Analysis of K-means, DBSCAN and OPTICS Cluster Algorithms on Al-Quran Verses

Mohammed A. Ahmed[1], Hanif Baharin [2], Puteri N.E. Nohuddin [3]
Institute of IR 4.0, Universiti Kebangsaan Malaysia
Bangi, Selangor, 43600, Malaysia

***Abstract*—Chapter Al-Baqarah is the longest chapter in the Holy Quran, and it covers various topics. Al-Quran is the primary text of Islamic faith and practice. Millions of Muslims worldwide use Al - Quran as their reference book, and it, therefore, helps Muslims and Islamic scholars as guidance of the law life. Text clustering (unsupervised learning) is a process of separation that has to be divided text into the same section of similar documents. There are many text clustering algorithms and techniques used to make clusters, such as partitioning and density-based methods. In this paper, k-means preferred as a partitioning method and DBSCAN, OPTICS as a density-based method. This study aims to investigate and find which algorithm produced as the best accurate performance cluster for Al-Baqarah's English Tafseer chapter. Data preprocessing and feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) have applied for the dataset. The result shows k-means outperformed even has the smallest of Silhouette Coefficient (SC) score compared to others due to less implementation time with no noise production for seven clusters of Al-Baqarah chapter. OPTICS has no noise with the medium of SC score but has the longest implementation time due to its complexity.**

Keywords*—K-means; DBSCAN; OPTICS; Al-Baqarah clustering; Silhouette Coefficient; Tafseer; text clustering

I. Introduction

The Quran is a significant religious text written in Quranic Arabic, followed by believers of the Islamic faith. The Quran means "perfect reading," in terms of language, which Muslims believed to be revealed to people as a guide in all aspects of life. Al - Quran is the message of Allah that was revealed and spread to the prophet Mohammed From the Prophet SAW 's time until now protected by Allah. Al-Quran wrote in Arabic but has translated into numerous languages around the world, as well as English. It is text data that may be further analyzed. Chapter Al-Baqarah is Al-Quran's longest chapter, so there are various themes in Al-Baqarah's chapter. Such themes are not written sequentially but depend on asbabunnuzul ayat (verses) while revealed. Hence, grouping verses of similar characteristics of a text they compose will form a cluster that could reflect any theme in Surah (chapter) Al–Baqarah [1].

Data mining began in the 1980s, made a lot of progress in the 1990s, and is growing in the early 2000s. Data mining can transform large-scale data set into knowledge to help meet significant issues; it can meet this need by providing data knowledge-based tools [2]. Discovering useful information from groups of text documents is known as text mining. Text mining has a meaningful effect on different applications, for example, social media data, opinion mining, and recommendation systems. Text mining is a common approach to uncover meaningful information from text collections, including clusters, outliers, and the evolution of clusters. The lack of ground-truths in real-world samples creates a demand to perform such analyzes in an unsupervised context [3].

Text clustering (unsupervised learning) is the process for the mining of text, which divides the similar text of documents into groups or clusters. It is typically done by finding patterns and trends through the use of several text manipulation techniques and specific algorithms. Text clustering is a part of data mining. Documentation loaded in the vector weight term become cluster objects. Therefore, there are many clustering algorithms and techniques used to make clusters, such as the partitioning method (k-means and k-medoid) [4], hierarchical method (Agglomerative and Divisive) [5], density-based method (DBSCAN and OPTICS) [6], and grid-based method (STING and CLIQUE) [2].

Most of the articles adapted one method to cluster the text translated Al-Quran (Tafseer) as the dataset for its experiments, such as [5] [7-10]. This paper has adapted two clustering methods (partitioning and density-based) and compared and analyzed Al-Baqarah's chapter text of English Tafseer as the experiment dataset. K-means preferred as a partitioning method and DBSCAN, OPTICS as a density-based method. This study aims to investigate and find which algorithm produced the best performance cluster for Al-Baqarah's chapter.

The rest of this paper is structured as follows: Section 2 discusses the related work of this research; Section 3 discusses the research methodology. Section 4 describes the experimental procedure and results. Finally, Section 5 presents the conclusions.

II. Related Work

There are many papers related to this article. The clustering experiment of [1] utilizes a mixture of k-means clustering techniques, k-medoid, and bisecting k-means, together with Jaccard similarity, correlation coefficients, and cosine similarity produce different validity values. The ideal cluster results, however, in chapter al - Baqarah clustering process given by cosine similarity of k-medoid. This research [11] tried to group Hadith texts of Indonesian text language to compare Fuzzy c-means and k-means algorithms with determined parameters and experiments. F-measure and Silhouette Coefficient calculations used as measure calculations to

evaluate clusters performances. Findings demonstrate that the Fuzzy c-means algorithm is more useful to group the hadith text based on consistency with the chapter and original data. Additionally, this paper suggested using DBSCAN and compare with k-means as future work, and this is one of the motivations to do the investigation related to this article.

This study [7] resulted in an initial practical move to understand the concepts of verses in the Holy Quran. The algorithm used to cluster 6236 total verses, using partitioning (k- means) for unsteamed, steamed words that formed three clusters. The paper [5] uses a hybrid of TF-IDF (Term Frequency-Inverse Document Frequency) and Network Analysis (map) approach to extract keywords and identify relationships between keywords and Tafseer chapters. Six short chapters of 130 keywords are taken from Malay translated Tafseer of Al-Quran. The proposed method was called KCRA. Reference [12] developed a semantic-based question answering (QAS) for Indonesian Quran translation, which asked the users three questions, then created a TF-IDE for each term belonging to the respective expected response category (also called entity group) to feed or provide a semantic interpreter on user request. The author has organized 222 ontology principles into 6, 24, and 77 of Time, Position, and Individual concepts, respectively. The research [10] aims to develop a web-based verse search (information retrieval) system for Al-Quran, which is integrated with the clustering algorithm (SPC) to facilitate Muslim discovery of relevant information in the Quran verse by grouping the Quran verse within their its similar group. This paper [13] discusses a study on generating weighted vectors for each concept in Indonesian Quran Translation (ITQ) and applying QAS such as [12]. Still, here the author provides more information on TF-IDE findings and has different work procedures.

In [14], various text clustering algorithms were studied. The main objective of this study is comparing different clustering algorithms and finding out which algorithm is most suitable for users using WEKA free open source software, with the advantages and disadvantages of each algorithm like DBSCAN and k-means. The author founds that the k-means clustering algorithm is the most straightforward compared to other algorithms. The author in [15] suggested an algorithm to estimate the optimum value of ϵ-neighbourhood and *MinPts* for the DBSCSN algorithm, based or used k-means algorithm.

The author in [16] has applied an OPTICS clustering on text data and provided valuable insight into the operation of OPTICS and its applicability to text information. The SCI algorithm introduced in this paper to create clusters from the OPTICS plot can be used as a benchmark to check OPTICS efficiency based on measurements of purity and coverage. The author in [17] suggested an ICA incremental clustering algorithm based on the OPTICS. Like OPTICS, the ICA also generates a dataset's cluster-ordering structure. The ICA is, however, smarter because no parameters are needed, which are very difficult to define for users who do not know the properties of datasets and delete some of these complex definitions. ICA is ideal for processing dynamic datasets. The author in [18] introduced a new similarity measure for sequential data and suggested an improved density-based clustering technique to find meaningful clusters in different web databases. Experiments compared DBSCAN clustering characteristics, OPTICS algorithm with the new advanced SSM-DBSCAN algorithm, and SSM-OPTICS on web sessions thought various similarity indicators such as Euclidean, Jaccard, Fuzzy, and Cosine.

III. Research Methodology

A. Text Preprocessing

Before the clustering algorithm is applied, the standard preprocessing procedures are used to preprocess text documents, which includes:

- Tokenization: Text data is divided into the basic sequence of independent units.
- POS tagging: The process of marking up a word in a text (corpus) as corresponding to a particular part-of-speech.
- Case folding or normalization: The process of converting all the characters in a document into the same case, either all upper case or lower.
- Stop word removal: Deleting particular common words that happened most often, such as 'is',' are',' that', 'an'….
- Stemming: Convert verbs to their origins for convenience.
- Term weighting steps: The texts are processed in numerical format or matrix in the preprocessing steps.

B. Feature Extraction

Term weighting used to extract terms or features after the preprocessing completed. Term weighting aims to convert text data into a numeric format. The literature contains many schemes for term weighting. For text document representation, the vector space model (VSM) calculates the term frequency-inverse document frequency scheme (TF-IDF) [19]. The VSM shows each document with a vector and weighs the cell values as follows:

$$d_i = \{F_{i,1}, F_{i,2}, \dots, F_{i,j}, \dots, F_{i,t}\} \quad (1)$$

where t refers to the features number and F_{ij} refers to the feature j weight in document i [20]. The following weighting scheme is used to measure the feature's weighting:

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

where $TF_{(i,j)}$ the feature j frequency in document i, and $DF(j)$ is all the documents containing feature j. Matrix of size $m \times n$ is used to represent the VSM as follows:

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

C. Cluster Techniques

This study adopted the three most popular clustering algorithms, described as follows:

1) k-means: The k-means algorithm is the most common clustering algorithm. It is a popular clustering technique introduced 50 years ago. Due to its simplicity, the k-means algorithm was commonly used to handle huge datasets. It is also easily implemented, enjoying low computational complexity and rapid convergence [21]. The process includes two main iterative steps. This process classifies the whole dataset into heterogeneous clusters. Across the years, many versions of this algorithm were developed to enhance its performance, such as k-medoids [22], kernel k-means [21], and k-harmonic-means [23].

k-means is the simplest and most fundamental version of partitioning cluster analysis. The k-means algorithm defines a cluster's centroid as the cluster's mean point value as follows. First, it selects a random ***kl*** of ***Dc*** items, each initially representing a mean or centre cluster. The k-means algorithm procedure is summarized as follows [2]:

Algorithm: k-means

Input: ***Dc***: a dataset comprising ***nu*** items,
kl: the number of clusters.
Output: A set of ***kl*** clusters.

Steps:
Step1: arbitrarily select ***kl*** items from ***Dc*** 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) DBSCAN: Hierarchical and partitioning clustering approaches are useful for spherical-shaped of clusters only. Density cluster methods are designed to solve the problem of locating arbitrary shape clusters like "S" shape and oval clusters. Such data would likely misidentify convex regions where noise or outliers are included in clusters, and this is the principal strategy behind clustering approaches based on density, which can find non-spherical form clusters. Density-Based Spatial Clustering of Applications with Noise referred to (DBSCAN) identifies central objects that have neighbourhoods of dense. It links centre objects and their neighbourhoods to construct dense areas as clusters. This algorithm requires two user-specified parameters as inputs, which are ϵ and MinPts [2]. The DBSCAN algorithm procedure is summarized as follows [2]:

Algorithm: DBSCAN

Input: ***Data*** = a data set including ***nu*** objects, *MinPts* = the threshold of neighbourhood density, ϵ = the radius parameter.
Output: A establishment of density-based clusters.

Steps:
Step1: sign all objects as **unchecked**;
Step2:**do**
Step3: at random choose an **unchecked** object ***b***;
Step4: sign ***b*** as **checked**;
Step5: **if** the ϵ-neighbourhood of ***b*** includes at least *MinPts*
Step6: establish a new cluster ***Cl***, then add together ***b*** to ***Cl***;
Step7: suppose ***M*** be the set of objects appearing in the ϵ-neighborhood of ***b***;
Step8: **for** every point ***b′*** in ***M***
Step9: **if** ***b′*** is equal **unchecked**
Step10: sign ***b′*** as **check**;
Step11: **if** the ϵ-neighbourhood of ***b′*** includes at least *MinPts* points, add the ***M*** points;
Step12: **if** ***b′*** isn't a cluster member yet, add ***b′*** to ***Cl***;
Step13: **ending for**
Step14: output ***Cl;***
Step15: **else** sign ***b*** as **noise;**
Step16:**until** no object is **unchecked**;

3) OPTICS: This method proposed to remove the problem of using a set of required variables in clustering analysis. Ordering Points To Identify the Clustering Structure referred to OPTICS, explicitly does not produce a dataset clustering but produces an ordering cluster. It is a continuous array of all objects within analysis which describes data clustering structure dependent on density, then the objects in a denser cluster listed in the ordering of the cluster. This order refers to the density-based clustering acquired from a large variety of parameters. OPTICS, therefore, does not require the users to give a certain density threshold. The cluster order can be applied to obtain necessary clustering data (e.g., arbitrary-shaped clusters or cluster centres), to derive and display the clustering composition. The basic idea is to define a specific database cluster and noise like DBSCAN. OPTICS identifies the cluster based on density [2].

OPTICS requires two essential aspects per item. Firstly, the (core distance) of item **b** is the smallest value ϵ' so that ϵ'-neighbourhood has at least MinPts items. That's the minimum distance threshold, which makes **b** a core item is ϵ'. If b isn't a core item, the core distance of **b** is indeterminate. Secondly, the minimum radius value (reachability-distance) to item **b** from **c** makes **b** density-reachable from **c**. Under the concept of density-reachability, **c** must be a core point, and b must be in **b**'s neighbourhood. And hence, the reachability distance is max{core-distance(**b**), dist(**b**, **c**)} from **c** to **b**. If **c** is not a core item, the reachability-distance from **c** to b is unspecified.

Several core items can directly reach an item **b**. Accordingly, **b** could have numerous reach-distances for different core items. The smallest reachability-distance of **b** is special importance as it provides the shortest route through which **b** links to a dense cluster [2].

D. Cluster Evaluation

Cluster validation or evaluation of the effects of clustering is useful for measuring the accuracy of the clustering or grouping. Many cluster validation methods can be used. If the ground truth of a dataset is not available such as this paper's experiments, an intrinsic method must be used to evaluate clustering accuracy. Generally, Intrinsic approaches measure clusters by evaluating well how-isolated clusters are and how dense clusters are. Most intrinsic approaches benefit a metric similarity between items in the dataset, such as the Silhouette Coefficient [2] [24].

1) Silhouette Coefficient (SC): A higher Silhouette Coefficient score refers to a more structured cluster model. The Silhouette Coefficient is defined and consists of two scores for each sample. If x is the mean distance between a sample and all other points in the same class, and y is the mean distance between a sample and all other points in the next closest cluster. Then, the Silhouette Coefficient sc for a single sample is given as [24]:

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

The score is set to -1 for inappropriate clustering and +1 for very dense clustering. Zero scores show overlapping clusters. The scoring is higher when the clusters are dense and well separated, which is a standard cluster concept [24] [2] [11].

2) Execution time: The algorithms execution time of this study is limited or related to use the hardware platforms of Intel Core i7-8550U CPU with @ 1.80 GHz and RAM of 8 GB and software platforms of MS Windows 10, Python 3.7.7.

IV. Experiments and Result Discussions

Fig. 1 describes the flowchart experiments of this study, the first part of this section discusses the source of the experiment's dataset, the statistics result before and after text prepossessing, output terms after the feature extraction process, and the parameters used to implement the algorithms. In contrast, the second part of this section discusses the results of three cluster algorithms and which is the optimal cluster algorithm for the experiment's dataset.

A. The Experiments

1) Used dataset: http:/tanzil.net/trans/ a website provides documentation about the Quran translation in various languages, for many translators from various Tafseer. Many such researchers used this website to collect data and adapted for its experiments [5] [7] [8] [13] [25-27]. Al-Baqarah chapter used in this research that consisted of 286 verses written by (Ahmed Ali) of English Tafseer Al-Quran, the text document contains 286 lines, each line represents Tafseer of one verse. The total words of this document are 11478, including all terms, names, numbers, symbols, and marks.

2) The preprocessing: Preprocessing of data was done in many stages, namely the document reading of 11478 words, tokenizing these words to become a sequence of independent units, stop word removal and normalize from becoming 1508 terms of features, and stemming from becoming 1221 features. Now the data is ready for the next step.

3) Features Extraction: After the text is preprocessed, the dictionary or corpus containing all the words in the document is assembled. Therefore, weight is calculated (generated) for each document feature (time weighting), or (equivalent weighting). TF-IDF used to assign a weight to each term or feature. All these weights will combine in a matrix to construct the corpus ready for the clustering process. Fig. 2 shows the bar chart for the most first 15 features of the Al-Baqarah chapter.

Fig 1. Flowchart of the Study Experiments.

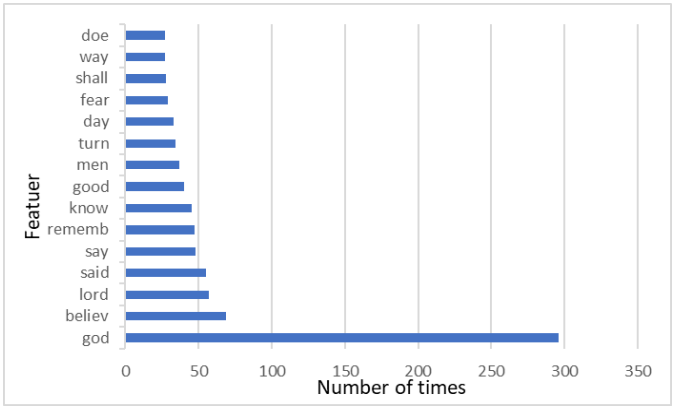

Fig 2. Features of Al-Baqarah Chapter.

4) Cluster algorithm parameters: The algorithm k-means clustering adapted from the partitioning method for this study. The value of k needs to be determined before implementation, which should be more than one. Based on the experiments of Choiruddin [28] and Huda [1], the optimal value of k to cluster the Al-Baqarah chapter using k-means is seven. Chapter Al-Baqarah has 53 subjects, although some still have the same subject as others. The verses with the same subject are then grouped into seven major themes or topics.

Based on the above and in this study experiments, the values MinPts and ϵ must be 7, 0.2, respectively to make the DBSCAN algorithm produced seven clusters with some noise, while the value of MinPts should be 9 for seven cluster output of the OPTICS algorithm.

B. Results and Discussion

The finding and discussion of this study can be summarized as follows:

- Depends on [1] and [28], k=7 for k-means to cluster chapter Al-Baqarah.
- Fig. 3 shows the visual clustering for the two/three most common features ('god', 'believ', and 'lord') applied by the k-means algorithm. These legends (seven cluster colors) represent the cluster for the two/three features.

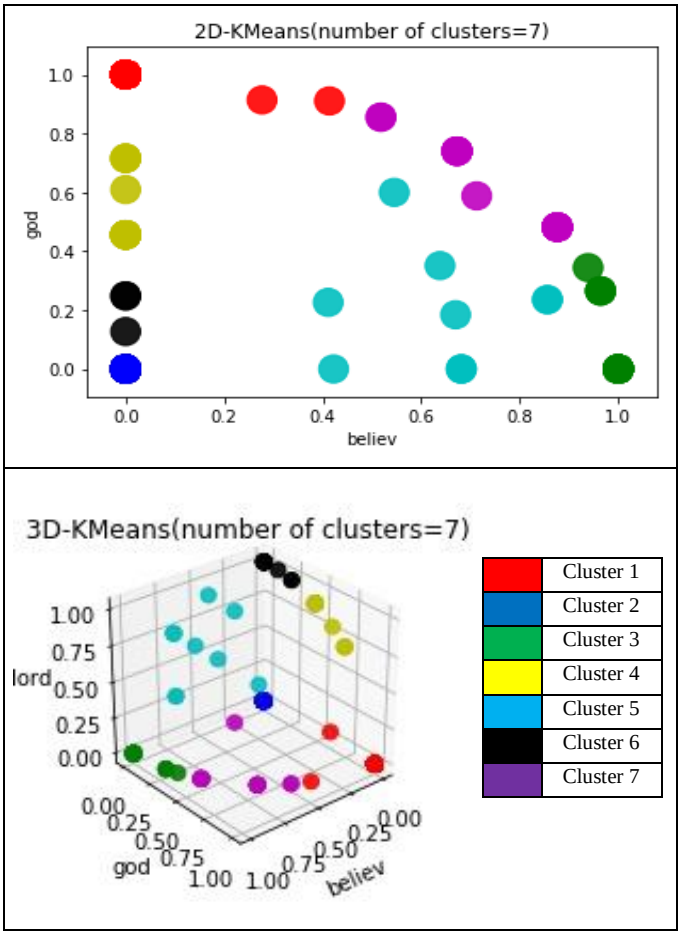

Fig 3. Cluster Visualization for the 2D/3D k-Means Algorithm.

- Fig. 4 gives the SC results for each cluster (seven) of the k-means, where the average for these results is equal to 0.8948 with 0.319 seconds of implementation time.

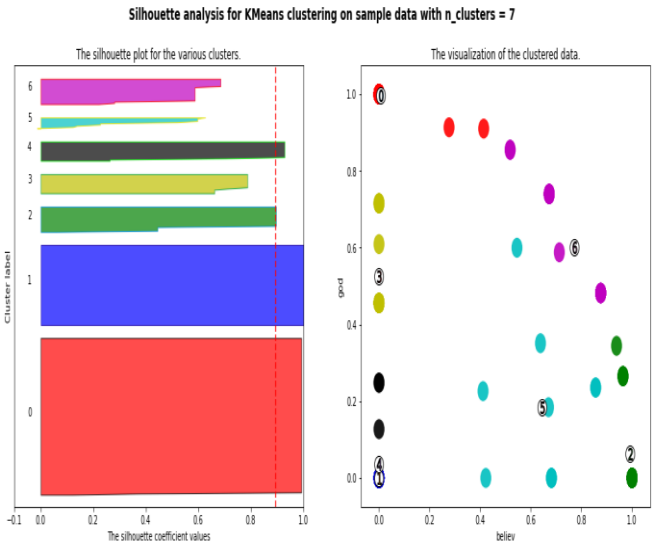

Fig 4. Silhouette Coefficient for Each Cluster with the Average for the k-Means Algorithm.

- Fig. 5 visuals the output of the OPTICS cluster algorithm for the two most common features ('god', 'believ'), where *MinPts* = 9 to make OPTICS produced seven clusters with no need to calculate ϵ due to algorithm property. The SC is equal to 0.9098, with 0.935 seconds of implementation time.

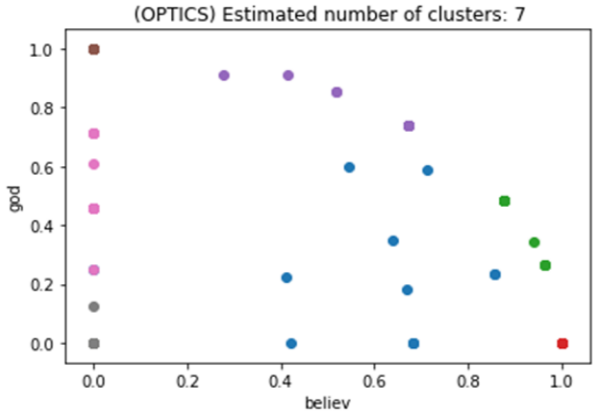

Fig 5. Cluster Visualization for OPTICS Algorithm.

- Fig. 6(A) shows the visual clustering for the two most common features ('god', 'believ') applied by the DBSCAN algorithm with some noise of black circles. From this study experiments, *MinPts* = 7, and ϵ = 0.2 to make DBSCAN produce seven clusters (Fig. 6(B) describes this cluster production number with regards to *MinPts* and ϵ). DBSCAN produces some noises, Fig. 6(C) illustrates these noises with regards to *MinPts* and ϵ. Fig. 6(D) shows the SC results obtained from this study trials. The SC is equal to 0.9129 with 0.327 second implementation time if DBSCAN produced the seven clusters.

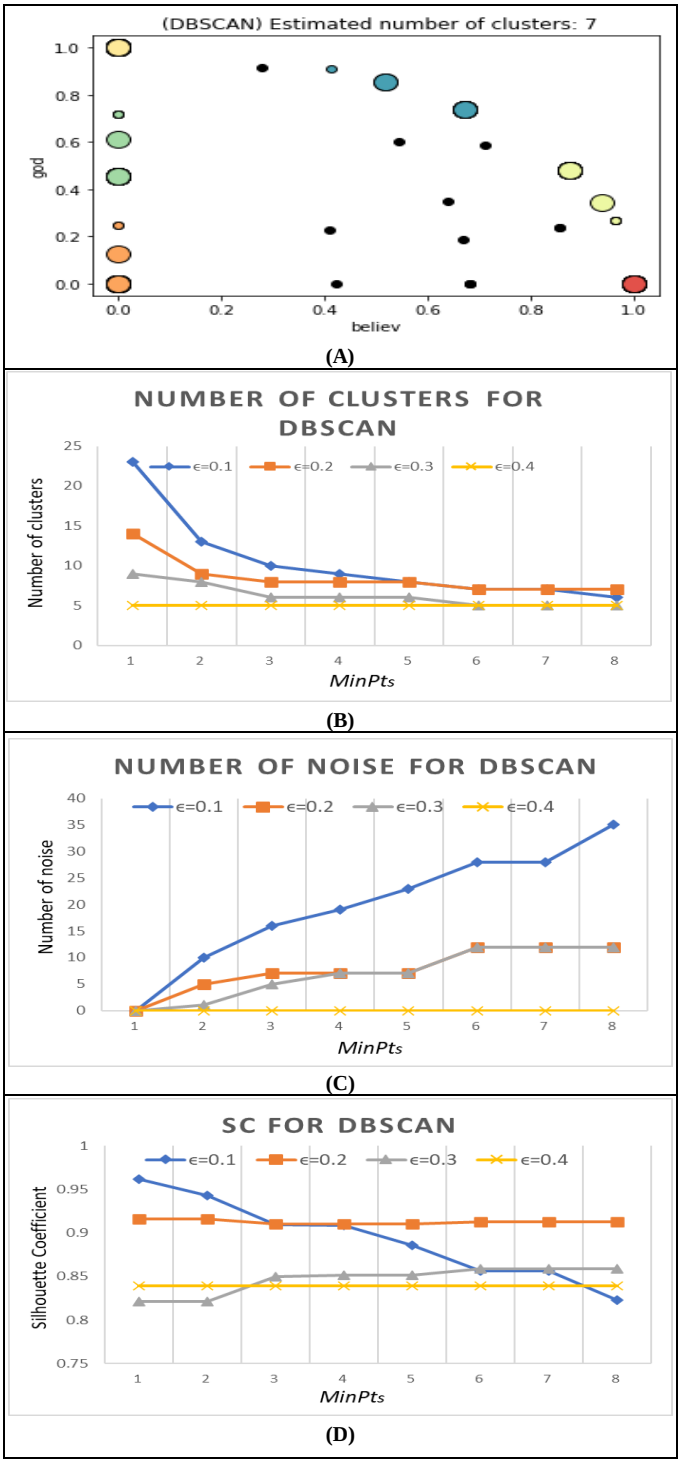

Fig 6. Cluster Visualization for the DBSCAN Algorithm.

- Table I and Fig. 7 illustrate the analysis between these three cluster algorithms regards to (time, SC, number of noises, *Minpts*, and ϵ). From these results, k-means outperforms even has the smallest value of SC compared to others due to less of implementation time with no noise production for seven cluster of Al-Baqarah chapter. OPTICS has no noise with the medium of SC value but has the longest implementation time due to its complexity.

TABLE I. CLUSTERING ANALYSIS BETWEEN THE THREE ALGORITHMS FOR SEVEN CLUSTERS

Matrices	K-means	DBSCAN	OPTICS
Time (seconds)	0.319	0.327	0.935
Silhouette Coefficient (SC)	0.8948	0.9129	0.9098
Number of noise	-	12	-
MinPts	-	7	9
ϵ	-	0.2	-

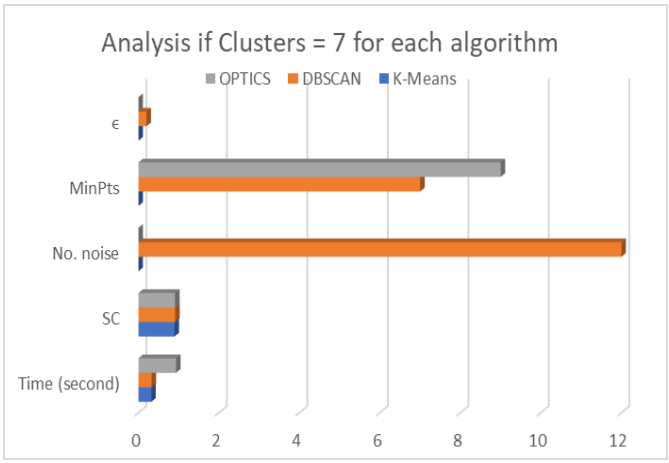

Fig 7. Clustering Analysis between the Three Algorithms for Seven Clusters.

V. CONCLUSIONS

The clustering experiments of this study adapted k-means, DBSCAN, and OPTICS text cluster algorithms to cluster English Tafseer of Al-Baqarah chapter to a seven cluster. Chapter Al-Baqarah has 53 subjects, although some still have the same subject as others. The verses with the same subject are then grouped into seven major themes or topics. This study aims to investigate and figure out which algorithm was the best performance cluster for Al-Baqarah's chapter. The visual and statistics result proves k-means outperforms even has the lowest score of SC compared to others due to less of implementation time with no noise production for seven cluster of Al-Baqarah chapter. DBSCAN has the highest SC score but produced some noises. OPTICS has no noise with the medium of SC value but has the longest implementation time due to its complexity.

This research contributed to researchers and Muslims because the Al-Quran is the reference book. Thus, it is beneficial for Muslims in general, and this research investigated document clustering using many available current techniques to achieve high accuracy and performance. For future work, the authors hope to implement more text cluster algorithms such as the hierarchical method (Agglomerative) and grid-based method (STING) and make more comparisons.

REFERENCES

[1] 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), 2019, pp. 1–5.

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

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

[4] C. Luo, Y. Li, and S. M. Chung, "Text document clustering based on neighbors," Data Knowl. Eng., vol. 68, no. 11, pp. 1271–1288, 2009.

[5] 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.

[6] Indah, N. G. N. Reza,K. Rice, B. V. R. S. Okta, A. Suwanto, S. W. P. N. Tuti, R. Yulia, and Robbi, "DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru," in Journal of Physics: Conference Series, 2019, vol. 1363, no. 1, p. 12001.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] N. Sharma, A. Bajpai, and M. R. Litoriya, "Comparison the various clustering algorithms of WEKA tools," facilities, vol. 4, no. 7, pp. 78–80, 2012.

[15] 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.

[16] P. Deepak and S. Roy, "Optics on text data: Experiments and test results," IBM India Res. Lab, 2006.

[17] J.-S. Fu, Y. Liu, and H.-C. Chao, "ICA: An incremental clustering algorithm based on OPTICS," Wirel. Pers. Commun., vol. 84, no. 3, pp. 2151–2170, 2015.

[18] K. Santhisree and A. Damodaram, "SSM-DBSCAN and SSM-OPTICS: Incorporating a new similarity measure for Density based Clustering of Web usage data," Int. J. Comput. Sci. Eng., vol. 3, no. 9, p. 3170, 2011.

[19] 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.

[20] R. Douglass, "A cluster-based approach to browsing large document collections," in Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992, pp. 318–329.

[21] S. J. Nanda and G. Panda, "A survey on nature inspired metaheuristic algorithms for partitional clustering," Swarm Evol. Comput., vol. 16, pp. 1–18, 2014.

[22] K. Subhadra, M. Shashi, and A. Das, "Extended ACO based document clustering with hybrid distance metric," in 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2015, pp. 1–6.

[23] Y. Kumar and G. Sahoo, "A hybrid data clustering approach based on improved cat swarm optimization and K-harmonic mean algorithm," AI Commun., vol. 28, no. 4, pp. 751–764, 2015.

[24] 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.

[25] 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.

[26] 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.

[27] I. Zeroual and A. Lakhouaja, "A new Quranic Corpus rich in morphosyntactical information," Int. J. Speech Technol., vol. 19, no. 2, pp. 339–346, 2016.

[28] H. Choiruddin, "Klasifikasi Kandungan Al-Qur'an." Jakarta: Gema Insani, 2005.