Annals of Emerging Technologies in Computing (AETiC)
Vol. 7, No. 4, 2023

Review Article

Text Clustering of Tafseer Translations by Using k-means Algorithm: An Al-Baqarah Chapter View

Mohammed A. Ahmed[1*,2], Hanif Baharin[1] and Puteri NE. Nohuddin[1,3]

[1]Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Malaysia.
p103761@siswa.ukm.edu.my; hbaharin@ukm.edu.my; puteri.ivi@ukm.edu.my
[2]Network Engineering Department, College of Engineering, Al-Iraqia University, 10053, Baghdad, Iraq.
mohammed.abdalmunam@aliraqia.edu.iq
[3]Higher Colleges of Technology, UAE
*Correspondence: p103761@siswa.ukm.edu.my

Received: 24th October 2022; Accepted: 12th April 2023; Published: 1st October 2023

Abstract: Al-Quran is Muslims' main book of belief and behaviour. The Al-Quran is used as a reference book by millions of Muslims worldwide, and as such, it is useful for Muslims in general and Muslim academics to gain knowledge from it. Many translators have worked on the Quran's translation into many different languages around the world, including English. Thus, every translator has his/her own perspectives, statements, and opinions when translating verses acquired from the (Tafseer) of the Quran. However, this work aims to cluster these variations among translations of the Tafseer by utilising text clustering. As a part of the text mining approach, text clustering includes clustering documents according to how similar they are. This study adapted the (k-means) clustering technique algorithm (unsupervised learning) **to illustrate and discover the relationships between keywords called features or concepts for five different translators on the 286 verses of the Al-Baqarah chapter. The datasets have been preprocessed, and features extracted by applying** TF-IDF (Term Frequency-Inverse Document Frequency). **The findings show two/three-dimensional clustering plotting for the first two/three most frequent features assigned to seven cluster categories ($k = 7$) for each of five translated Tafseer. The features 'allah/god', 'believ', and 'said' are the three most features shared by the five Tafseer.**

Keywords: *Al-Baqarah chapter; k-means algorithm; Text clustering; Text mining; Tafseer translation*

1. Introduction

Text mining is a branch of science that includes data analysis, statistics, knowledge processing, machine learning, and computational linguistics. A lot of information is generated every day, including digital libraries, research articles, news pieces, books, reviews, and online sites. Getting high-quality data from that whole text data is an important task. Consequently, work was very involved in text mining. Typical text mining duties consist of text classification, extraction of the entity/ concept, text categorisation, sentiment analysis, production of granular taxonomies, summary documentation, and clustering [1-2].

Text clustering is a procedure used in text mining to group documents that are similar in structure into many segments of related documents. Usually, this is performed by finding trends and patterns, which can be accomplished through approaches such as statistical language modelling, statistical pattern learning or topic modelling. In most cases, text mining approaches involve (preprocessing) the input text before it can be applied or implemented [2-3].

Text clustering techniques are similar to data mining techniques. Clustering objects are documents that have been inserted into the weight vector term. Typically, the approaches for clustering algorithms are

Mohammed A. Ahmed, Hanif Baharin and Puteri NE. Nohuddin, "Text Clustering of Tafseer Translations by Using k-means Algorithm: An Al-Baqarah Chapter View", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 27-34, Vol. 7, No. 4, 1st October 2023, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2023.04.003, Available: http://aetic.theiaer.org/archive/v7/v7n4/p3.html.

categorised as partitioning (k-means) [4], hierarchical (network analysis map) [5], density-based (DBSCAN) [6], and grid-based (STING) [2].

Al-Quran is a source of textual facts that requires additional understanding. The original Arabic text of the Quran has been translated into a wide range of languages, including English. Muslims believe Allah revealed the Quran to the Prophet Mohammed (SAW). Surah Al-Baqarah represents the largest chapter of the Quran.; its verses cover a wide variety of subjects. These topics are not listed in chronological order but are related to asbabunnuzul ayat (verses). The verses of Al-Baqarah tend to group in predictable ways because of the text's ability to represent any topic based on its similarity [7].

This study aims to discover relationships and cluster keywords referred to as features or concepts of the Al-Baqarah chapter for five different translators using the clustering technique (k-means) algorithm.

The article is organised as follows: the introduction is presented in section one. Section two discusses this research's relevant literature. Section three explains the methodology used in the research. Section four details experiments. Section five contains the outcomes. Finally, Section six summarises the findings.

2. Related Work

Three cluster algorithms were used in the analysis [7] to identify Qur'anic verses in chapter Al-Baqarah: k-medoid, bisecting, and k-means. Al-Baqarah, which has 286 verses, was interpreted as a text taken from the Qur'an's English translation Tafseer. Three tests of similarity were conducted. Finally, the optimal result was obtained by using cosine similarity with k-medoid.

Slamet *et al.* [8] study established a real preliminary step for learning the patterns of the Holy Quran's verses. Using (k-means) of unstemmed/stemmed words, the algorithm clustered 6236 total verses using (k-means) of unstemmed/stemmed terms, resulting in the construction of three clusters.

To extract keywords and discover connections between them of Tafseer chapters, Chua *et al.* [5] use TF-IDF (Term Frequency-Inverse Document Frequency) (text mining) in a mixture with a network analysis (map) technique. Six brief chapters (Surah) containing 130 keywords were chosen for this experiment from the Malay translation of the Tafseer of the Quran. The KCRA framework is the proposed method.

The article [9] used Java of free software (WEKA) to create a combined Holy Quran answer and question corpus, then visualised the clusters. Data from four separate websites are grouped into four clusters by a probabilistic clustering algorithm.

Putra *et al.* [10] established a semantic-based Question Answering System (QAS) for Indonesian Translation of the Holy Quran (ITQ). The developers asked the participants three research questions and then constructed a weighted vector (TF-IDF) for every concept associated with the expected answer type (also known as an entity group). In order to provide semantic interpreters for the user questions. The author clustered 222 ontological concepts into six, twenty-four, and seventy-seven concepts for Time, Location, and Person.

The research [11] intends to build a website based verses search tool (retrieval of information) for the Al-Qur'an that is linked with (SPC) clustering algorithm to assist Muslims in discovering appropriate information in Al-Quran verses by clustering related Holy Quran verses together.

The paper [12] describes an effort to create a weighted vector for each concept in the ITQ and to develop a semantic-based QAS similar to the research described in [10]. However, the author gives additional information on the TF-IDF results and follows separate procedures of work.

Ahmed *et al.* [13] performed a comparative analysis for OPTICS, k-means, and DBSCAN clustering algorithms using TF-IDF for feature extraction and preprocessing. The results proved that the k-means outperformed other algorithms using time and Silhouette Coefficient (SC) metrics. Moreover, Ahmed *et al.* [14] implemented a comparative analysis for the k-means and the variation of the k-means (Mini-Batch) clustering algorithms using Principal Component Analysis (PCA) and TF-IDF. The experiments presented that the Mini-Batch k-means outperformed k-means in the implementation time. However, both studies [13-14] used the English translation of Tafseer Surah Al-Baqarah as the input dataset.

Finally, Razaque *et al.* [15] used more than one technique, including clustering (k-means), to produce or discover clusters inside the input data to analyse home energy consumption. It offered utilities valuable information to develop customised electricity charges and healthier-targeted energy efficiency initiatives.

3. Methodology of the Research

3.1. The Preprocessing

Data preprocessing is crucial in system construction to ensure that the output produced is more efficient. Data from the study is converted into a format that gives more accurate results [16]. The preprocessing consists of tokenising, case folding, POS Tagging, stemming, and removal of stop-words. Then, applied standardisation process (change all terms in the documents into the lower case) [7].

3.2. The Weighting Processing (Feature Selection)

A feature is defined as a document term, which is significant as important for the document. Text mining applications commonly use features since they are the foundational elements of a document. A set of commonly used term weighting functions in text mining is employed to assess the importance of a candidate feature. There are several popular functions for finding out how essential a term is in the context of a document. The total number of word occurrences in a document (d) is referred to as the (t). The IDF of a text, on the other hand, tests the commonness of a word among all documents. If the term infrequently appears in the document, the IDF number will be increased. [5] [16]. TF- IDF can be expressed in the following equation [3]:

$$TF - IDF(t, d, D) = tf(t, d) * \log(\frac{|D|}{d(t)}) \quad (1)$$

Where *d* and *D* are a documents collection, *t* is the number of times of term occur in the document [17].

3.3. The Clustering Algorithm (k-means)

The clustering (unlike classification) in which the data is divided into groups depends on its similar attributes where data has no previous label (unsupervised learning) [2]. The clustering techniques include partitioning methods such as (k-means) [4], hierarchical (network analysis map) [5], density-based (DBSCAN) [6], and grid-based (STING). This paper adopted the partitioning algorithm (k-means) to accomplish the research aims [2].

k-means clustering is the most effective partial clustering algorithm. This approach uses a partitioning strategy in the clustering process to reduce the gap between data from each cluster core iteratively. The methodology begins with finding a random point of departure for the cluster, converging iteratively (The cluster does not really change) [2].

The steps taken by the k-means approach shall be as follows:

1) Define the value of k (clusters' number). In fact, the value of k should be more than two, and there is no standard method on how many clusters to determine.
2) Define the initial centre (centroid) of the cluster. The original cluster centre is determined by randomly selecting the experimental data.
3) Distance measurement. Using the Euclidean distance equation, determine the data distance from the core clustering point as follows [2]:

   $$D_n(X, Y) = \sqrt{\sum_{n=1}^{m} (X_n - C_n)^2} \quad (2)$$

   where $C_n$is cluster point (n) attribute (n), X_n is data (n) in the attribute (n), m is the total attributes of features, and D_n is the distance of data and cluster.
4) Calculate the shortest distance. After establishing a sufficient space between the data and cluster core, the minimum distance value to become a cluster member for each document is determined. (the assignment of the cluster data).
5) Determination of new cluster centres. When the first iteration generates a cluster, and its members, the new cluster or centre value will be changed by splitting the weights on the same cluster as the following formula [2]:

   $$V_{ij} = \frac{1}{N_i} \sum_{i=1}^{N_i} X_{ij} \quad (3)$$

   where X_{ij} is the k-data value of the cluster for j attribute/variable. j is an index of variables/attributes. i, k are the index of clusters. N_i denotes the quantity of data included within the i cluster, and $V_{ij}$is centroid/i cluster average for j attribute.

6) The iteration process has come to an end. Steps 3 to 6 should be repeated as many times as necessary until the value for either the cluster member or the centroid does not change. The outputs are clusters of similar documents.

4. Experiments

The software platform used to implement this study experiment is the Python version (3.7.3). The experiments described in this paper are divided into four stages. The experiment's flowchart is depicted in Fig 1.

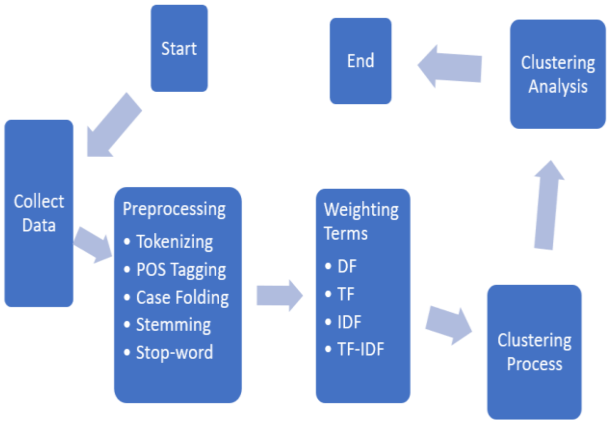

Figure 1. The Flowchart of Experimental Scenario

4.1. The Dataset

Tanzil[1] is an internet site that offers translation documents of the Holy Quran for various languages, including English, by several translators translated from different Tafseer. Many researchers' authors use the data collected from this website [5] [8] [12-14]. The objective of this study is to establish relationships between features and to implement clustering for five different translators, so each translator will be one document; the name of five translators' documents chosen were (1. Ahmed Ali, 2. Ahmed Raza Khan, 3. A. J. Arberry, 4. Abdul Majid Daryabadi, and 5. Muhammad Taqi-ud-Din al-Hilali and Muhammad Muhsin Khan) represents the Tafseer (number) in the experiments of this paper. This study chose chapter Al-Baqarah, which consisted of 286 verses [7].

4.2. Preprocessing

After reading the data, preprocessing involves several stages, including POS tagging, tokenising, stemming, case folding, and stop-word deletion. This stage represents an essential step to make the final result clustering more accurate. Table 1 gives indications about the total number of terms before and after this process for each Tafseer. In each preprocessing stage, this table shows the reduction of terms. Moreover, the table describes the number of terms before and after the stop-word and stemming stages.

4.3. Weighting Terms

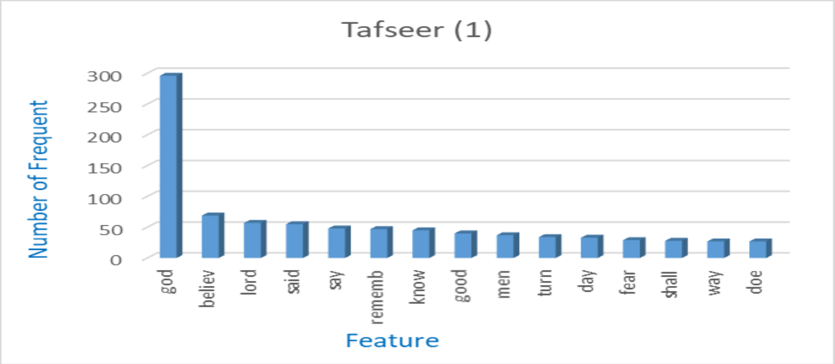

Figure 2. The Most Frequent Terms of Tafseer 1

[1] http://tanzil.net/trans/

The corpus or dictionary will be compiled or assembled after the document goes through the preprocessing. TF-IDF is often used to provide a numerical value (weight) to every feature or term. All weights are combined into a matrix to provide the corpus prepared for clustering. Five corpora were generated for this study's experiments. Fig 2 shows fifteen features of the Tafseer 1 corpus.

4.4. The Clustering Algorithm Parameter

As mentioned, the k-means algorithm is used for the clustering operation. This algorithm needs to define the *k* value; this value should be more than one [2] [7]. In the experiment of this study, the *k* value is seven based on Huda *et al.* [7], Ahmed *et al.* [13-14], and Choiruddin [16] studies for chapter Al-Baqarah. This chapter comprises 53 subjects, some of which share a common theme with others. The verses that share a theme are then combined together to create seven major themes, each with its own number of verses. Fig 3 shows these seven topics or themes.

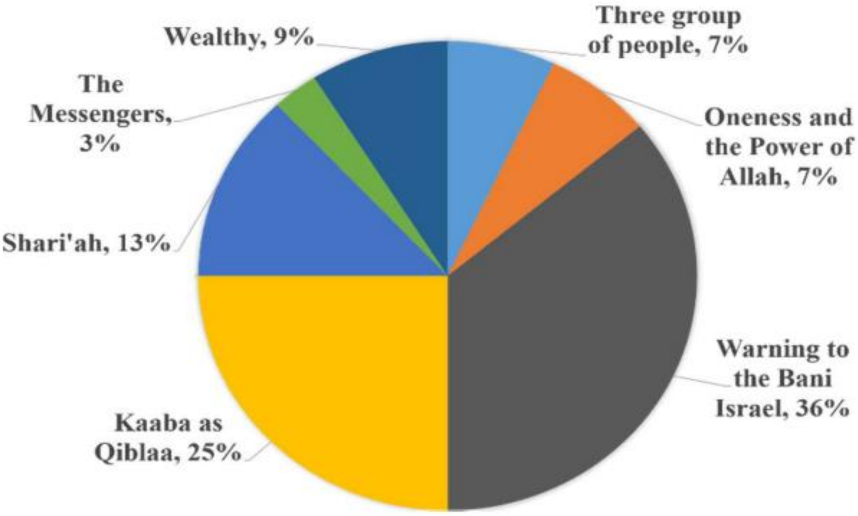

Figure 3. The seven themes of the Surah Al-Baqarah [7]

5. Result and Discussion

The practical summary of the study's findings and results using k-means are:

- The amount of data is five Tafseers (documents).
- Each document has 286 lines or verses from chapter Al-Baqarah Tafseer.
- Implements the preprocessing operation; Table one illustrates the numbers of terms before and after applying the preprocessing operation for the Al-Baqarah chapter.

Table 1. Number of Features and the Preprocessing Steps

Tafseer Number	Total terms	Total terms (stop-words)	Total terms (stop-words + stemming)
1	1742	1508	1221
2	1808	1597	1301
3	1466	1278	1033
4	1591	1392	1198
5	1885	1667	1312

- The parameter *k*=7 has been used for the k-means algorithm.
- Implement a weighting term process. From this, a table that contains the terms with their weight has been produced. Table 2 shows the eight most frequently sorted terms (features) with the number of frequent for each Tafseer.
- From Table 2, ('allah/god', 'believ', and 'said') are the three most common features shared by these five Tafseer, which all refer to 'taqwa' or faith to Allah.

Table 2. The Most Eight Frequent Sorted Terms for Each Tafseer

Tafseer Number	Features' name (frequent number)
1	**god**(296), **believ**(69), **lord**(57), said(55), say(48), rememb(47), know(45), good(50),…
2	**allah**(264), **said**(69), **know**(66), believ(54), people(54), prophet(49), lord(48), say(48),…
3	**god**(288), **shall**(111), **believ**(73), said(70), say(60), lord(49), know(45), thou(44),…
4	**allah** (276), **unto** (225), **ye** (214), shall(151), verili(102),said(73), thou(68),believ(66),…
5	**allah**(349), **believ**(84), **said**(72), say(67), shall (62), know(56), lord(54), muhammed(45),….

- Assign the first most three frequent features to the seven categories (*k*=7) using the k-means clustering algorithm for each Tafseer and draw the results (each verse will automatically get one category).
- Fig 4 illustrates three-dimensional plotting for these three features, assigning to seven cluster categories of each Tafseer.

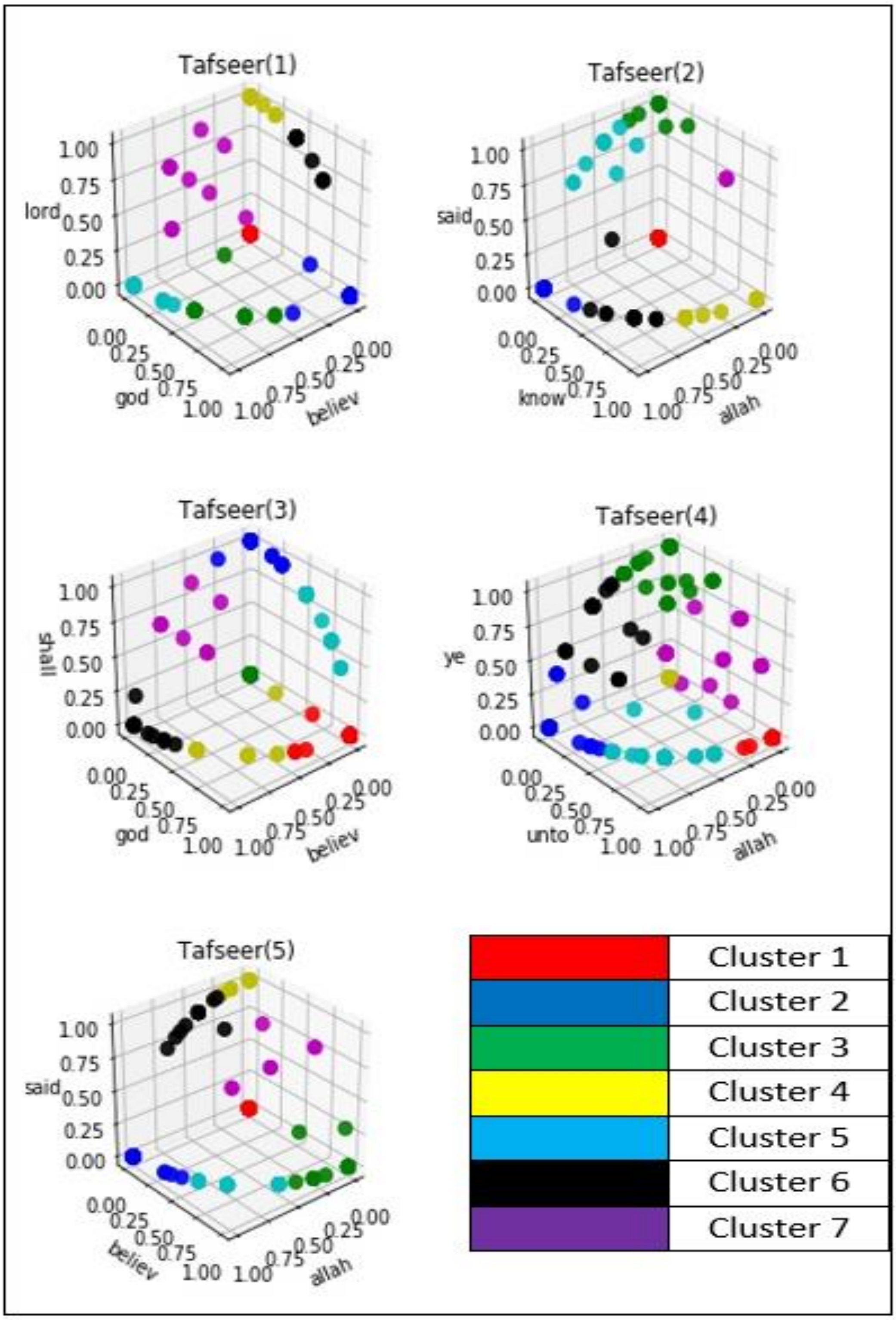

Figure 4. The 3D Clusters Plotting for the five Tafseer Represented First Three Features

- Fig 5 illustrates two-dimensional plotting (scatter) for these two features, assigning to seven cluster categories of each Tafseer. These legends (7 cluster colours) indicate the cluster for features of each Tafseer. Each part of the last two figures illustrates the Tafseer number and this distribution of clusters related to the features of Tafseer.

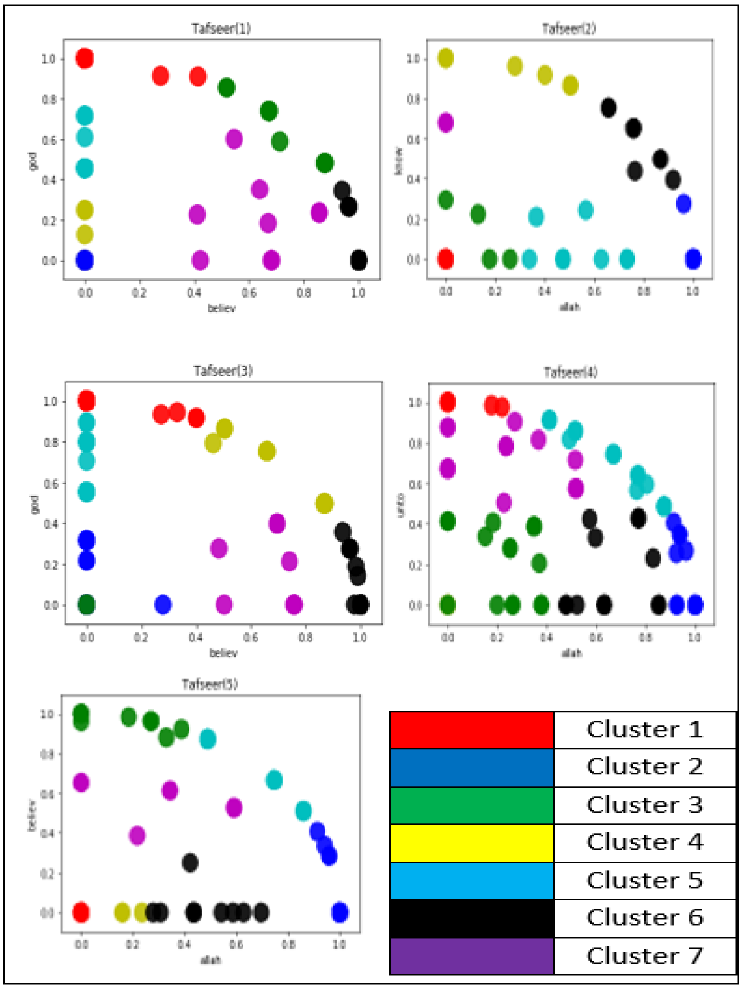

Figure 5. The 2D Clusters Plotting for Two Features of Each Tafseer

6. Conclusion

This paper used one of the text clustering algorithms, the k-means, adopted for the experiments of this study because it is an unsupervised learning algorithm. Al-Baqarah chapter verses are the data used as input to this algorithm. Five documents of five different translators have been applied, and Translators translated verses from Arabic to English using Tafseer. The results illustrate statistics about the frequent term or feature after the preprocessing and weighting processes. Two/three-dimensional clustering plotting implemented showed the first most two/three frequent features assigned to seven cluster categories ($k = 7$) on each of five Tafseer. Features ('allah/god', 'believ', and 'said') are the three most features shared by the five Tafseer.

Since the Al-Quran is the guidebook, this research benefited both researchers and Muslims. Thus, it benefits Muslims in general and Islamic scholars in particular by providing knowledge about the English Tafseer language. The authors hope to extend to other Al-Quran chapters, increase the translator number regardless of the language, and prioritise the Tafseer for a specific Surah for future work.

Acknowledgement

This research is supported by Geran Universiti Penyelidikan (GUP-2018-076) from Universiti Kebangsaan Malaysia. 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] Ammar Kamal Abasi, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, Syibrah Naim, Zaid Abdi Alkareem Alyasseri *et al* , "A novel hybrid multi-verse optimiser with K-means for text documents clustering", *Neural*

Computing and Applications, 2020, DOI: 10.1007/s00521-020-04945-0, Available: https://link.springer.com/article/10.1007/s00521-020-04945-0.

[2] Jiawei Han, Jian Pei and Micheline Kamber, *Data mining: concepts and techniques*, 3rd ed. Massachusetts, USA: Morgan Kaufmann, 2012, ISBN: 9780123814807.

[3] Chengqing Zong, Rui Xia and Jiajun Zhang, *Text Data Mining* , 1st ed. Singapore: Springer, 2021, ISBN: 978-981-16-0099-9, DOI: 10.1007/978-981-16-0100-2, Available: https://link.springer.com/book/10.1007/978-981-16-0100-2.

[4] Congnan Luo, Yanjun Li and Soon M. Chung, "Text document clustering based on neighbors", *Data & Knowledge Engineering*, pp. 1271–1288, Vol. 68, No. 11, 2009, DOI: 10.1016/j.datak.2009.06.007, Available: https://www.sciencedirect.com/science/article/abs/pii/S0169023X09000974.

[5] S. Chua and P. N. E. Nohuddin, "Relationship Analysis of Keyword and Chapter in Malay-Translated Tafseer of Al-Quran", *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, pp. 185–189, Vol. 9, No. 2–10, 2017, Available: https://jtec.utem.edu.my/jtec/article/view/2724/1771.

[6] Mustakim, Reza N. Gayatri Indah, Rice Novita, Oktaf Brillian Kharisma, Rian Vebrianto *et al.*, "DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru", *Journal of Physics: Conference Series 2019*, p. 12001, Vol. 1363, No. 1, DOI: 10.1088/1742-6596/1363/1/012001, Available: https://iopscience.iop.org/article/10.1088/1742-6596/1363/1/012001/meta.

[7] A. F. Huda, M. R. Deyana, Q. U. Safitri, W. Darmalaksana, U. Rahmani *et al.*, "Analysis Partition Clustering and Similarity Measure on Al-Quran Verses", in *Proceedings of the 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, 2019, Yogyakarta, Indonesia, pp. 1–5, DOI: 10.1109/ICWT47785.2019.8978215, Published by IEEE, Available: https://ieeexplore.ieee.org/abstract/document/8978215.

[8] Cepy Slamet, Ali Rahman, Muhammad Ali Ramdhani and Wahyudin Darmalaksana, "Clustering the verses of the Holy Qur'an using K-means algorithm", *Asian Journal of Information Technology*, Online ISSN: 1682-3925, pp. 5159–5162, Vol. 15, No. 24, 2016, Published by Medwell Journals, Available: http://digilib.uinsgd.ac.id/id/eprint/5117.

[9] Bothaina Hamoud and Eric Atwell, "Quran question and answer corpus for data mining with WEKA", in *Proceedings of the 2016 Conference of Basic Sciences and Engineering Studies (SGCAC)*, 2016, Khartoum, Sudan, pp. 211–216, DOI: 10.1109/SGCAC.2016.7458032, Available: https://ieeexplore.ieee.org/abstract/document/7458032.

[10] Syopiansyah Jaya Putra, Ria Hari Gusmita, Khodijah Hulliyah and Husni Teja 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, Available: https://dl.acm.org/doi/abs/10.1145/3011141.3011219.

[11] Zul Indra, Arisman Adnan and R. Salambue, "A Hybrid Information Retrieval for Indonesian Translation of Quran by Using Single Pass Clustering Algorithm", in *Proceedings of the 2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019, Semarang, Indonesia, pp. 1–5, DOI: 10.1109/ICIC47613.2019.8985737, Published by IEEE, Available: https://ieeexplore.ieee.org/abstract/document/8985737.

[12] Syopiansyah Jaya Putra, Khodijah Hulliyah, Nashrul Hakiem, Rayi Pradono Iswara and Asep Fajar 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, Jakarta, Indonesia, pp. 293–297, DOI: 10.1145/3011141.3011218, Available: https://dl.acm.org/doi/abs/10.1145/3011141.3011218.

[13] Mohammed A. Ahmed, Hanif Baharin and Puteri NE. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster Algorithms on Al-Quran Verses", *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 11, No. 8, pp. 248-254, 2020, Available: https://thesai.org/Downloads/Volume11No8/Paper_32-Analysis_of_k_means_DBSCAN_and_OPTICS.pdf.

[14] Mohammed A. Ahmed, Hanif Baharin and Puteri NE. Nohuddin, "Mini-Batch k-Means versus k-Means to Cluster English Tafseer Text: View of Al-Baqarah Chapter", *Journal of Quranic Sciences and Research (JQSR)*, E-ISSN: 2773-5532, Vol. 2, No. 2, pp. 48-53, 19 December 2021, DOI: 10.30880/jqsr.2021.02.02.006, Available: https://publisher.uthm.edu.my/ojs/index.php/jqsr/article/view/9924.

[15] Fahad Razaque, Nareena Soomro, Javed A. Samo, Huma Dharejo and Shoaib Shaikh, "Analysis of Home Energy Consumption by K-Mean", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 1-6, Vol. 1, No. 1, 2017, DOI: 10.33166/AETiC.2017.01.001, Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3504157.

[16] Aris Tri Jaka Harjanta, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining", *Jurnal Informatika Upgris*, Vol. 1, No. 1, 2015, DOI: 10.26877/jiu.v1i1%20Juni.804, Available: http://journal.upgris.ac.id/index.php/JIU/article/view/804.

[17] Choiruddin Hadhiri S. P, *Klasifikasi Kandungan Al-Qur'an*, 3rd ed. Jakarta, Indonesia: Gema Insani Press, 1994, Available: https://opac.perpusnas.go.id/DetailOpac.aspx?id=229562.