

# Text mining approaches for analyzing an Indonesian tafseer and translation of the Holy Quran

Media Anugerah Ayu, Edi Irawan, Teddy Mantoro

Department of Computer Science, Faculty of Engineering and Technology, Sampoerna University, Jakarta, Indonesia

## Article Info

### Article history:

Received Sep 14, 2021

Revised Jan 5, 2022

Accepted Jan 12, 2022

### Keywords:

Association rule

K-means clustering

Most frequent words mining

Tafseer text mining

Text mining

## ABSTRACT

The Indonesian tafseer and translation of Holy Quran is an important source of information and knowledge for Indonesian muslims, since not many Indonesian muslims understand Arabic language in the Quran. However, the tafseer is full of the commentaries and explanation of each surah (chapter) and/or ayah (verse), which form a large document and not so easy to be accessed. Thus, the challenge is how to refer to both tafseer and translation in faster and accurate ways as one needs to always refer to them back and forth. Hence, this study proposes several text mining approaches, i.e. most frequent words, K-means clustering, and association rules, to analyze an Indonesian tafseer and translation of Quran and provide insights of hidden knowledge and relationships based on statistical information derived from it. These insights could be useful for muslims in general and for people that doing research in related areas. This study shows interesting results from combined analysis of the approaches used which can help people accessing information in tafseer more efficiently. As well, interesting relationships have been drawn from terms in the tafseer which could provide further and deeper knowledge on messages in the Quran.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Media Anugerah Ayu

Department of Computer Science, Faculty of Engineering and Technology, Sampoerna University

L'Avenue Building, Street Raya Pasar Minggu Kav. 16 Pancoran, South Jakarta, Indonesia

Email: media.ayu@sampoernauniversity.ac.id

## 1. INTRODUCTION

In the recent years, natural language processing (NLP) has been widely used for the automation related with translation or interpretation. Within NLP there is text mining which considered as one of its branches as it is using some fundamental methods in NLP but with different goals. Unlike NLP which cares about semantics information in the text, in the text mining there is also a method which treats the text as the 'bag of word', meaning the semantics information is not explored. The main goal in text mining is to analyze both unstructured and structured large text dataset so that one does not have to read the whole text [1]. This has lead Text Mining in becoming a valuable research area as the existing improvement of artificial intelligence (AI) has been on the level where the extraction of information in a textual data has to be automated. The result from text mining is the information of the terms and words analysis. Many large text data artifacts have become the data source for research in the text mining area. One of those large text data sources is the Holy Quran.

The Holy Quran is the most valuable book for muslims, i.e. people with Islamic religion, as they believe it is containing the words of God. Inside the Quran, there are fundamental categories of knowledge which have to be understood and recited by all muslims [2]. The original language of Quran is Arabic. Since many muslims do not understand Arabic properly, Quran has been translated into many languages, including

Indonesian language, to make its contents easier to be understood. However, to some extent, translation of Arabic Quran only is not enough for general people to really understand the exact meaning of messages in the Quran. That is why there are some honorable and knowledgeable people who write and create some commentary books of the Quran called mufassir and the books called as tafseer. The explanation and commentary inside the tafseer must not base on the individual opinion. The contents of the Quran must stay the same so all the commentary must refer to the explanation from the Prophet Muhammad [3].

As tafseer and translation of the Quran dealing with long sentences and words, it becomes a challenge to extract the valuable information from both of them. In this technological era, people tend to leave conventional thing such as refer to a thick book by opening one page to another page. The invention of information retrieval algorithm and text mining in natural language processing (NLP) has enabled people to mine valuable information inside large text documents faster automatically and might be a possible answer to the referencing tafseer and translation challenge. There are several Quran-related NLP studies, for examples the ones by [4]-[6]. However, there are rarely found NLP studies on the Indonesian tafseer of the Quran, whereas this tafseer has great importance for muslims in understanding the contents of the Quran, especially the ones with little arabic language knowledge. Thus, this research study aims to utilize text mining techniques to retrieve the insights of the Indonesian tafseer, to uncover hidden knowledge and relationships of materials discussed in the Quran.

The organization of this paper starts with an introduction in section 1, which then followed by section 2 which presents reviews on some related works from previous research in the literature. After that, methodology applied in the study is discussed in section 3, and then section 4 presents the results of the conducted research complete with its discussions. Last section is the conclusion which presents the summary of the key findings and its takeaways.

## 2. LITERATURE REVIEW

Generally, there are three types of research areas in the text mining, i.e. techniques for preprocessing, comparative studies about machine learning for both classification and clustering as well as the feature extraction algorithm comparison, and the study about the text dataset exploration result for the mining. Many of the studies on text mining in general are concentrated on the preprocessing stage of the text mining. This is due to the needs for further improvement in preprocessing since it is a crucial stage which can affect to the result significantly.

The preprocessing includes tokenization, normalization and substitution. Besides the preprocessing, the selection of the methods also is one of the trends in the research area. The researchers usually compare two or more common method in text mining, whether it is about clustering or classification [7], [8]. Another research area for text mining is to implement the text mining to a specific dataset with the focus on that dataset like a research work done by Alhawarat *et al.* [9] on Arabic language dataset and conducted research by Matsumoto *et al.* [10] on combining numerical and text dataset. Moreover, there are also studies in comparing two or more distance calculation techniques in determining the similarities when doing clustering or classification [11].

Quran has also been a subject of text mining as one of dataset sources. However, text mining research in Quran are not only focus on the dataset, it can also accommodate all of those general three types mentioned earlier and a combination between them. Researchers can study the algorithm used for the Quran text mining. Within this type of research, the researchers can compare two or more algorithm to extract the most valuable information inside the Quran. Several text mining studies on Quran explored and analyzed the classification of its content as reported by [12]-[16]. Another type of research in Quran text mining is focus in the specific dataset and analyzing the text mining result acted to the datasets, which are Indonesian Tafseer and Translation. There are also previous related works about text mining for Quran and Tafseer related with different goals among them as the works by [2], [5], [17], [18].

As the Quran contains chapters and already decided in the past, researchers want to explore the rule that made the division of the Quran. A good example is the work done in [5] with the goal to do the analysis on the frequent patterns that can be found in the chapters of a Malay translated tafseer of Quran; the techniques are frequent pattern mining, non-trivial patterns and interesting relations. The findings of the study were the processed dataset: 6 documents and 17 terms. The term weighting is term frequency-inverse document frequency (TF-IDF). Three most frequent terms are “Allah”, “Muhammad”, and “wahai”. The different type of research is presented by Khadangi *et al.*, [4] which intended to study the similarity of topics in Quranic surahs; the methodology was natural language processing methods which are word2vec and roots' accompaniment in Verses. The finding was the knowledge that the choice of the surah's title is based on rational logic, the surahs hold the inner coherence between the concepts so that they have formed on a single topic or a few topics tightly related to each other [4].

An analysis of a text mining algorithm on Quran is presented by Qi *et al.* [19] looked through the semantic information inside the Quran. The objective was to contribute in building an algorithm with semantic analysis and automatic identification areas. The research compared and analyzed semantically between Chinese and Arabic written language of Quran. The algorithm used in the research study was Semantic annotated corpus and semantic knowledge base.

There was also a study which explored the Quran Tafseer in Malay Language. The aim was to provide classification algorithm for Quran Tafseer verses automatically. This research study by Hamoud and Atwell [18] used K-nearest neighbor (KNN) or classifier and cosine similarity as the distance. The result of the study was a contribution to Malay Quran tafseer category classification. From this study we can learn that one way to contribute in NLP study of Quran, is to strengthen the algorithm in building a good tidy corpus. Another study went to that direction and did research in the exploration of making the corpus to build the tagging algorithm for creating a prototype which is able to extract collocation of N-gram words [17]. This N-gram words consist of 2 until 6 words from Arabic Quran corpus ordered by part of speech tagging. The result showed that the proposed system succeeded to make the users select a sequence of tags (2-6 gram) and scope of the corpus source. In addition, a study to reveal frequent patterns in Holy Quran (Arabic) using text mining has been reported in [20] that can be used to analyze further the Quran and bring more comprehensive understanding. Among those explored research studies within NLP-text mining related to Quran, we have not found the one which focuses in Indonesian tafseer of Quran. Since Indonesia is a country with the biggest number of muslims in the world, and not many Indonesia can understand Arabic well, then a technology-based approach like text mining that can help in extracting hidden knowledge from Quran through its tafseer will be beneficial.

### 3. METHODOLOGY

There are several steps conducted for the text mining process applied in this study, as presented in Figure 1. This whole process was conducted for tafseer and translation with the same steps. The dataset used was from: KEMENAG Indonesian tafseer and translation, “tahlili” 2011 version all Juz. The tool used for feature selection until frequent term mining is R and RStudio 3.6.3 as the IDE.

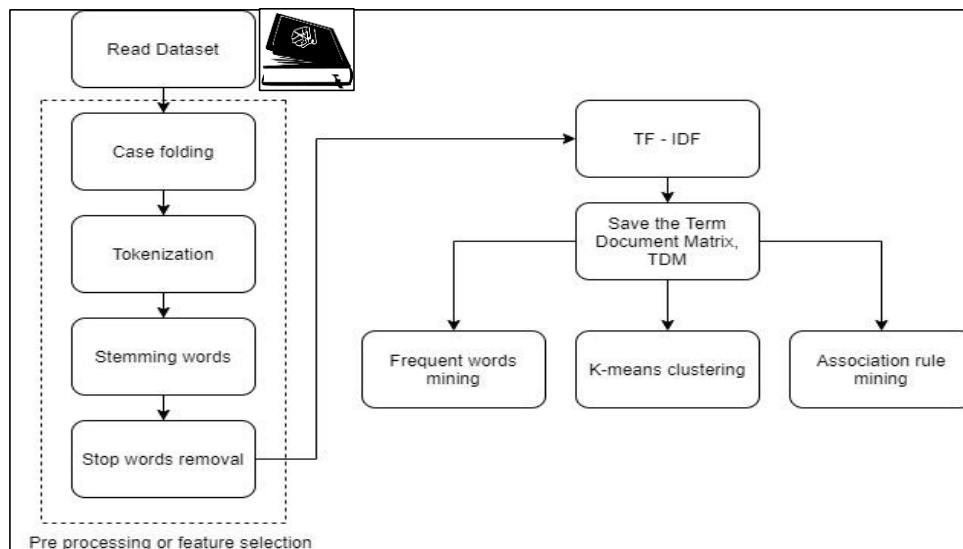


Figure 1. The text mining process for the Indonesian tafseer and translation of the Quran

#### 3.1. Preprocessing or feature selection

The preprocessing or feature selection stage includes case folding, tokenization, stemming words and stop words elimination. Preprocessing is needed to reduce the unwanted words which have no significant meaning, noise, into text mining. This step also done to reduce the redundancy and repetition. Those steps are reversible and can go back to any step if it is required.

#### 3.2. Feature extraction using TF-IDF

The TF-IDF is considered as one of the most powerful feature extractions [21]; it is because unlike the bag of word method, this method is not only seeing the most frequent terms so that the undominant word

is eliminated; the TD-IDF is also weighting the terms based on how frequent the term in a document compared to how frequent the term in the whole documents. By doing TF-IDF, the most frequent word is rescaled. The mathematical model for the TF-IDF is shown in (1) [21].

For a term  $i$  in the document  $j$ :

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

where:

$tf_{i,j}$  = Number of occurrence in  $i$  and  $j$

$df_i$  = Number of documents containing  $i$

$N$  = Total of documents

### 3.3. Most frequent words mining

In this stage, the most frequent words are extracted from both tafseer and translation. The result of the most frequent words measured by TF is represented and visualized in the form of word clouds. The other presentation of the result, which is the frequency measured by TF-IDF is in the form of the bar plot of each tafseer and translation result. Not only seeing the most frequent words, the result is also evaluated in terms of its correlation using pearson correlation coefficient.

### 3.4. K-means clustering

The clustering in this study was performed based on Euclidian distance between terms or words. The Euclidian distance in (2).

$$\|A - B\| = \sqrt{\sum_{i=1}^d (a_i - b_i)^2} \quad (2)$$

where A and B are points in d dimensional space such that:  $A = [a_1, a_2, \dots, a_d]$  and  $B = [b_1, b_2, \dots, b_d]$ .

After getting each distance, then the clustering methods are applied. The K-means algorithm is one of the partitional clustering, meaning the clusters dataset are fully divided from the others and treated as different cluster. The first thing to do in K-means clustering is assigning the number of clustering,  $k$ . After that, initially, the random centroid for  $k$  cluster is chosen. The iteration of K-Means is done until the mean of each training data to the centroid met the stopping criterion, whereas the smallest Euclidean distance from a sample is the nearest centroid for the sample to be the one with [22], [23].

In order to present the best clustering results, preliminary experiments were done. One of the approaches to know the optimal number of  $k$  is by seeing the elbow of sum square of error (SSE) of cluster center plot. Thus, in the  $k$ -means clustering stage, preliminary experiments were conducted to get the best valuer of  $k$ , before the main clustering process was done.

### 3.5. Association rules mining

Originally, frequent pattern (FP) growth algorithm is used for knowing the association rules in the relational database of transaction. The formal definition of association rule was presented by Agrawal *et al.* [24] as the following description. Let  $I = I_1 + I_2 + \dots + I_m$  be a set of items or binary attributes. Let  $D$  be a set of all transactions where each transaction  $T$  is a set of items such that  $T \subseteq I$ . Let  $X, Y$  be a set of items such that  $X, Y \subseteq I$ . From those definitions, there is the association rule implication which presented in the form  $X \Rightarrow Y$ , where  $X \subset I, Y \subset I, X \cap Y = \emptyset$  [24].

When dealing with association rules, there are two values which need to be analyzed, which are support and confidence values. In the case of *Support*, if  $s\%$  of transactions in  $D$  contain  $X \cup Y$  then association rule for  $X \Rightarrow Y$  be having  $s$  as the support value; whereas for the case of *Confidence*, if  $c\%$  of the transactions in  $D$  that contain  $X$  also contain  $Y$  then the association rule for  $X \Rightarrow Y$  be having  $c$  as the confidence value. Association rules mining can also be used to capture positive and negative association among the items based on their frequency of appearance, eventhough major association rules tend to go for the positive association [25].

## 4. RESULTS AND DISCUSSION

This section presents the results of the conducted research and discussion related to it. There are five sub-sections here, where each sub-section discusses results from each step performed in the text mining of the tafseer and translation of Quran. The five sub-sections are preprocessing results, feature extraction results, most frequent words mining results, K-means clustering results, and association rules results.

**4.1. Preprocessing or feature selection results**

When it comes to feature selection result, the datasets would not form any meaningful sentences anymore as some words taken away from the datasets. Figure 2 shows samples of the results from feature selection stage on data from tafseer in Figure 2(a) and translation in Figure 2(b). It can be seen from the presented samples that each word has been tokenized, cased folded into uppercase, and stemmed. The data in Figure 2 also shows that there are some different words and some similar words as the results of the pre-processing onto the Tafseer and the Translation. Further processes, like clustering (in Section 4.4) and association rules (in Section 4.5) will be able to show what can be revealed from those differences and similarities.

SURAH	PENDAPAT	ULAMA	ABU	AHLI	ALLAH	AYAT
134.71901	134.11199	66.08931	158.08545	98.83268	808.36051	553.46830
MEDINAH	SALAT	HADIS	NABI	ALBUKHARI	MUSLIM	RIWAYAT
50.42218	196.86536	362.28900	527.68617	115.59964	154.48807	218.49652
UMAR	MEKAH	ABBAS	RASULULLAH	AHMAD	HURAIRAH	SAHABAT
75.60364	125.21297	75.93314	393.16143	76.56092	82.72732	63.75376
ALQURAN	MELARANG	MANUSIA	MUHAMMAD	PERINTAH	TUHAN	WAHYU
322.96000	58.01810	437.15789	342.15315	236.57726	332.61349	87.30771

(a)

DOSA	NYATA	BESAR	KIAMAT	KITAB	MENGIKUTI	TAKUT	MENDUSTAKAN	DUNIA
107	139	175	174	249	110	146	138	181
KAFIR	LAKILAKI	RAHMAT	KEHIDUPAN	AIR	MALAM	SURGA	KEKAL	MELIHAT
414	136	118	177	143	116	164	102	232
PAHALA	NERAKA	CIPTA	MATI	REZEKI	TAQWA	BERPALING	MENGATAKAN	KEBAJIKAN
101	274	162	119	109	221	108	115	129
MENDENGAR	HATI	AKHIRAT	ANAK	GOLONGAN	KAUM	FIRAUN	MUSA	NEGERI
119	194	139	150	120	207	111	235	128
SALAT	HARTA	FIRMAN	PERBUATAN	MALAIKAT	IBRAHIM	SETAN	PEREMPUNAN	
100	133	154	107	133	126	110	176	

(b)

Figure 2. Samples of preprocessing or feature selection results of tafseer and translation, (a) feature selected samples of tafseer and (b) feature selected sample translation

**4.2. Feature extraction results**

TF-IDF algorithm was used in this feature extraction process. Table 1 shows the matrix property of the term document matrices (TDM) of the tafseer and translation dataset. The tafseer contains 488 significant terms for the TF-IDF calculation while translation have 116 terms. These terms are presentend as the columns of the term document matrix and the occurrence of each term is weighted from each document. The total documents, or in this case sentences, of the tafseer was 18450 and the translation was 6234. The non-sparse entries of each matrix show as the nonzero entries and the sparse entries are as the zeros entries. The maximal length in tafseer was14 words of each document and 13 words of each document for translation. The visualizations of word TF-IDF are presented in Figure 3 for both tafseer in Figure 3(a) and translation in Figure 3(b). The two figures show similar curve for the TF-IDF values. There are around 4 words or terms which have significant difference values compared to the others. Further discussion about those numbers is presented on the next section, i.e. most frequent word mining results.

**4.3. Most frequent words mining results**

Since in the feature extraction stage TF-IDF was used for weighting the term frequency, then this most frequent words mining is another automatic result from the TF-IDF algorithm. Figure 4 shows the bar plot of the 30 most frequent words in the tafseer Figure 4(a) and translation Figure 4(b) measured by TF-IDF, respectively. Based on the TF-IDF definition, those words are the most likely to appear in each sentence of the tafseer and translation. Previous work studying frequent items in tafseer of the Quran in Malay [6] has reported 17 words that frequently appeared in the tafseer, which were: aku, Allah, apabila, berlindung, katakanlah, kejahatan, makhluk, manusia, masuk, menguasai, Muhammad, orang, pula, sekalian, tuhan, ugama, dan wahai. The study also reported that “Allah”, “muhammad”, and “wahai” are the most frequent ones among those 17 items. Comparing to our results as presented in Figure 4, there are some words which are intersection between them: Allah, Muhammad, tuhan, manusia, agama (note: ugama in Malay). Only those five words are found in both works. This indicates the importance of those five words in the Quran and its tafseer in different languages. Whereas for other words which are not in the intersection, it could be due to the difference in the way of explaining the meaning of the ayah, which made the words usage was not the same as well.

Results in Figure 4 present frequent words in tafseer Figure 4(a) and translation Figure 4(b) of Quran. In order to know the level of correlation between tafseer and translation, the calculation of pearson correlation coefficient needs to be done. The correlation observation is performed on the mutual words between the tafseer and translation, to see whether the pattern is the same or not. The pattern observation is on how much the tendency of the frequency of a particular word in tafseer and translation being affected by each other.

Table 1. The matrix property of TDM of the tafseer and translation

Data source	Terms	Documents	Non-sparse entries	Sparse entries	Maximal length
Tafseer	488	18450	234693	8768907	14
Translation	116	6234	18815	704329	13

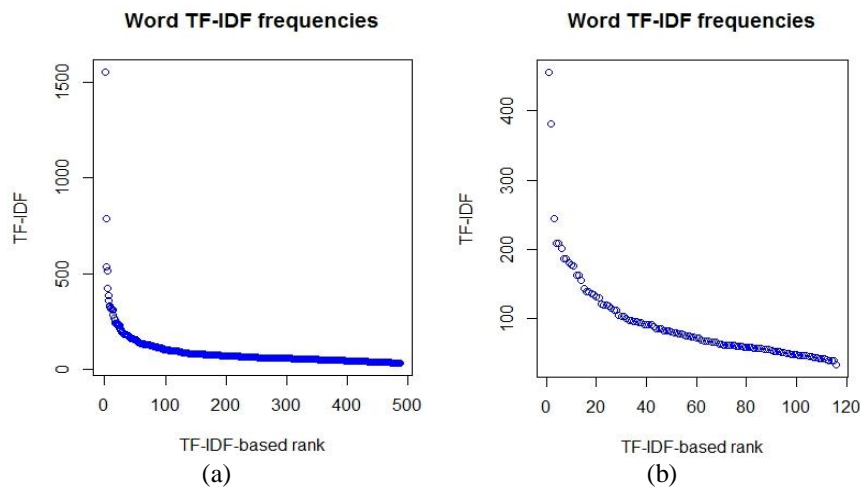


Figure 3. TF-IDF term frequency plots of the tafseer and translation, (a) tafseer TF-IDF plot and (b) translation TF-IDF plot

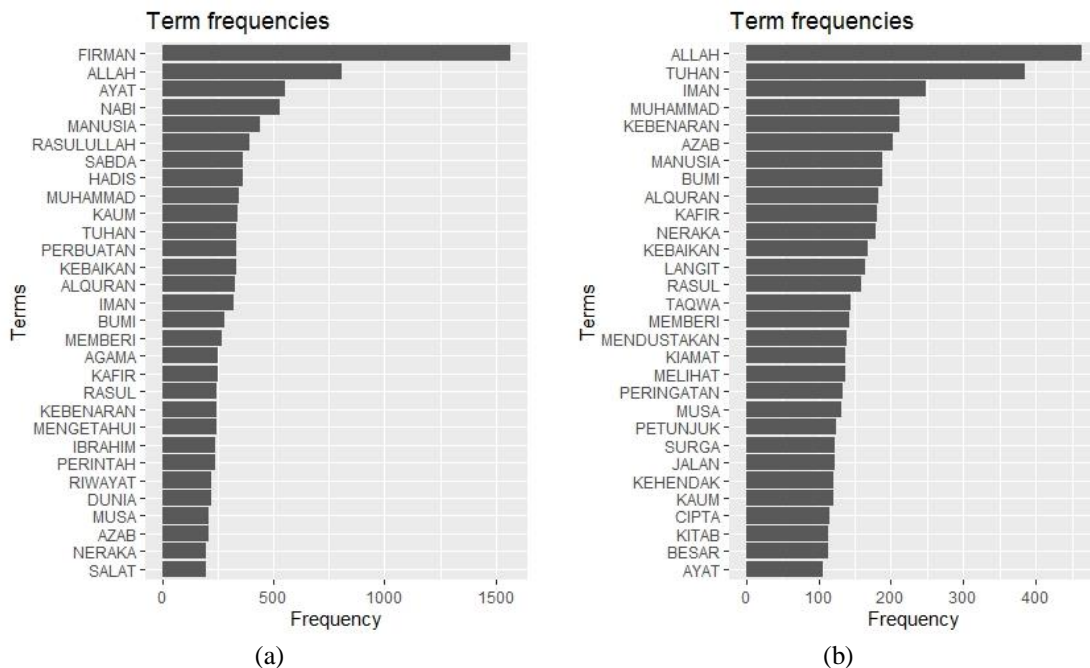


Figure 4. The 30 most frequent words in the tafseer and translation, (a) most frequent words in tafseer and (b) most frequent words in translation

The result of the pearson correlation coefficient value is 0.5306. The 0.5306 value means a positive moderate correlation of words occurred inside both the tafseer and the translation. This means that there is tendency that the higher frequency of the word occurred in tafseer, the higher frequency of that word occurred in translation, and vice versa. Thus, even though there are some differences on the most frequent words between the tafseer and translation, there is always a tendency that the same words occurred in both of them. This information is beneficial in ensuring that the tafseer and translation version are having the same directions. In other words, one can trust to refer from both tafseer and translation of this version due to the same pattern of the words.

**4.4. K-means clustering results**

In the K-means clustering phase, the initial step was to determine the K value that would be used in the clustering process. The determination of the best K value was done based on its SSE evaluation. Figure 5 shows the SSE Cluster Center Plot for tafseer in Figure 5(a) and translation in Figure 5(b), respectively. As can be seen on presented graphs in Figure 5, the best K value for the tafseer is on K=10 and for translation is on K=8. Thus, this study focuses on analyzing the results of 8 and 10 clusters for both of the datasets.

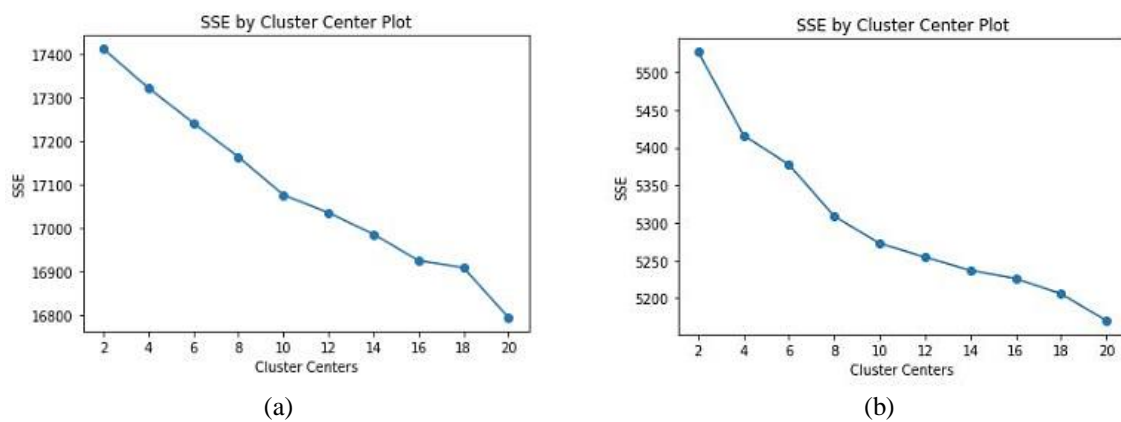


Figure 5. SSE vs cluster centers plot for tafseer and translation, (a) tafseer and (b) translation

Table 2 shows the clustering result for K=8 of the tafseer and translation. It should be noted that the clustering number is not ordered and does not matter in the clustering case. The cluster 1 of the tafseer shows words “kitab”, “anak”, “sihir”, “mesir”, “agama”, “harun”, “Tuhan”, “Firaun”, “Bani Israil”, and “Musa”. This shows a good example of one clustering. Referred to the Quran, the Prophet Musa story is narrated. The story is about the duty of Prophet Musa to remind Firaun and accompanied by Prophet Harun. The place was in Egypt or “Mesir” where the bad magic or “Sihir” was popular at that time.

Table 2. Clustering results with K=8 clusters

Cluster	Tafseer	Translation
1	Kitab, Anak, Mukjizat, Sihir, Mesir, Agama, Harun, Tuhan, Firaun, Israil, Bani, Musa	Petunjuk, Rasul, Benar, Mengingkari, Kafir, Azab, Mendustakan, Kitab, Quran.
2	Hati, Peringatan, Ajaran, Hukum, Petunjuk, Agama, Kitab, Quran	Kitab, Kiamat, Peringatan, Rasul, Musa, Kafir, Neraka, Azab.
3	Tanah, Planet, Gunung, Siang, Hujan, Kekuasaan, Bulan, Malam, Benda, Alam, Tanda, Matahari, Air, Menciptakan, Bintang, Mahluk, Malaikat, Langit, Bumi	Baik, Kebesaran, Hujan, Tanda, Air, Gunung, Langit, Bumi
4	Tirmidzi, Imam, Ibnu, Ismail, Ahmad, Bukhari, Hurairah, Abu.	Karunia, Kafir, Hati, Petunjuk, Muhammad, Hamba, Rasul, Beriman.
5	Balasan, Pahala, Kehidupan, Berhala, Dosa, Hamba, Kafir, Nikmat, Neraka, Surga, Amal, Azab, Akhirat, Dunia.	Kenikmatan, Bertaqwa, Kebajikan, Kekal, Petunjuk, Lurus, Sungai, Surga, Manusia.
6	Hati, Peringatan, Ajaran, Hukum, Petunjuk, Agama, Kitab, Quran	Petunjuk, Janji, Jalan, Azab, Firaun, Kitab, Rasul, Muhammad, Tanda, Benar
7	Dosa, Kiamat, Perempuan, Laki, Tempat, Kafir.	Berdoa, Rahmat, Quran, Azab, Pengampun, Pengasih, Penyayang.
8	Istidraj, Istiadat, Israfil, Israil, Istana, Istilah, Isteri, Petunjuk, Malaikat	Beriman, Istri, Azab, Akhirat, Yatim, Dunia, Harta, Nikmat, Perempuan, Laki



Seeing the members of the cluster 3 in the tafseer, this cluster contains astronomical terms, for example “Planet”, “Bulan”, “Bintang”, “Matahari”, “Langit” and “Bumi”. This cluster might be a partition about the perspective of universe creation from Quran. Another interesting cluster result in tafseer is the cluster 4 which are the name of hadith narrators, which make a good cluster as well. In cluster 5 of the tafseer, the words quite interesting, as it contains pair of opposite words, such as “neraka” and “surga”, “hamba” and “kafir”, or “pahala” and “dosa”. It can be seen here that in the tafseer, the bad and the good are narrated together in one case so they become close and get into one cluster. Cluster 6 also shows a good example of one cluster theme because of their topic closeness, about Quran and Kitab as laws of Muslims which already discussed in the literature review.

The results of clustering from the translation are not as clear as the tafseer. There are some same words appeared on each cluster, for example, the word “Azab” and “Petunjuk” which makes it difficult to decide the main topic of each cluster. Then, the information which can be acquired is that the distance of each word in translation are not really far from each other. Meaning, using TF – IDF weighting method, the term most likely appears on each document the same amount of times.

Next, observation was done to the results presented in Table 3 for K=10. For the case of tafseer, there are five clusters which are similar with the previous result. Whereas for the case of translation, it starts to get clearer for some clusters. As for examples, the words “mata”, “air”, “balasan”, “baik”, “taman”, “buah”, “penghuni”, “kenikmatan”, “mengalir”, “kekal”, “sungai”, and “surga” are go to one cluster in translation. However, overall, determining theme of the translation cluster result is still not easy to be decided.

Table 3. Clustering results with K=10 clusters

Cluster	Tafseer	Translation
1	Munafik, Larangan, Kemenangan, Musuh, Yahudi, Ibrahim, Mekah, Perang, Musyrik, Kafir.	Hati, Langit, Hamba, Tobat, Bumi, Quran, Muhammad, Beriman
2	Istidraj, Isteri, Israfil, Israil, Istana, Petunjuk, Malaikat	Gembira, Perjalanan, Kafir, Celakalah, Manusia, Beriman, Kebenaran, Peringatan.
3	Negeri, Tanda, Nuh, Setan, Nikmat, Hati, Kiamat, Kafir, Neraka.	Malaikat, Hamba, Benar, Sahaya, Istri, Anak, Perempuan, Laki.
4	Nikmat, Hamba, Quran, Ajaran, Surga, Umat, Baik, Kesenangan, Kebahagiaan, Neraka, Kafir, Kehidupan, Azab, Hidup, Akhirat, Dunia.	Sapi, Takut, Bumi, Malam, Firaun, Harun, Kekuasaan, Kebesaran, Tanda.
5	Isa, Hud, Esa, Musyrik, Sembah, Berhala, Patung, Tuhan.	Diutus, Azab, Umat, Yatim, Nuh, Harta, Anak, Rasul.
6	Dawud, Abdullah, Umar, Tirmidzi, Imam, Ahmad, Ibnu, Bukhari, Muslim, Abu, Hurairah, Sabda.	Mata, Air, Balasan, Baik, Taman, Buah, Penghuni, Kenikmatan, Mengalir, Kekal, Sungai, Surga.
7	Jalan, Kebajikan, Buruk, Balasan, Sifat, Isteri, Ibu, Surga, Saleh, Pahala, Perempuan, Laki, Hamba, Harta, Amal, Dosa, Anak.	Air, Golongan, Dunia, Gunung, Waktu, Negeri, Muhammad, Malaikat, Baik, Kiamat, Manusia, Azab, Neraka.
8	Harun, Hati, Sibir, Umat, Petunjuk, Kaum, Mukjizat, Kebenaran, Taurat, Firaun, Bani, Israil, Kitab, Musa.	Petunjuk, Nikmat, Muhammad, Dustakan, Azab, Kafir.
9	Tumbuhan, Planet, Bulan, Kekuasaan, Tanda, Benda, Tanah, Gunung, Alam, Matahari, Hujan, Cipta, Mahluk, Binatang, Air, Langit, Bumi.	Azab, Kafir, Kerajaan, Tanda, Janji, Rasul, Bumi, Langit, Besar
10	Ibrahim, Menyampaikan, Hamba, Mahluk, Utusan, Laki, Wahyu, Jibril, Lut, Adam, Malaikat.	Puji, Zalim, Disembah, Esa, Langit, Bumi, Azab, Pengasih.

Results from K-means clustering have shown that with K=8 the created clusters from the tafseer have converged to obvious themes. However, the case for translation was different, where the created clusters have not shown clear themes. The major reason for tafseer to show clear grouping in each cluster is that tafseer usually narrated and described similar topics into one story, such as story of Musa and Firaun, Quran as Muslims Law, and Astronomical Creatures. These kinds of structures were not the case for the translation. Translation more into just translating the saying from Arabic to Indonesian for each ayah in the surah, which not always converge to similar topic.

#### 4.5. Association rules results

Results of interesting association within the tafseer and translation are presented in this section. Figure 6 shows the association of the word “Allah” from translation dataset. Except the word “kafir”, all of associations are showing positive sentiments. High support values are shown from the association of word “Allah” with “memberi” which means “to give”, “petunjuk” which means ‘guidance’, and “penyayang” which means “loving”. The support value of the association with word “kafir” which means “non-believer” is 0.004 and the confidence is 0.957, meaning from the whole translation documents, 0.4% occurrence together



the term “Allah” and “kafir” in one document. In addition, 95.7% of the documents in the translation contain term “Allah” also contain “kafir”. To see the meaning of this, further referencing is done by looking up into the translation dataset. One sample from this association is Surah An-Nahl Ayah 106–107. The ayahs show that Allah always narrate on how the bad fate would come to kafeer, which are people who deny the truth of Allah.

The other thing is, there are words having several different support and confidence values when they are associated with different words as well. For example, when the word “jalan” associated with only the word “Allah”, the support and confidence value is 0.005 and 0.810. However, when the word “jalan” associated with word “Allah” and also word “kebaikan”, those values are 0.003 and 0.833. The other association of the word “jalan” is with “kafir” by values 0.04 and 0.957. This kind of occurrences also applied to some other words.

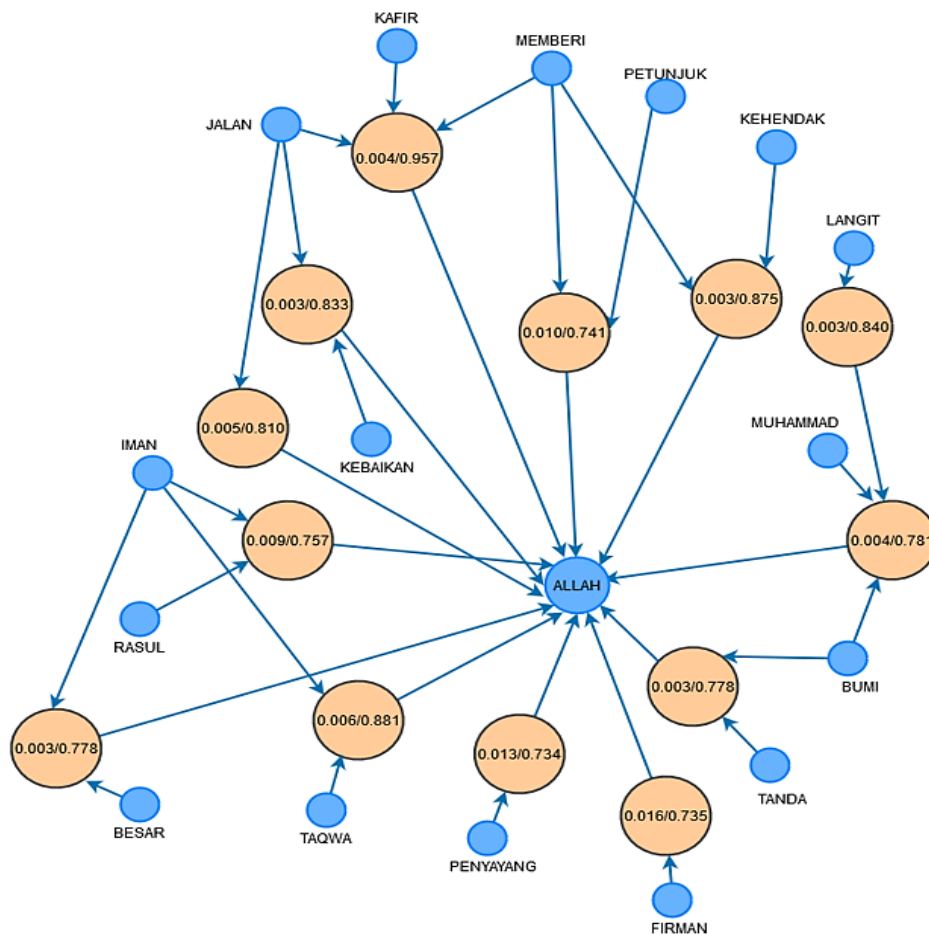


Figure 6. A result of association rules process on translation dataset

For the case of tafseer, Figure 7 shows the results from the association rules. The word “israil”, “musa”, and “bani” are in the same values of support and confidence, which is 0.1% of the whole documents contain their union and 54% of the documents contains those words. To compare with the previous clustering result, this is also related to the cluster 1 which contains those words and the word “Harun”.

From both association rule results, it can be observed that there are relations with the previous results, frequent pattern mining and clustering result. The sequence example of the information retrieval from this result is, after knowing that “Musa” is one of the most frequent word in the tafseer and translation, then one can find the cluster “Musa” in the clustering result. Next, further information about the association of each word in that cluster can be determined by this result. By doing this sequence, knowledge that the Prophet Musa did a duty from Allah to remind the Firaun can be revealed. As well, about Musa who then

asked Allah Azza Wa Jalla, that he wanted his brother, Prophet Harun to accompany him in this duty. Also, information that those took place in Mesir which is Egypt now.

Several benefits can be drawn from knowing these association rules results. The first possible benefit is to enable the Islamic scholars and muslims to know and/or reveal connections in a certain topic that they would like to learn further. For example, say one wants to know about Prophet Musa by referring to Indonesian Tafseer. Without knowing the association rule, he/she might just focus only to the word “Musa” in the tafseer and have to read the whole sentences in the tafseer about “Musa” to be able to draw valuable information about Prophet Musa. However, by knowing and having the association rules list of the word “Musa”, insight knowledge will be able to be gained faster. For instance, using (“Musa”, “Mesir”)-> “Agama” or (“Musa”, “Bani Israil”)-> “Agama”, one can take a look at those words to focus in searching information about Prophet Musa.

Another benefit is in business, specifically online bookstores or libraries. Say one user accesses to an online book store or library and is interested in book tagged “Musa” as the keyword. Then, the systems could be able to give what kind of books that might interests the user and create a preference book for the user. Because in tafseer the word “Musa” has association rule with “Mesir” and/or “Bani Israil”, the systems can give suggestion and recommendation for books which tagged with words “Mesir” and/or “Bani Israil”. Of course, to be able to do that, it needs further process. However, that is the general thing that the association rules can provide further assistance in business area.

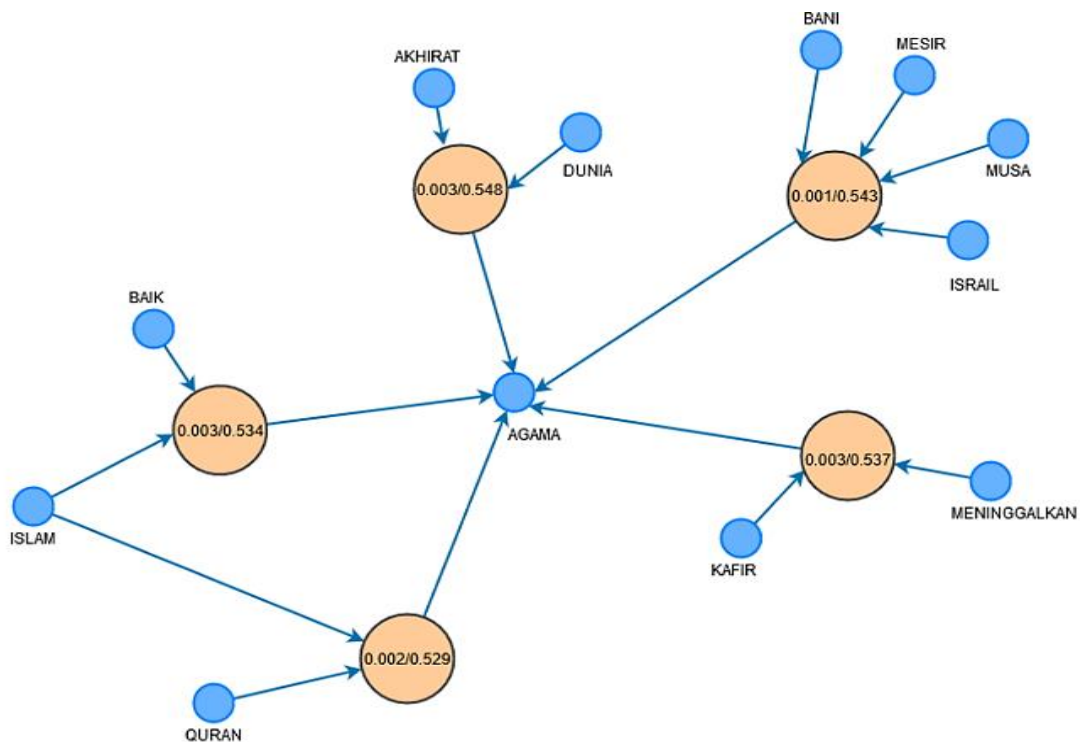


Figure 7. A result of association rules process on tafseer dataset

## 5. CONCLUSIONS




This research study has conducted a text mining on Indonesian tafseer and translation of Quran through several approaches, i.e. most frequent words, K-means clustering, and association rules. Valuable information from tafseer and translation is succeeded to be obtained through the text mining perspective. The 30 most frequent words inside the tafseer and translation were presented and showing 17 mutual words from tafseer and translation occurred in the 30 ranking. The correlation result shows that the mutual words from tafseer and translation having 0.5306 value, meaning there is tendency that the higher frequency of the word occurred in tafseer, the higher frequency also occurred in translation, vice versa. This result tells us that the tafseer and translation of this version most likely to exchange information with the similar meaning and there is less natural language processing problem in both datasets. Then, the clustering results of tafseer and translation are obtained using the K-Means technique. The best partition result shown by the tafseer with

K=8 clusters. However, the result from translation dataset did not show as good as tafseer partition. The clustering result from tafseer solved the unstructured dataset problem. By knowing the clusters, one can differentiate each topic based on their similarities, thus the information retrieval can be more relevant and efficient. Furthermore, the results from the association rules show some interesting relations between terms or words in the tafseer and translation. Results from all those three approaches actually supported each other. The frequent words mining shows word “Musa”, for example, and it appears in one of the good cluster partitions containing “Musa”, “Firaun”, and others. Then, the association among those words can be seen and measured by association rule result. The sample of rules in tafseer showed by the association (“Musa”, “Bani”, “Israil”) -> “Agama”. These show that combining the three approaches could lead to ways on retrieving more information and revealing more insight knowledge from Quran and its tafseer.




## REFERENCES

- [1] M. Zhou, N. Duan, S. Liu, and H. Y. Shum, “Progress in neural NLP: modeling, learning, and reasoning,” *Engineering*, vol. 6, no. 3, pp. 275–290, Mar. 2020, doi: 10.1016/j.eng.2019.12.014.
- [2] S. Yilmaz Genç and H. Syed, “Quranic principles of universal law on the Quranic exegesis,” *Bilimname*, pp. 165–186, Dec. 2019, doi: 10.28949/bilimname.592756.
- [3] S. M. Al-Qaththan, *MABAHITS FI ULUMIL QURAN*. Pustaka Al-Kautsar, 2000.
- [4] E. Khadangi, M. M. Fazeli, and A. Shahmohammadi, “The study on Quranic surahs’ topic sameness using NLP techniques,” in *2018 8th International Conference on Computer and Knowledge Engineering, ICCKE 2018*, Oct. 2018, pp. 298–302, doi: 10.1109/ICCKE.2018.8566248.
- [5] G. Sabah, S. Khaotijah, and F. Mahdi, “Categorization of ‘Holy Quran-tafseer’ using K-Nearest Neighbor Algorithm,” *International Journal of Computer Applications*, vol. 129, no. 12, pp. 1–6, Nov. 2015, doi: 10.5120/ijca2015906909.
- [6] S. Chua and P. N. E. Binti Nohuddin, “Frequent pattern extraction in the Tafseer of Al-Quran,” in *2014 the 5th International Conference on Information and Communication Technology for the Muslim World, ICT4M 2014*, Nov. 2014, pp. 1–5, doi: 10.1109/ICT4M.2014.7020667.
- [7] R. N. Waykole and A. D. Thakare, “A review of feature extraction methods for text classification,” *International Journal of Advance Engineering and Research Development*, vol. 5, no. 04, pp. 351–354, 2018.
- [8] P. K. Jayasekara and K. S. Abu, “Text mining of highly cited publications in data mining,” in *IEEE 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, ETLIS 2018*, Feb. 2018, pp. 128–130, doi: 10.1109/ETLIS.2018.8485261.
- [9] M. Alhawarat, M. Hegazi, and A. Hilal, “Processing the text of the Holy Quran: A text mining study,” *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 2, 2015, doi: 10.14569/ijacsa.2015.060237.
- [10] T. Matsumoto, W. Sunayama, Y. Hatanaka, and K. Ogohara, “Data Analysis support by combining data mining and text mining,” in *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*, Jul. 2017, pp. 313–318, doi: 10.1109/IIAI-AAI.2017.165.
- [11] D. Agnihotri, K. Verma, and P. Tripathi, “Pattern and cluster mining on text data,” in *Proceedings - 2014 4th International Conference on Communication Systems and Network Technologies, CSNT 2014*, Apr. 2014, pp. 428–432, doi: 10.1109/CSNT.2014.92.
- [12] B. S. Arkok and A. M. Zeki, “Classification of Quranic topics based on imbalanced classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 678, May 2021, doi: 10.11591/ijeecs.v22.i2.pp678-687.
- [13] N. S. Elmitwally and A. Alsayat, “The multi-class classification for the first six surats of the Holy Quran,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 327–332, 2020, doi: 10.14569/ijacsa.2020.0110141.
- [14] A. Adeleke, N. A. Samsudin, Z. A. Othman, and S. K. Ahmad Khalid, “A two-step feature selection method for quranic text classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 730–736, Nov. 2019, doi: 10.11591/ijeecs.v16.i2.pp730-736.
- [15] M. N. Al-Kabi, H. A. Wahsheh, and I. M. Alsmadi, “A topical classification of Hadith Arabic Text,” in *NOORIC 2013: Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, 2013.
- [16] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, “Comparative analysis of text classification algorithms for automated labelling of Quranic verses,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 4, pp. 1419–1427, Aug. 2017, doi: 10.18517/ijaseit.7.4.2198.
- [17] W. Alromima, I. F. Moawad, R. Elgohary, and M. Aref, “Extracting N-gram terms collocation from tagged Arabic corpus,” in *2014 9th International Conference on Informatics and Systems, INFOS 2014*, Dec. 2015, pp. NLP10–NLP15, doi: 10.1109/INFOS.2014.7036700.
- [18] B. Hamoud and E. Atwell, “Quran question and answer corpus for data mining with WEKA,” in *Proceedings of 2016 Conference of Basic Sciences and Engineering Studies, SGCAC 2016*, Feb. 2016, pp. 211–216, doi: 10.1109/SGCAC.2016.7458032.
- [19] J. Qi, Y. Yu, L. Wang, J. Liu, and Y. Wang, “An effective and efficient hierarchical K -means clustering algorithm,” *International Journal of Distributed Sensor Networks*, vol. 13, no. 8, p. 155014771772862, Aug. 2017, doi: 10.1177/1550147717728627.
- [20] A. Aslani and M. Esmaeili, “Finding frequent patterns in Holy Quran using text mining,” *Signal and Data Processing*, vol. 15, no. 3, pp. 89–100, 2018, doi: 10.29252/jsdp.15.3.89.
- [21] S. Qaiser and R. Ali, “Text mining: Use of TF-IDF to examine the relevance of words to documents,” *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.
- [22] V. Friedemann, “Clustering a customer base using twitter data,” *Cs*, vol. 229, no. 1, pp. 1–5, 2015.
- [23] K. Korovkinas, P. Danenas, and G. Garšva, “SVM and K-means hybrid method for textual data sentiment analysis,” *Baltic Journal of Modern Computing*, vol. 7, no. 1, pp. 47–60, 2019, doi: 10.22364/bjmc.2019.7.1.04.
- [24] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, Jun. 1993, doi: 10.1145/170036.170072.
- [25] S. Mahmood, M. Shahbaz, and A. Guergachi, “Negative and positive association rules mining from text using frequent and infrequent itemsets,” *Scientific World Journal*, vol. 2014, pp. 1–11, 2014, doi: 10.1155/2014/973750.




**BIOGRAPHIES OF AUTHORS**

**Media Anugerah Ayu**    is currently a Professor in Computer Science Department, Sampoerna University. She earned a PhD degree in Information Science and Engineering from the Australian National University (ANU), Australia. She has published many research papers in international journals, conferences, book chapters and books in IT related areas, where 104 of them have been indexed by Scopus. She has served as committee members in international conferences and reviewers for international journals. She can be contacted at email: [media.ayu@sampoernauniversity.ac.id](mailto:media.ayu@sampoernauniversity.ac.id).



**Edi Irawan**    was graduated from Sampoerna University with a bachelor degree in Computer Science. He has actively involved in several research in machine learning area. Beside machine learning topic, he also ever involved in research about Multi-Agents System. As well, he has published some papers through international conferences. In addition, he was also a Class Tutor for Object Oriented Programming course. He is currently working as a Technical Consultant for a business process management and technology company that is singularly focused on the travel and aviation sector. He can be contacted at email: [irawan.edi12@gmail.com](mailto:irawan.edi12@gmail.com).



**Teddy Mantoro**    a Professor in Computer Science at Sampoerna University, Indonesia. His research interest is in the area of Pervasive Computing, Wireless Sensor Networks, Context-Aware Computing, Mobile Computing, Information Security, and Intelligent Environment/IoT. He has been working in Intelligent Environment which uses Computational Intelligence for quite a while. He has published 180 conference/journal papers. He obtained a PhD, an MSc, and a BSc, all in Computer Science, and his PhD was from School of Computer Science, the Australian National University (ANU), Canberra, Australia. He can be contacted at email: [teddy.mantoro@sampoernauniversity.ac.id](mailto:teddy.mantoro@sampoernauniversity.ac.id).