **until** Only one maximal cluster remains.

D. The Validation of the Clustering (Stage 3)

According to the experiment's datasets of the study, as shown in the flowchart of Figure 3, only internal cluster validations plus the execution time were applied by the proposed method, as follows:-

1. *Time of Implementation (**ET**)*

During this stage, the value of time required to complete each algorithm's computation is calculated. The utilisation of hardware platforms that consist of an Intel Core i7-8550U CPU that runs at a rate of 1.80 GHz and is complemented by 8 GB of RAM puts a constraint on the amount of time that can be spent executing the algorithms that were investigated in this study. Microsoft Windows 10 and Python 3.7.7 are representations of the software platforms that are utilised.

2. internal Validation Metrics

A greater **Silhouette Coefficient (SC)** score indicates a cluster model with a higher level of organisation. Let x denote the average distance between a given sample and all other points within the same class, while y represents the average distance between the sample and all other points within the next neighbouring cluster; the SC is then given for a single sample as follows [39]:-

$$SC = \frac{y-x}{max(x,y)} \quad (4)$$

For inappropriate clustering, the score is -1, and for very dense clustering, +1. Zero scores show clusters that overlap. When the clusters are dense and separate well, the result is higher, which is a standard conception for the cluster [15] [39] [28].

Davies-Bouldin Index (DBI) aims to find sets of well-separated and compact clusters. The Davies-Bouldin Index is defined as [40]:

$$DBI = \frac{1}{cc}\sum_{i=1}^{cc} Max_{i\neq j}\left\{\frac{d(X_i)+d(X_j)}{d(cc_i,cc_j)}\right\} \quad (5)$$

Where *cc* represents the cluster number, *i,j* are labels of the cluster, then $d(X_i)$ and $d(X_j)$ are every sample in clusters *i* and *j* to their corresponding cluster centroids $d(cc_i, cc_j)$ is the space between these centroids. The lesser value of DBI indicates a "better" clustering solution.

3. Feature Extraction (PCA)

In feature extraction, features are projected into a new, lower-dimensional space. Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Singular Value Decomposition (SVD) are a few examples of feature extraction techniques. Both feature selection and extraction are dimensionality reduction approaches that can reduce computational complexity, improve learning performance, build more generalisable models, and reduce storage requirements. However, Regarding readability and interpretability, feature selection demonstrates superior performance compared to feature extraction due to its ability to retain the original feature values within the decreased space. On the other hand, feature extraction is a process that converts the data from its initial space into a different space characterised by a reduced dimensionality, which is not directly associated with the features present in the original space. [36].

The research's experiments used PCA [41], [42] for feature reduction to make 2D clusters visualization of each algorithm output. PCA is the most widely used unsupervised method. The covariance matrix $C_D = \frac{1}{N-1}DD^T$ is diagonalized into $\frac{1}{N-1}(PD)(PD)^T$ and eliminates lesser principal components, i.e., decreased P to size $k \times N$. Where, P is the principal component matrix, and the rows are the eigenvectors of DD^T.

IV. RESULTS OBTAINED

A. Text Cleansing and Internal Validation Metrics Results

According to the studies [26], [29], [30], [43], chapter Al-Baqarah can be clustered using $k = 7$, hence the study experments applied this value for the all five translotor (TR1-TR5). Table 3 shown k-means algorithm results of internal validation metrics (SC, DBI) and the execution time (ET) in addition to features statistics. Table 4 shows the same results but for the agglomerative hierarchical cluster algorithm.

Table 3. k-means internal validation metrics results and feature statistics for each translator.

k-means (Partition)						
Translator No.	**ET(second)**	**SC**	**DBI**	**Total Features**	**Stop Words Removal**	**Stemming**
TR1	0.24534	0.00678	6.97075	2008	1775	1388
TR2	0.28424	0.00729	6.75686	1591	1392	1254
TR3	0.14062	0.00625	6.42995	1592	1322	1043
TR4	0.23334	0.00714	6.55994	1659	1438	1126
TR5	0.14057	0.01225	6.09906	1563	1332	1060

Table 4. Agglomerative internal validation metrics results and feature statistics for each translator.

Agglomerative (Hierarchical)						
Translator No.	**ET(second)**	**SC**	**DBI**	**Total Features**	**Stop Words Removal**	**Stemming**
TR1	0.12546	0.00722	5.21939	2008	1775	1388
TR2	0.1275	0.01052	4.63428	1591	1392	1254
TR3	0.04689	0.00728	4.4171	1592	1322	1043
TR4	0.11003	0.00812	5.46683	1659	1438	1126
TR5	0.09373	0.01166	3.86967	1563	1332	1060

The study aims to prioritise the five translators of chapter Al-Baqarah. Table 5 shows the rank of these five translators depending on (ET, SC, and DBI) of the k-means algorithm. From this table, only rank numbers (1) and (3) demonstrate the same rank. Table 6 shows the rank of the same five translators obtained from the Agglomerative algorithm. Each (ET, SC, and DBI) has a different rank from this table. Finally, it is essential to use a modern technique (future work) to get the optimal union rank (e.g. MCDM (Multi-Criterria Decision-making Analysis) [44]). The utilization of MCDM can be employed to address a variety of case studies. The publications [45]–[49] provide variations in terms of their case study applications, aggregation functions, and criteria weighting kinds. The ranking of the five TRs is based on the findings of machine results. Therefore, a potential future research recommendation involves acquiring the study dataset rank based on the perspective of an expert specializing in Islamic studies with extensive knowledge and experience in Arabic to English translations. This suggestion involves contrasting the intended research rank with the human expert rank. The argument may have connections to other academic fields, such as the field of Islamic studies.

Table 5. The rank of the translator according to the k-means algorithm results.

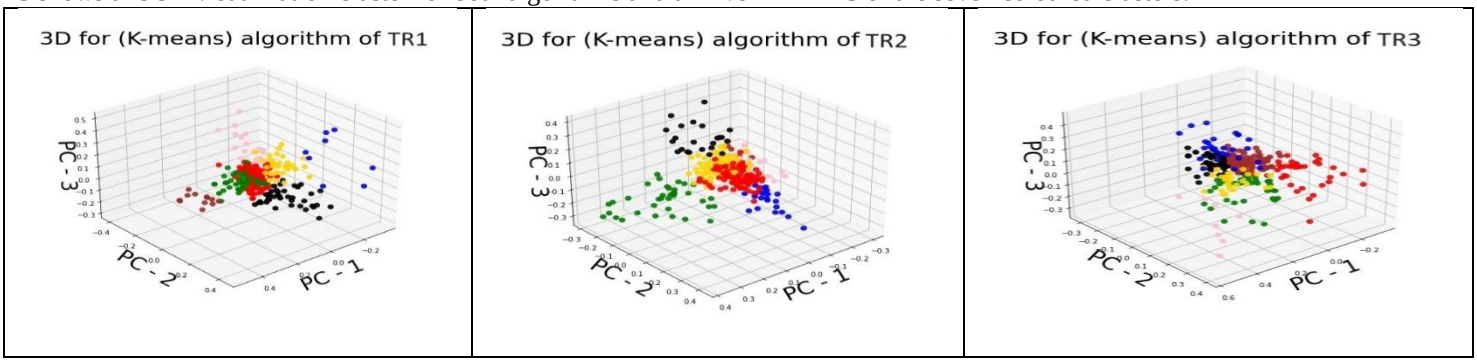
Rank	Time(second)	SC	DBI
1	**TR5**	**TR5**	**TR5**
2	TR3	TR2	TR3
3	**TR4**	**TR4**	**TR4**
4	TR1	TR1	TR2
5	TR2	TR3	TR1

Table 6. The rank of the translator according to the Agglomerative algorithm results.

Rank	Time(second)	SC	DBI
1	TR3	TR5	TR5
2	TR5	TR2	TR3
3	TR4	TR4	TR2
4	TR1	TR3	TR1
5	TR2	TR1	TR4

B. Feature Extraction (PCA) Results

Experiments in the study employed PCA for feature reduction to generate 2D cluster visualisations of each algorithm's output. Figure 5 shows this 3D visualization cluster for both algorithms and all five TR1-TR5 of the seven coloured clusters.



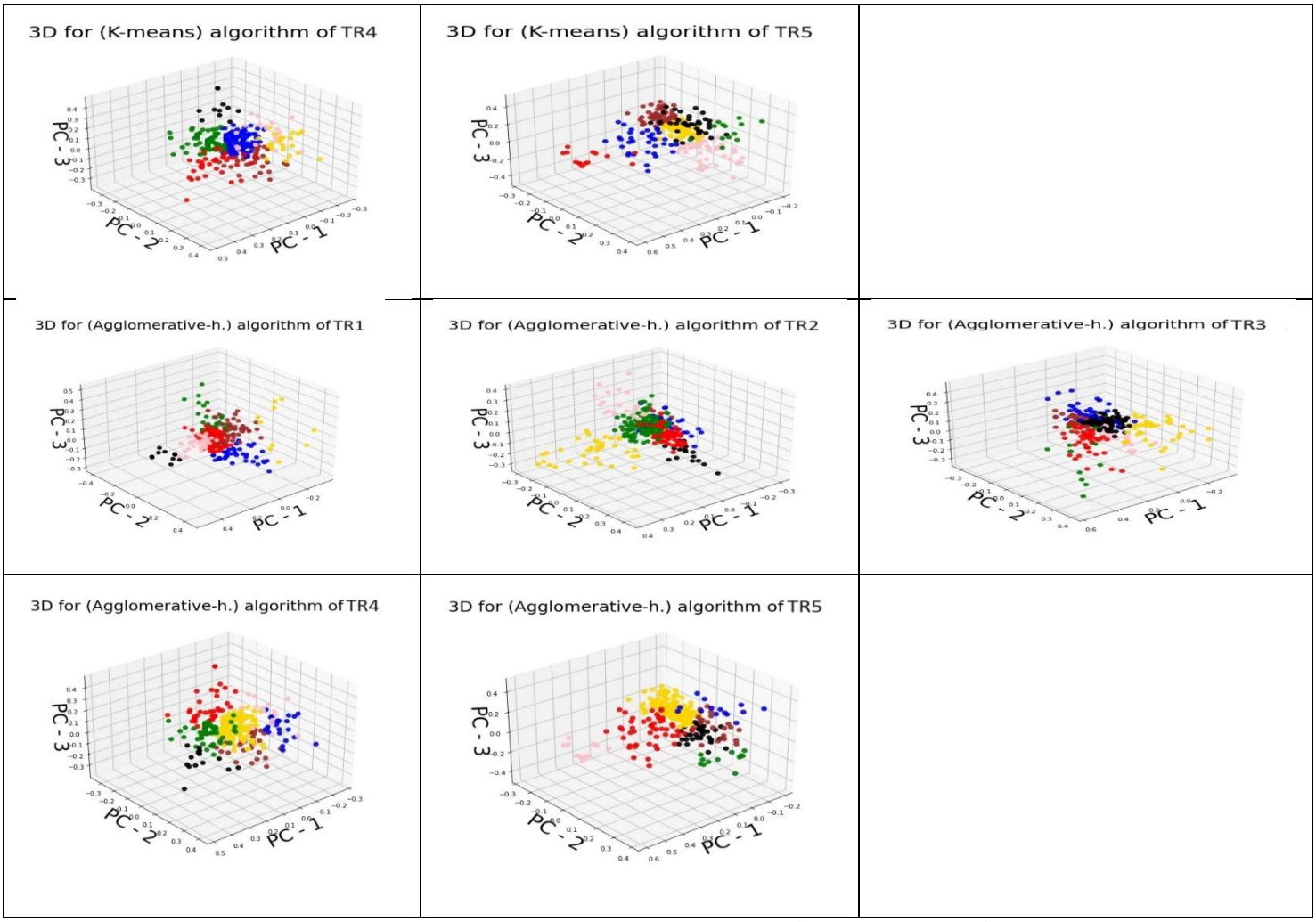


Figure 5. k-means and Agglomerative 3D cluster visualization algorithms of seven coloued clusters for (TR1-TR5).

V. CONCLUSION

The study was conducted in three stages. During the first stage of preprocessing, the input datasets comprised a choice of 286 verses from the Al-Baqarah chapter English Tafseer text. These verses were associated with five distinct translators, denoted as TR1-TR5. This stage employs text cleansing (e.g. tokenization). In addition, TF-IDF and VSM are used to produce the five corpora. During the subsequent stage of the analysis, the clustering process involved the execution of the k-means and Agglomerative cluster algorithms. The use of seven clusters (k=7) was based on the findings from the existing literature for each of the five datasets. In the third and final stage, called clustering validation, the ET is computed with the SC and DBI as internal cluster validation metrics. Furthermore, PCA was employed in that stage to demonstrate and visually represent the output datasets/algorithms of the seven clusters.

Based on the results (ET, SC, and DBI) of the k-means algorithm, only rank (1) and (3) indicate a similar ranking for these five translators. Conversely, the Agglomerative algorithm displays the ranking of the same five translators; each (ET, SC, and DBI) has a unique rank. Therefore, an advanced technique such as MCDM (Multi-Criteria Decision-making Analysis) must be utilized to determine the optimal union rank. The application of MCDM can be applied to solve many different case studies. These articles have different case study applications, aggregation functions, and criteria weighting types.

In future, and in addition to MCDM, the authors suggest implementing additional clustering algorithms. Increasing the number of datasets to more than five translators, analyzing and finding the results, and increasing the internal validation metrics numbers used.

Moreover, obtaining the study dataset rank from the opinion of somebody specialising in Islamic studies with a good and wide knowledge experience of Arabic to English translations and comparing the s proposed study rank and the obtained rank is a future work suggestion. This suggestion may be related to other faculties, such as the faculty of Islamic studies. The Islamic faculty student can compare their findings to the proposed research.

ACKNOWLEDGMENT

Universiti Kebangsaan Malaysia supports this research. Moreover, We would like to thank the staff of the collage of engineering of Al-Iraqia for presenting support to complete this work as shown in an official paper.

REFERENCES

[1] K. Zebiri, "Neal Robinson: Discovering the Qur'an: a contemporary approach to a veiled text. xiv, 332 pp. London: SCM Press Ltd., 1996.£16.95.," *Bull. Sch. Orient. African Stud.*, vol. 61, no. 3, pp. 538–540, 1998.

[2] A. H. M. Ragab and A. S. Bajnaid, "An Effective-Adaptive E-learning System Based on Multi-Styles Assessment," in *7th Annual Symposium on Learnining and Technology, the Edutainment Effat Univ. King AbdulAziz University, Jeddah, Saudi Arabia*, 2009, pp. 10–11.

[3] D. E. Smith, "The structure of al-Baqarah," *Muslim World*, vol. 91, no. 1/2, p. 121, 2001.

[4] C. Hadhiri, *Klasifikasi Kandungan Al-Qur'an*. Jakarta: Gema Insani, 1993.

[5] S. Sadiq, *A comparative study of four English translations of Sûrat Ad-Dukhân on the semantic level*. Cambridge Scholars Publishing, 2010.

[6] C. Blake, *Text mining*, vol. 45. 2011. doi: 10.1002/aris.2011.1440450110.

[7] M. N. Al-Kabi, H. A. Wahsheh, and I. M. Alsmadi, "A Topical Classification of Hadith Arabic Text," in *Proceedings - 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, NOORIC 2013Taibah University International Conference on Advances in Information Technology*, 2013, pp. 252–257.

[8] C. Qi, L. Jianfeng, and Z. Hao, "A text mining model based on improved density clustering algorithm," in *2013 IEEE 4th International Conference on Electronics Information and Emergency Communication*, 2013, pp. 337–339.

[9] M. Alhawarat, M. Hegazi, and A. Hilal, "Processing the text of the Holy Quran: a text mining study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 262–267, 2015.

[10] A. Fahad *et al.*, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 267–279, 2014.

[11] A. Aslani and M. Esmaeili, "Finding Frequent Patterns in Holy Quran UsingText Mining," *Signal Data Process.*, vol. 15, no. 3, 2018, doi: 10.29252/jsdp.15.3.89.

[12] S. J. Putra, T. Mantoro, and M. N. Gunawan, "Text mining for Indonesian translation of the Quran: A systematic review," *3rd Int. Conf. Comput. Eng. Des. ICCED 2017*, vol. 2018-March, pp. 1–5, 2018, doi: 10.1109/CED.2017.8308122.

[13] C. Luque, J. M. Luna, M. Luque, and S. Ventura, "An advanced review on text mining in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 3, p. e1302, 2019.

[14] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, Z. A. A. Alyasseri, and S. N. Makhadmeh, "A novel hybrid multi-verse optimizer with K-means for text documents clustering," *Neural Comput. Appl.*, 2020, doi: 10.1007/s00521-020-04945-0.

[15] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, 3rd ed. Elsevier, 2012.

[16] W. A. Mohotti, "Unsupervised text mining: Effective similarity calculation with ranking and matrix factorization," Queensland University of Technology, 2020.

[17] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.

[18] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. theory*, vol. 28, no. 2, pp. 129–137, 1982.

[19] W. H. E. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J. Classif.*, vol. 1, no. 1, pp. 7–24, 1984.

[20] S. J. Putra, R. H. Gusmita, K. Hulliyah, and H. T. Sukmana, "A semantic-based question answering system for indonesian translation of Quran," in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, 2016, pp. 504–507.

[21] C. Slamet, A. Rahman, M. A. Ramdhani, and W. Darmalaksana, "Clustering the verses of the Holy Qur'an using K-means algorithm," *Asian J. Inf. Technol.*, vol. 15, no. 24, pp. 5159–5162, 2016.

[22] S. J. Putra, K. Hulliyah, N. Hakiem, R. P. Iswara, and A. F. Firmansyah, "Generating weighted vector for concepts in indonesian translation of Quran," in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, 2016, pp. 293–297.

[23] H. T. Sukmana, R. H. Gusminta, Y. Durachman, and A. F. Firmansyah, "Semantically annotated corpus model of Indonesian Translation of Quran: An effort in increasing question answering system performance," in *2016 4th International Conference on Cyber and IT Service Management*, 2016, pp. 1–5.

[24] B. Hamoud and E. Atwell, "Quran question and answer corpus for data mining with WEKA," in *2016 Conference of Basic Sciences and Engineering Studies (SGCAC)*, 2016, pp. 211–216.

[25] S. Chua and P. N. E. Nohuddin, "Relationship Analysis of Keyword and Chapter in Malay-Translated Tafseer of Al-Quran," *J. Telecommun. Electron. Comput. Eng.*, vol. 9, no. 2–10, pp. 185–189, 2017.

[26] A. F. Huda, M. R. Deyana, Q. U. Safitri, W. Darmalaksana, U. Rahmani, and others, "Analysis Partition Clustering and Similarity Measure on Al-Quran Verses," in *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*,

2019, pp. 1–5.
[27] Z. Indra, A. Adnan, and R. Salambue, "A Hybrid Information Retrieval for Indonesian Translation of Quran by Using Single Pass Clustering Algorithm," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019, pp. 1–5.
[28] 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)*, 2019, pp. 1–5.
[29] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 248–254, 2020, doi: 10.14569/IJACSA.2020.0110832.
[30] 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," *JOURNALOF QURANIC Sci. Res.*, vol. 2, no. 2, pp. 48–53, 2021.
[31] S. J. Putra, R. H. Gusmita, K. Hulliyah, and H. T. Sukmana, "A semantic-based question answering system for indonesian translation of Quran," in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, 2016, pp. 504–507. doi: 10.1145/3011141.3011219.
[32] S. Chua and P. N. E. binti Nohuddin, "Frequent pattern extraction in the Tafseer of Al-Quran," in *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, 2014, pp. 1–5. doi: 10.1109/ICT4M.2014.7020667.
[33] M. Z. Husin, S. Saad, and S. A. M. Noah, "Syntactic rule-based approach for extracting concepts from quranic translation text," in *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, 2017, pp. 1–6.
[34] W. Wu, Z. Xu, G. Kou, and Y. Shi, "Decision-making support for the evaluation of clustering algorithms based on MCDM," *Complexity*, vol. 2020, 2020.
[35] G. Forman and E. Kirshenbaum, "Extremely fast text feature extraction for classification and indexing," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 1221–1230.
[36] C. C. Aggarwal and C. K. Reddy, *Data clustering*. Citeseer, 2014.
[37] B. Bansal and S. Srivastava, "Hybrid attribute based sentiment classification of online reviews for consumer intelligence," *Appl. Intell.*, vol. 49, no. 1, pp. 137–149, 2019.
[38] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Text Clustering of Tafseer Translations by Using k-means Algorithm : An Al-Baqarah Chapter View," *Ann. Emerg. Technol. Comput.*, vol. 7, no. 4, pp. 27–34, 2023, doi: 10.33166/AETiC.2023.04.003.
[39] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[40] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 2, pp. 224–227, 1979.
[41] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *J. Educ. Psychol.*, vol. 24, no. 6, p. 417, 1933.
[42] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.
[43] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "k -means variations analysis for translation of English Tafseer Al-Quran text," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 3, pp. 3255–3265, 2023, doi: 10.11591/ijece.v13i3.pp3255-3265.
[44] R. L. Keeney, H. Raiffa, and others, *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.
[45] O. S. Albahri *et al.*, "Multidimensional benchmarking of the active queue management methods of network congestion control based on extension of fuzzy decision by opinion score method," *Int. J. Intell. Syst.*, 2020, doi: 10.1002/int.22322.
[46] A. S. Albahri *et al.*, "Integration of fuzzy-weighted zero-inconsistency and fuzzy decision by opinion score methods under a q-rung orthopair environment: a distribution case study of COVID-19 vaccine doses," *Comput. Stand. \& Interfaces*, vol. 80, p. 103572, 2022.
[47] M. M. Salih, Z. T. Al-Qaysi, M. L. Shuwandy, M. A. Ahmed, K. F. Hasan, and Y. R. Muhsen, "A new extension of fuzzy decision by opinion score method based on Fermatean fuzzy: A benchmarking COVID-19 machine learning methods," *J. Intell. \& Fuzzy Syst.*, no. Preprint, pp. 1–11, 2022, doi: 10.3233/JIFS-220707.
[48] A. H. Alamoodi *et al.*, "New Extension of Fuzzy-Weighted Zero-Inconsistency and Fuzzy Decision by Opinion Score Method Based on Cubic Pythagorean Fuzzy Environment: A Benchmarking Case Study of Sign Language Recognition Systems," *Int. J. Fuzzy Syst.*, pp. 1–18, 2022, doi: 10.1007/s40815-021-01246-z.
[49] A. H. Alamoodi *et al.*, "Based on neutrosophic fuzzy environment: a new development of FWZIC and FDOSM for benchmarking smart e-tourism applications," *Complex \& Intell. Syst.*, pp. 1–25, 2022, doi: 10.1007/s40747-022-00689-7.