

Arabic Text Clustering Methods and Suggested Solutions for Theme-Based Quran Clustering: Analysis of Literature

Qusay Bsoul

Cybersecurity Department, Science and IT College, Irbid National University, Irbid, Jordan
E-mail: q.bsoul@inu.edu.jo

Jaffar Atwan

Prince Abdullah Bin Ghazi Faculty of ICT, AL-Balqa Applied University, Salt, Jordan
E-mail: jaffaratwan@bau.edu.jo

Rosalina Abdul Salam*

Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai, Negeri Sembilan, Malaysia
E-mail: rosalina@usim.edu.my

Malik Jawarneh

Faculty for Computing Sciences, Gulf College, Muscat, Sultanate of Oman
E-mail: malik@gulfcollege.edu.om


ABSTRACT

Text clustering is one of the most commonly used methods for detecting themes or types of documents. Text clustering is used in many fields, but its effectiveness is still not sufficient to be used for the understanding of Arabic text, especially with respect to terms extraction, unsupervised feature selection, and clustering algorithms. In most cases, terms extraction focuses on nouns. Clustering simplifies the understanding of an Arabic text like the text of the Quran; it is important not only for Muslims but for all people who want to know more about Islam. This paper discusses the complexity and limitations of Arabic text clustering in the Quran based on their themes. Unsupervised feature selection does not consider the relationships between the selected features. One weakness of clustering algorithms is that the selection of the optimal initial centroid still depends on chances and manual settings. Consequently, this paper reviews literature about the three major stages of Arabic clustering: terms extraction, unsupervised feature selection, and clustering. Six experiments were conducted to demonstrate previously un-discussed problems related to the metrics used for feature selection and clustering. Suggestions to improve clustering of the Quran based on themes are presented and discussed.

Keywords: text mining, Arabic text clustering algorithms, terms extraction, un-supervised feature selection, optimal initial centroid

Received: May 4, 2021
Accepted: October 10, 2021

Revised: September 22, 2021
Published: December 30, 2021

***Corresponding Author:** Rosalina Abdul Salam
 <https://orcid.org/0000-0002-0893-2473>
E-mail: rosalina@usim.edu.my



All JISTaP content is Open Access, meaning it is accessible online to everyone, without fee and authors' permission. All JISTaP content is published and distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>). Under this license, authors reserve the copyright for their content; however, they permit anyone to unrestrictedly use, distribute, and reproduce the content in any medium as far as the original authors and source are cited. For any reuse, redistribution, or reproduction of a work, users must clarify the license terms under which the work was produced.

1. INTRODUCTION

The use of text clustering as an analysis tool for detecting and understanding Arabic text enables people to understand Quran text. Users can start understanding the Quran by going through the words or verses. The Quran consists of thirty chapters and each chapter includes many Quranic verses. Each chapter belongs to more than one theme such as Zakat (alms), Inheritance, Prayers, and others. Verses of each chapter are required to be grouped based on the clustering of themes. Therefore, Arabic text clustering becomes very important as an analysis tool for everyone in understanding the Quran. This is the motivation for clustering the themes of the Quran. The main reason why text clustering has not yet been used to group the Quran text is the degraded poor performance of the three tasks: terms extraction, unsupervised feature selection. Some tasks use text mining, such as sentiment analysis (Touahri & Mazroui, 2021), Twitter topic detection (Motaghinia et al., 2021), concept map construction (Qasim et al., 2013), and systematic review support (Ananiadou et al., 2009), but it is still of great significance to increase the performance of text clustering. The Quran is an essential guide to the life of all humankind, both Muslims and non-Muslims. It covers all aspects of human life, including biology, information technology, law, social order, politics, business, economics, and individual responsibilities (Rostam & Malim, 2021). In short, the Quran represents a sea of knowledge (Yauri et al., 2013).

The great knowledge incorporated in the Quran stresses the importance of exploring such holy texts using automated applications. In fact, the information technology is a system consisting of a medium, infrastructure, and methods for gaining, transferring, accessing, interpreting, saving, organizing, and using data meaningfully (Azad & Deepak, 2019). People who seek to understand Islam could find easier ways to find answers to their various queries (Abualkishik et al., 2015). However, understanding the Quran requires more than just reading its translation. Hence, making the content of the Quran available to everyone is still a big challenge. Therefore, it is important to take one's level of experience into account when presenting the content of the Quran.

The aim of this study is to investigate the limitation of text clustering, because there is a need to increase its quality. Although text clustering has been studied for many years, it is still an important research domain and its methods require further improvements (Alghamdi & Selamat, 2019). The rest of the paper is structured as: Sec-

tions 2 and 3 provide a review of text clustering and the problems of its application. However, Section 4 presents possible research directions and the proposed solutions for extraction, unsupervised feature selection, and clustering. Finally, Section 5 concludes the paper.

2. TEXT CLUSTERING FOR ARABIC LANGUAGE AND THE QURAN

Text clustering is one of the most commonly used methods for detecting themes or types of documents (Alghamdi & Selamat, 2019). It involves three main processes (Bsoul et al., 2014, 2016a, 2016b): The first process is document pre-processing, which is needed for any type of document such as finance, medical, or law documents. It removes unimportant words and symbols from the documents. The second process is to extract the most important terms and weights from the features of the documents then calculate the similarities among them. Then, feature selection is employed to select the optimal features for the clustering algorithm. The last process is clustering. There are two main types of cluster techniques, hierarchical and partitioning, as shown in Fig. 1. The organization of the paper is shown in Fig. 2. This work focuses on the domain of Quran themes and the main weaknesses of Arabic text clustering.

In fact, there are a number of aspects that are involved in understanding the Quran, such as the reasons of revelation, the knowledge of *Makkee* and *Madanee* (that is, the places where a chapter "surah" was revealed), the knowledge of the various forms that were revealed, understanding the Quran's abrogated rulings and verses, the various classifications of its verses, the knowledge of the themes, and the analysis of its grammar (Harrag, 2014). Recently, many studies have explored how Quranic text can be stored, processed, and extracted to further understand its content. The Quranic ontology (Beirade et al., 2021) and other approaches can be found in Raharjo et al. (2020) and Farhan et al. (2020). Other studies have used information retrieval (IR) (Raharjo et al., 2020; Safee et al., 2016) or text classifiers to detect verses or chapters in the Quran (Safee et al., 2016). The main objective of this study is to provide people with better solutions to understand the Quran and to disseminate knowledge using recent methods and technology. Hence, this study focuses on employing Arabic text clustering to cluster the themes of the Quran based on the highest similarity between the verses. In the next section, the process of Arabic text clustering and its limitations are presented in detail.

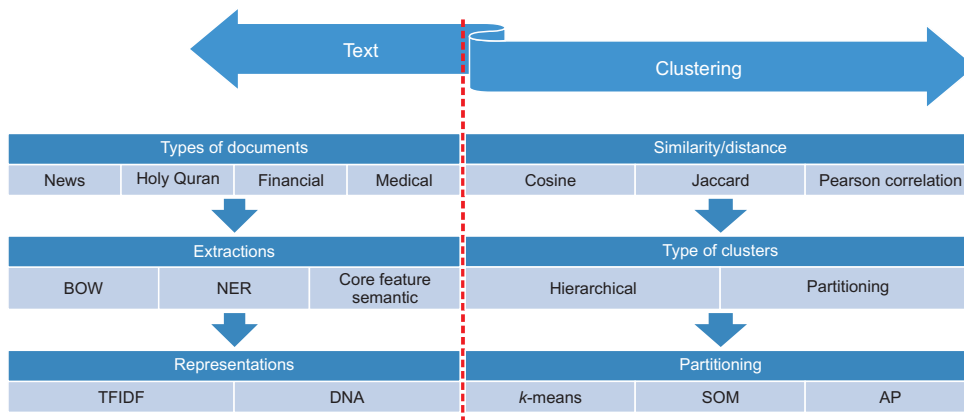


Fig. 1. Text clustering structure. BOW, Bag of Words; NER, named entity recognition; TFIDF, term frequency-inverse document frequency; DNA, deoxyribonucleic acid; SOM, self-organizing map; AP, affinity propagation.

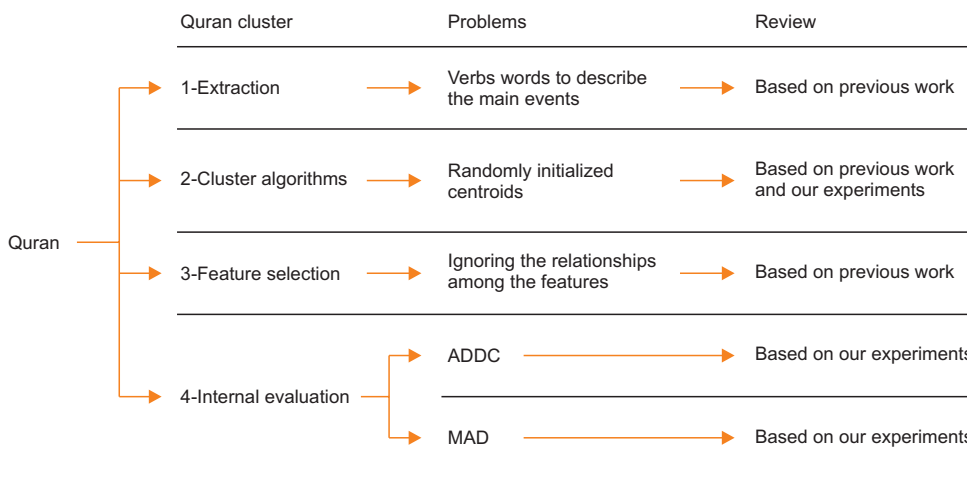


Fig. 2. Organization of the paper. ADDC, average distance of documents to the cluster centroid; MAD, mean absolute difference.

3. TEXT CLUSTERING PROCESS

This section reviews comprehensive previous related work, namely the three phases of text clustering: terms extraction, unsupervised feature selection, and clustering.

3.1. Terms Extraction

The extraction of Arabic words is one of the processes undertaken in Arabic Cluster (Alghamdi & Selamat, 2019), which functions to extract the most important terms needed to retrieve pertinent information. Consequently, various techniques have been suggested in the literature for such purposes, which are categorized into three main categories: Bag of Words (BOW), Bag of Concepts (BOC), and Bag of Narratives (BON) (Saloot et al., 2016). Several researchers have focused on extracting information from the terms by employing named entity recognition (Mesmia et al., 2018), bag-of-words (BOW) (Hmeidi et al., 2008), n-grams (Al-Salemi & Aziz, 2011), and lemmatization algorithms (Al-Shammari & Lin, 2008;

Al-Zoghby et al., 2018). To make the terms extraction stage become effective, researchers used ontology-based extraction (Al-Zoghby et al., 2018), semantic extraction (Al-Zoghby et al., 2018), Arabic word sense disambiguation (Elayeb, 2019; Salloum et al., 2018), semantic word embedding (El Mahdaouy et al., 2018), and semantic relationships (Benabdallah et al., 2017). However, text data often presents challenges because of their high dimensionality and ambiguous or overlapping word senses. Previous studies have proposed various terms extraction methods such as BOW, N-grams, and named entity recognition to address the high dimensionality of the terms. These methods are called syntactic extraction. Overlapping word senses like the semantic words found in Word-Net are called semantic extraction. In this paper, we will discuss syntactic extraction and why researchers are moving towards developing and using semantic extraction.

Al-Smadi et al. (2019) have worked specifically on extraction based on morphological, syntactic, and semantic extract Arabic terms using supervised machine

learning, which is founded that the syntactic better than morphological, and semantic. Besides this, Harrag (2014) employed Named entity approach to extract the most important Arabic terms, this approach have limitation in distinguished between two similar sentences. For that limitation Helwe and Elbassuoni (2019) their approach relies only on word embedding from the Arabic name entity to made difference between two similar sentences.

Al-Salemi and Aziz (2011) have assessed the BOW method and three levels of N-gram (3, 4, and 5) for extraction, whereby the outcomes obtained via Naïve Bayes classifier have revealed the BOW to outperform all of the N-gram levels tested. Similarly, Mustafa (2005) has positioned a technique for Arabic text searching in Arabic IR by implementing two methods for information extraction, namely contiguous N-grams and hybrid N-grams; the hybrid N-grams are better than contiguous N-grams and both methods extracted the nouns. Furthermore, Al-Shammari and Lin (2008) have stemmed nouns and verbs by using Educated Text Stemmer (ETS) as the stemmer for the roots of Arabic words. The algorithm is aimed towards noun and verb selection from Arabic documents according to prepositions, as well as certain rules regarding other linguistic elements, such as the definite article “the.”

In their previous work on the second stage of BOC (Al-Smadi et al., 2019; Cambria & White, 2014; Menai, 2014), their models were employed to extract the semantic relations between Arabic words based on the stage of BOW. Al-Smadi et al. (2019) employed morphological, syntactic, and semantic processes as Arabic terms extraction and in their result they mentioned the worst result using semantic extraction. Another problem in terms extraction is that additional non-related words are extracted (Al-Smadi et

al., 2019; Al-Zoghby et al., 2018; Menai, 2014); these extra words are termed as highly dimensional features and there is a need to reduce the number of features.

In contrast, using syntactic methods to extract terms can hide the core meaning that is the main meaning of the sentence. This leads to overlapping meanings among different terms such as “bank,” which has ten meanings: for example, a financial institution or sloped land. This overlap in meaning is the main problem of syntactic extraction. Semantic approach can overcome this weakness of syntactic extraction (Al-Zoghby et al., 2018; Elayeb, 2019; Salloum et al., 2018). El Mahdaouy et al. (2018) proposed a word embedding semantic extraction for IR that outperforms BOW. Benabdallah et al. (2017) proposed a method for ontology relation and complex term extraction. In their research a complex term consists of four words. The method was compared with BOW and their proposed method showed an improvement in recall and precision. However, the overall performance of the IR showed that BOW outperformed their method. Table 1 summarizes previous work on terms extraction.

To improve Arabic clustering performance with a reduced number of features and extracting a subset of the disambiguated terms with their relations with high accuracy is highly desirable. Hence, the next sections discuss the extraction of the core features (feature selection) and the clustering problem.

3.2. Unsupervised Feature Selection

Feature selection is a way to alleviate the curse of dimensionality. It reduces dimensionality by eliminating unnecessary and redundant features from the problem, which in turn improves the learning performance. Feature

Table 1. Previous works on extraction

Author	Dataset used	Baseline extraction	Outperform extraction	Domain
Al-Shammari & Lin (2008)	Dataset manual	Khoja and Larkey stemmers	Noun with verbs for stemmer	Arabic clustering
Al-Salemi & Aziz (2011)	TREC-2002	N-gram 3, 4, and 5 level	BOW	Arabic classifiers
Helwe & Elbassuoni (2019)	Arabic Wikipedia corpus	Name entity and their system	Name entity and their our system are equivalent	Arabic classification
El Mahdaouy et al. (2018)	Arabic TREC collection	BOW	Word embedding similarities	Information retrieval
Benabdallah et al. (2017)	Dataset manual	Ontology ‘synonym, antonym, hypernym’ and complex extraction	BOW	Marker learning algorithm

BOW, Bag of Words.

subset selection is a common problem in text clustering (Alweshah et al., 2021; Mafarja & Mirjalili, 2017) because of the high dimensionality of text in documents. Therefore, feature selection is necessary to reduce text dimensionality and to select a reasonable number of high quality features that affect performance. The development of new approaches to handle feature selection and the curse of dimensionality is still an active area of research, particularly for text clustering. The purpose of feature selection includes performance improvement such as accuracy, data visualization and simplification for model selection, and dimensionality reduction to remove noise and irrelevant features (Mafarja & Mirjalili, 2017).

The selection of the features and distribution of the data have a high impact on the performance of clustering algorithms (Abualigah et al., 2016) and text clustering processes such as extraction (Harrag, 2014). The latter tends to obtain local minima rather than the global minimum. The obtained results are often very good, especially when the initial features are fairly far apart. This is because the algorithm can usually distinguish the main category or class in a given data. Moreover, a clustering algorithm's main process and the quality of extraction methods are both affected by the initial feature selection. Thus, the initial features can enhance the quality of the results (Mafarja & Mirjalili, 2017). For instance, failure of the clustering algorithm to recognize the features of the main category in certain data sets is possible if the features are close or similar. This failure can also occur particularly if the feature selection algorithm is left uncontrolled such as in the filter method (Zhang et al., 2019). The feature selection is split into two parts, filter and wrapper. The filter method ignores feature dependencies (Abualigah et al.,

2016; Ahmad et al., 2021; Mafarja & Mirjalili, 2017; Zhang et al., 2019). The wrapper method using optimization as feature selection can introduce a good initial feature selection and leads to a better performance when refining the features and finding the optimal feature selection (Zhang et al., 2019). Table 2 shows the summary of previous work on unsupervised feature selection.

However, only a few studies use optimization as unsupervised feature selection. Tabakhi et al. (2014) employed ant colony optimization as an unsupervised feature selection method for classifiers and used the absolute value (Bharti & Singh, 2014) as the similarity to optimize between features. The efficiency and effectiveness of the ant colony optimization was better than other feature selection methods used for English classifiers. Abualigah et al. (2016) employed a genetic algorithm for feature selection and used the mean absolute difference (MAD) (Bharti & Singh, 2014) as the similarity between features. They compared their proposed method with k-means clustering without feature selection and showed that their proposed method increases the performance of English text clustering. Subsequently, Zhang et al. (2019) have implemented particle swarm optimization unsupervised feature selection combined with filter approach; they compared the proposed approach with filter approach and the PSO. Their results mentioned that the particle swarm optimization with filter approach was better than others and reduced the number of features in English clustering. In contrast, employing the particle swarm optimisation (PSO) for feature selection by Abualigah et al. (2016) employed particle swarm optimization (PSO) for feature selection using MAD (Bharti & Singh, 2014). Their results showed that the proposed feature selection method out-

Table 2. Previous work on unsupervised feature selection

Reference	Dataset used	Baseline algorithm	Outperform algorithm
Tabakhi et al. (2014)	UCI dataset	Laplacian score, Term variance, Random subspace method, Mutual correlation, and Relevance-redundancy feature selection	Ant colony optimization as an unsupervised feature selection
Abualigah et al. (2016)	Text dataset	Harmony Search, an unsupervised feature selection and without feature selection	Genetic algorithm an unsupervised feature selection
Zhang et al. (2019)	UCI datasets and text data	UFS, and ant colony optimization	Particle swarm optimization an unsupervised feature selection
Abualigah et al. (2016)	Text dataset	Genetic algorithm, and Harmony Search	Particle swarm optimization an unsupervised feature selection
Abualigah & Khader (2017)	Text dataset	Genetic algorithm, Harmony Search, particle swarm optimization	Hybrid particle swarm optimization an unsupervised feature selection

UCI, University of California, Irvine; UFS, unsupervised feature extraction.

performs other feature selection methods such as genetic and harmony search algorithms for English clustering. Finally, Abualigah and Khader (2017) employed a hybrid of a PSO algorithm and genetic algorithm for feature selection using MAD (Bharti & Singh, 2014). Their results showed that the proposed feature selection method outperforms PSO and other methods for English clustering. The main gap related to the previous work is the fitness function used, that is, the internal evaluation. The fitness function used by previous work does not have a relation between the internal evaluations metric, MAD, and the external evaluations metric, the F-measure. Theoretically the best score of the internal method should get the best score of the external method. However, in the experiments that we conducted and from the previous work, this does not happen.

The Arabic datasets used and BOW were applied as an extraction method using harmony search feature selection (HSFS) and then harmony search clustering (HSClust) to evaluate the F-measure. This was the first work that employed multi-objective optimization for two problems with the objective functions of MAD and best average distance of documents to the cluster centroid (ADDC), using HSClust with the best parameters from HSFS (Forsati et al., 2013) for Arabic clustering. The number of features are represented by the recommended algorithm in which all the features collection is codified in a vector of length m , where m represents the feature number, as demonstrated in Table 3. Every component of such vector is regarded as

a label where the features are dropped or chosen. Table 3 shows the representation of features for three different solutions with their cost function of MAD. Solution 3 gave the highest score. In this example, twelve features {1, 2, 5, 9, and 11} are chosen while the others {3, 4, 6, 7, 8, 10, 12} are dropped and so forth.

On the other hand, the Harmony Search (HS) as a cluster employs certain representations to code the document. All of the partition clusters, along with a vector of length n that denotes the number of documents, are presented in Table 4. For this vector, every element acts as the label that a single document belongs to. For instance, if the total number of clusters is represented by K , then each solution vector's element gives an integer value that falls in the range of $[K]=\{1, \dots, K\}$. In this example the cost function of ADDC was used. Four documents {2, 3, 8, and 10} originate from the cluster that has been assigned Label 1. The cluster assigned as Label 3 includes three documents {1, 7, 12} and the same goes for the other groups. The best score for ADDC cost function is the lowest value. The Pseudo code of harmony search as a feature selection is shown in Fig. 3 and the Pseudo code of harmony search as cluster is shown in Fig. 4.

The Arabic dataset was taken from (Al-Salemi & Aziz, 2011); more details about the dataset are in Section 4.1. Fig. 5 shows the results of 1,000 iterations for HSFS: each iteration uses new feature selections and uses HSClust to evaluate the F-measure. Based on our experiment, the internal MAD cost function is used as a metric that is the

Table 3. Representation of features

Solutions	Features (m)												MAD cost function
	1	2	3	4	5	6	7	8	9	10	11	12	
Solution 1	1/S	1/S	0/N	0/N	1/S	0/N	0/N	0/N	1/S	0/N	1/S	0/N	0.59
Solution 2	0/N	0/N	1/S	1/S	0/N	1/S	1/S	1/S	0/N	1/S	1/S	0/N	0.55
Solution 3	1/S	1/S	0/N	1/S	0/N	0/N	1/S	0/N	1/S	1/S	1/S	0/N	0.66

MAD, mean absolute difference; 0, non-selected; 1, selected.

Table 4. Representation of clusters as groups from 1 to 5

Solutions	Documents (n)												ADDC cost function
	1	2	3	4	5	6	7	8	9	10	11	12	
Solution 1	3	1	1	4	5	2	3	1	4	1	4	3	0.11
Solution 2	2	3	4	5	4	5	2	3	3	2	5	1	0.33
Solution 3	1	3	2	3	3	4	1	2	1	4	3	2	0.29

ADDC, average distance of documents to the cluster centroid.

best one, and has the highest value. However, it does not represent the highest F-measure value. For instance, at iteration 1 the MAD metric in HSFS was around 0.102 and the value of F-measure is around 0.84. In contrast, at iteration 560, the MAD metric in HSFS was around 0.167, which is better than at iteration 1, but the F-measure value was 0.56, which is worse. The same experiment was conducted for ADDC as shown in Fig. 6. The internal ADDC

metric that is the best one and has the lowest value does not represent the highest F-measure value.

3.3. Clustering Algorithms

The clustering process (Mehra et al., 2020) is comprised of grouping objects based on similarity. Such a method includes two essential kinds of clustering, namely, partitioning and hierarchical. The approaches of hierarchical clustering are believed to be better; nevertheless, at the start of the clustering process, they cannot usually recognize documents which are expected to be misclassified (Yahya, 2018). In addition, the complexity of time of hierarchical approaches is quadratic in terms of the number of objects of data (Wu et al., 2018). The approaches of partitioning clustering are better due to their fairly low computational complexity that enables them to deal with quite big datasets (Wu et al., 2018). The k-means approach has a crucial role in partitioning clustering (Bsoul et al., 2016), and it also has been utilized in partition-based clustering with linear time complexity (Wu et al., 2018). Furthermore, Hartigan (1981) stated that the essential aim of the k-means algorithm is that the documents mean assigned to such a cluster is used to symbolize each of the k clusters. Such a mean is named the cluster centroid. However, the k-means algorithm does not have sensitivity to the initialization and a priori clusters are needed. Moreover, the primary centroids play an essential role in the performance of clustering and might lead the algorithm to be stuck in a locally optimum solution (Selim & Ismail, 1984).

A number of researchers have conducted empirical research on the algorithms of clustering on different data sets while others have carried out research to identify the number of clusters or best clusters. One instance of iden-

```

Objective function  $f(x_i) = i=1$  to  $N$ 
Define HS parameters:  $HMS, HMCR, PAR$  and  $BW$ 
Generate initial harmonics (for  $i=1$  to  $HMS$ ) see table 3
Evaluate  $f(x_i)$  Using MAD unsupervised feature selection
While (until terminating condition)
  Create a new harmony:  $x_i^{new} i=1$  to  $N$ 
  If  $(U(0,1) > HMCR)$ ,
     $x_i^{new} = x_i^{old}$ , where  $x_j^{old}$  is a random from  $\{1, \dots, HMS\}$ 
  Else if  $(U(0,1) \leq PAR)$ ,
     $x_i^{new} = x_L(i) + U(0,1) \times [x_U(i) - x_L(i)]$ 
  Else
     $x_i^{new} = x_j^{old} + BW [(2 \times U(0,1)) - 1]$ , where  $x_j^{old}$  is a random from  $\{1, \dots, HMS\}$ 
  End if
  Evaluate  $f(x_i^{new})$  Using MAD unsupervised feature selection
  Accept the new harmonics (solution) if better
End while
  Fine the current best estimates
    
```

Fig. 3. Pseudo code of harmony search as a feature selection.

```

Objective function  $f(x_i) = i=1$  to  $N$ 
Define HS parameters:  $HMS, HMCR, PAR$  and  $BW$ 
Generate initial harmonics (for  $i=1$  to  $HMS$ ) see fig. 3
Evaluate  $f(x_i)$  Using ADDC Clustering
While (until terminating condition)
  Create a new harmony:  $x_i^{new} i=1$  to  $N$ 
  If  $(U(0,1) > HMCR)$ ,
     $x_i^{new} = x_i^{old}$ , where  $x_j^{old}$  is a random from  $\{1, \dots, HMS\}$ 
  Else if  $(U(0,1) \leq PAR)$ ,
     $x_i^{new} = x_L(i) + U(0,1) \times [x_U(i) - x_L(i)]$ 
  Else
     $x_i^{new} = x_j^{old} + BW [(2 \times U(0,1)) - 1]$ , where  $x_j^{old}$  is a random from  $\{1, \dots, HMS\}$ 
  End if
  Evaluate  $f(x_i^{new})$  Using ADDC Clustering
  Accept the new harmonics (solution) if better
End while
  Fine the current best estimates
    
```

Fig. 4. Pseudo code of harmony search as a clustering.

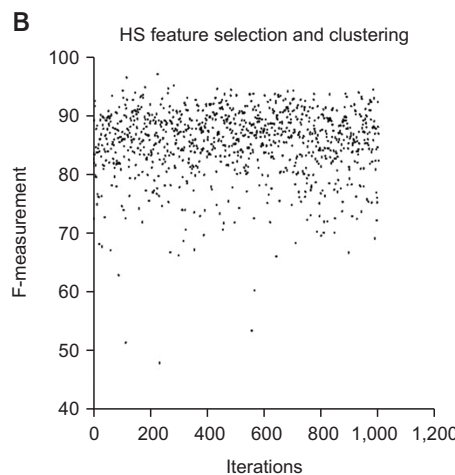
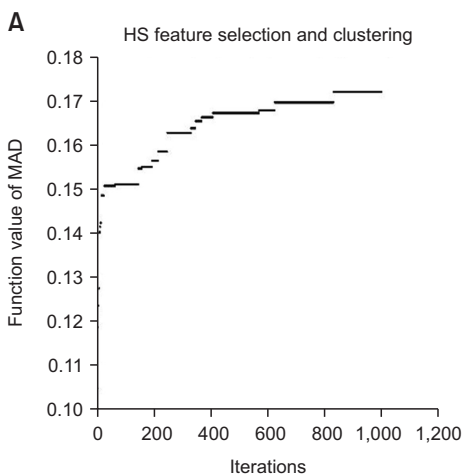


Fig. 5. Comparison of (A) MAD and (B) the F-measure using harmony search for feature selection and clustering. MAD, mean absolute difference.

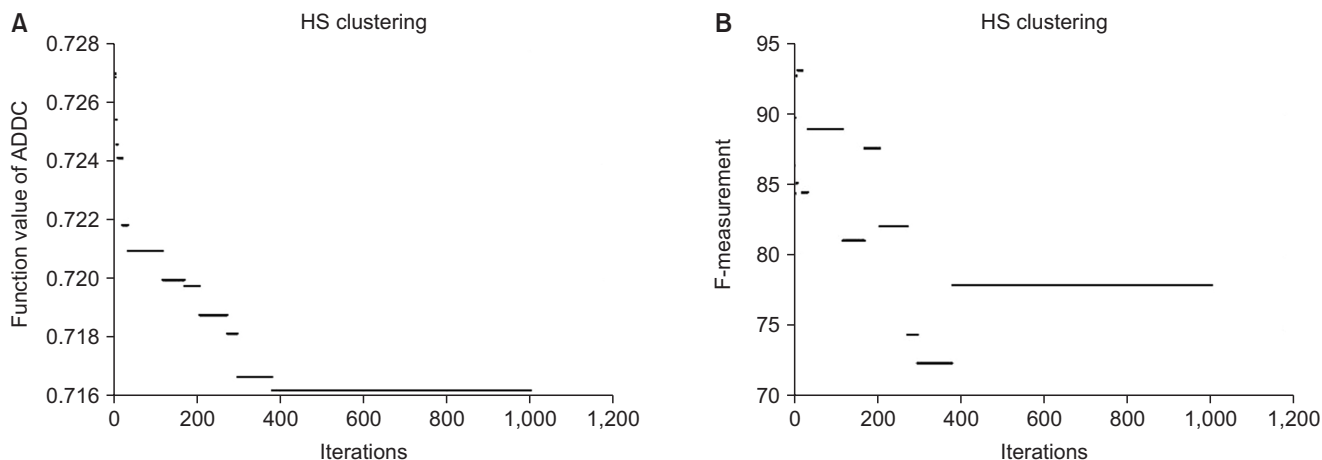


Fig. 6. Comparison of (A) ADDC and (B) the F-measure using harmony search for clustering. ADDC, average distance of documents to the cluster centroid.

tifying the best cluster is Dai et al. (2010a), which streamlined agglomerative hierarchical clustering by considering the significance of a document title. When the term existed in the title, it was assigned as a higher weight. The findings of their studies revealed that the proposed approach was efficient in clustering documents related to financial news. Nevertheless, some studies have investigated clustering themes. For instance, Bação et al. (2005) utilized self-organizing maps (SOMs) in order to cluster images “IRES,” for clustering data (as opposed to the clustering of text), and compared their approach with k-means. It was revealed that the SOM performance is better than that of k-means.

Other studies like Shahrivari and Jalili (2016) have made a comparison among complete single pass cluster, k-means, and others cluster algorithms in a data set. The findings of these studies revealed that k-means is more efficient than the other algorithms. Besides this, Bouras and Tsogkas (2010) utilized clustering approaches such as maximum, single, and centroid linkage hierarchical clustering, k-medians, regular k-means, and k-means++. The results of their study revealed that using k-means does not only produce the best results in terms of the internal metric of the clustering index function, but also provides better results on the experimentation of real users.

Other studies have also made a comparison among single-pass algorithms, k-means, and other algorithms for news topics clustering. For example, Jo (2009) found that k-means performed better than single-pass clustering. Besides this, Dai et al. (2010a) examined hybrid algorithms for clustering and presented a two-layer text clustering method which can detect the themes of retrospective news

employing affinity propagation (AP) clustering. Furthermore, initial layer text clustering studies utilized AP clusters so as to produce the number of groups. Subsequently, they used common agglomerative hierarchical clustering to produce the final themes of news. Finally, such studies adopted classic k-means and usual agglomerative hierarchical clustering (AHC) as comparative approaches. The results showed that the proposed approach obtained the highest precision, followed by AP clustering, k-means clustering, and AHC, respectively. In terms of recall, it was revealed that the proposed approach as well as k-means achieved the highest results followed by the AHC as well as AP clustering approaches.

Velmurugan and Santhanan (2011) also made a comparison among three clustering algorithms on a geographic map data set. The findings of their study revealed that k-means was better with small data sets; k-medoids was better with huge data sets, and fuzzy c-means obtained qualitative results, which fell between those of k-means as well as k-medoids. Dueck and Frey (2007) also proposed AP clustering that could automatically produce many clusters. In fact, AP performs better than k-means in terms of precision, based on the findings of Dai et al. (2010a). However, an assessment of F-measure showed that k-means performs better than AP due to the fact that it has better recall. In addition, Qasim et al. (2013) made a comparison among four clusters utilizing 65 documents as a dataset. The findings of this study revealed that the best clustering approach is AP and then spectral, hierarchical, and ordinary k-means, respectively. Table 5 provides a summary of the studies on clustering.

Furthermore, Silhouette Width (SW) (Alghamdi & Se-

Table 5. Previous work on cluster algorithms

Reference	Dataset used	Baseline algorithm	Outperform algorithm
Dai et al. (2010a)	Financial news	AHC	Enhance AHC
Baço et al. (2005)	IRIS	k-means	SOM
Shahrivari & Jalili (2016)	KDD	Scalable, complete k-means	k-means
Bouras & Tsogkas (2010)	Web	Single, maximum, centroid AHC, and k-medians, k-means++	k-means
Dueck & Frey (2007)	News	k-means, AHC, AP	APAHC
Velmurugan & Santhanam (2011)	Geographic map	Fuzzy <i>c</i> -means	k-means, k-medoids
Qasim et al. (2013)	65 documents news	Spectral, Hierarchical and K-means	AP

IRIS, flower dataset; KDD, knowledge discovery and data mining; AHC, agglomerative hierarchical clustering; AP, affinity propagation; SOM, self-organizing map; APAHC, affinity propagation agglomerative hierarchical clustering.

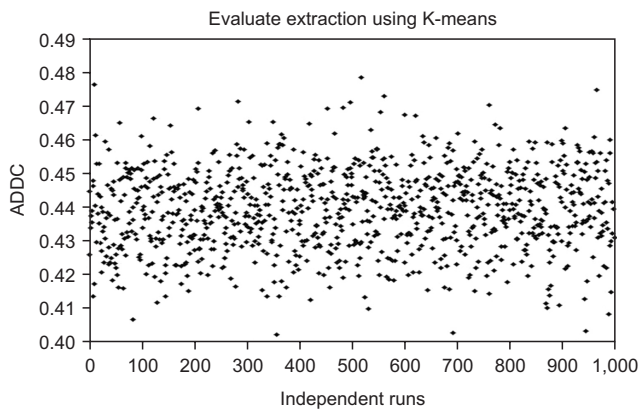


Fig. 7. BOW extraction using k-means for 1,000 independent runs. BOW, Bag of Words; ADDC, average distance of documents to the cluster centroid.

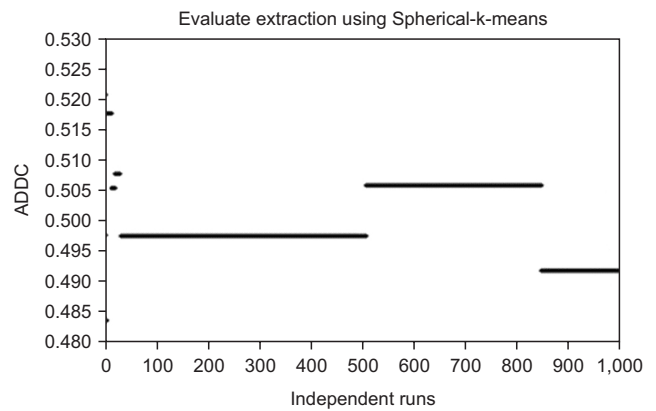


Fig. 8. BOW extraction using spherical k-means for 1,000 independent runs. BOW, Bag of Words; ADDC, average distance of documents to the cluster centroid.

amat, 2019; Campello & Hruschka, 2006) can be used to detect the numbers of clusters. A higher value of SW justifies better discrimination among clusters. Nevertheless, the biggest values of SW justify the best clustering (number of cluster). Numbers ranging between 2 and 10 are used to obtain the number of optimal cluster. When two clusters are recognized, the best SW value reached is 0.1710, but it is not helpful to discover the trend of two categories of crime because the number of clusters is insufficient to be analyzed. Thus, they utilized four clusters instead of two clusters.

Qasim et al. (2013) used AP clustering method to generate an optimal number of clustering and grouped 65 documents as datasets. However, they did not compare SW and Bisecting k-means with other algorithms, and they did not used big datasets to show the performance of AP on bigger datasets than 65 documents. We did three

experiments to address the problem related to three clustering algorithms, namely: Spherical k-means, AP, and k-means cluster. The results of experiments for Arabic clustering are shown in Figs. 7, 8, and 9 and it illustrated the problem of local optima. The results showed that 1,000 independent runs can produce either good or bad performance. The performance of the k-means algorithm is mainly affected by the specified clusters number as well as the random selection of primary cluster centers as revealed in the previous experiments. The means of Spherical k-means were worse than AP and k-means while AP outperform the Spherical k-means and k-mean as illustrated in Table 6. Nevertheless, it does not work on huge data as noticed in Fig. 10. Thus, the present study focuses on dealing with the latter issue by proposing efficient algorithms so as to generate results that are less reliant on the selected primary cluster centers, and thus are more

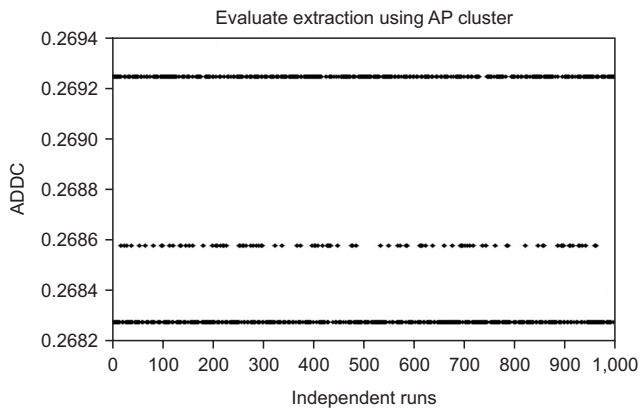


Fig. 9. BOW extraction using affinity propagation for 1,000 independent runs. BOW, Bag of Words; ADDC, average distance of documents to the cluster centroid.

Table 6. Experiment to compare between five cluster algorithms using internal evaluation

Cluster algorithm	Best case	Worst case	Different
HSCLust	0.7815	0.783	0.002
HKM	0.4	0.4	0
K-means	0.401	0.478	0.077
SPK-means	0.484	0.522	0.038
AP	0.26825	0.26925	0.001

HSCLust, harmony search clustering; HKM, hadoop K-means; SPK-means, spherical K-means; AP, affinity propagation.

stabilized. Most of the existing work utilized Harmony Search optimization as clustering. Such studies are mainly conducted for English. However, similar studies have not been conducted for Arabic.

The choice of the primary cluster centers as well as the data distribution has a great effect on k-means and other algorithms. Therefore, it could get stuck in local search rather than global search. The achieved results are usually very good in most cases, particularly when the centroids of initial centers cluster are sufficiently selected far apart, where it would be usually possible to differentiate the major clusters in a certain data set. Moreover, the essential processes of k-means apart from the final partition of the dataset quality are both influenced by the cluster centroids initialization. Therefore, the first points have a great role in the quality of results. For example, the failure of the k-means algorithm to recognize the main clusters features in a certain data set is possible if they are similar or close and particularly if the k-means algorithm is left unsupervised. In addition, to be less reliant on initialization of a certain

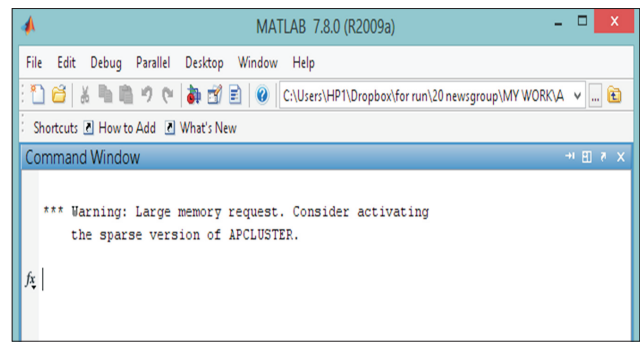


Fig. 10. The main drawback related to affinity propagation.

dataset, and to improve the performance of k-means algorithm, it is important to incorporate the k-means algorithm with an optimization. This will allow better initial clustering centroids, improve the number of the clustering centroids refinement, and recognize the centers of optimal clustering (Gbadoubissa et al., 2020; Mehra et al., 2020).

Meta-heuristics optimization, is widely known to be effective. There are two types of popular meta-heuristics optimization, namely, single-solution based meta-heuristics, also known as trajectory methods, and population-solution based meta-heuristics. Single-solution based meta-heuristics start with a single initial solution and navigate the search space as a trajectory. The main trajectory methods include simulated annealing, Tabu search, the Greedy randomized adaptive search (GRASP) method, variable neighborhood search, guided local search, iterated local search, and their variants. In contrast, population-solution based meta-heuristics do not handle a single solution but rather a set of solutions that is a population. The most widespread population-based approaches are related to evolutionary computation and swarm intelligence (SI). Darwin's theory of evolution provided the concepts needed for evolutionary computation algorithms, because it includes recombination and mutation operators, which allow a population of individuals to be modified by exploiting simple analogues of social interaction, rather than purely individual cognitive abilities. The main aim of SI is to create computational intelligence. The use of nature-inspired meta-heuristic algorithms has been widely adopted in several domains, including computer science (Kushwaha & Pant, 2021), data mining (Grislin-Le Strugeon et al., 2021), texture (Paci et al., 2013), agriculture (Cisty, 2010), computer vision (Connolly et al., 2012), forecasting (Kaur & Sood, 2020), medicine and biology (Arle & Carlson, 2021), scheduling (Guo et al., 2009), economy (Zheng et al., 2020), and engineering (Sayed et al., 2018).

Through the literature it was found that existing clustering algorithms are still inadequate for the Quran themes clustering. Mainly, this is due to the Arabic clustering. In text clustering, similarity measures for documents are utilized and many terms, or features, are specified for each document. In contrast, distance measures between points in a space are used by data mining. The dissimilarity between text mining and data mining has also been mentioned by Hearst (1999). Note that data mining uses non-textual data as opposed to text mining, which utilizes textual data (Guo et al., 2009). However, a few works involve optimization as clustering. Mahdavi and Abolhassani (2009) employed harmony search as an English clustering technique and merged it with *k*-means on diverse datasets. They found the best ADDC measure to be harmony *k*-means, when compared with harmony search clustering, *k*-means, genetic *k*-means, the Mises-Fisher generative model-based algorithm, and PSO clustering. Forsati et al. (2013) generated three versions of harmony search with *k*-means and evaluated 13 scenarios to choose the best parameters related to harmony search for English clustering. The best proposed method was related to the combination of *k*-means with harmony search as cluster. Other works used other type of meta-heuristics. Cagnina et al. (2014) employed Practical Swarm Optimization as English clustering method on short text datasets and compared it with *k*-majorClust, CLUDIPSO, *k*-means, and CHAMELEON as a local search clustering. The best algorithm was an improved version of PSO.

In general, existing methods are summarized in Table 1 for the extraction of Arabic clustering and Table 6 for English text clustering, respectively. The limitations are as follows:

- (1) Syntactic extraction is not adequate for extracting information from Quran themes. Semantic meanings are useful for Quran themes. As an example: Extracting nouns and verbs with their semantic meanings, where events have nouns to describe locations, names, themes, dates, and events also have verbs to describe themes, types, and reasons while simultaneously avoiding the extraction of unimportant features (Atwan et al., 2014).
- (2) *k*-means can be the best clustering algorithm (Jain, 2010). However, in the case of Quran themes, it is not sufficient. *k*-means with a hybrid method with harmony search or spherical *k*-means could be considered. Furthermore, other researchers have provided a solution by integrating with other al-

gorithms such as SW (Alghamdi & Selamat, 2019; Campello & Hruschka, 2006) to determine the number of clusters. However, their findings for SW (Alghamdi & Selamat, 2019; Campello & Hruschka, 2006) and integrated algorithms (Dai et al., 2010b) showed that their approach can be further improved in determining the correct number of clusters.

- (3) Internal evaluations of ADDC for clustering and internal evaluation of MAD for unsupervised feature selection are acceptable to measure optimal centroids and optimal features. However, our experiments in Sections 3.2 and 3.3 showed that the F-measure of this internal evaluation needs to be increased for Arabic text clustering. This can be seen in the experiments shown in Fig. 5 and Fig. 6.
- (4) All the clustering-based works solved the challenges of clustering documents by enhancing part of the process, where the output of each stage affects the accuracy of the next stage (Atwan et al., 2014). Section 4.2 shows the disadvantages of *k*-means clustering where it can affect the evaluation of the extraction methods. In our study this is discovered in Arabic clustering. Therefore, there is a need to evaluate the extraction methods using optimization algorithms that are less dependent on initial centroids. However, existing works do not used optimization-based clustering for Arabic text clustering to evaluate Arabic text extraction. The focus of our research is on the Arabic optimization clustering and optimization as feature selection for Arabic clustering. This is similar to work that has been done in English optimization clustering and optimization as a feature selection for English clustering.
- (5) In addition, previous approaches do not employ multi-objective optimization as feature selection and as clustering using the same algorithm. In other words, it is necessary to employ an optimization method to simultaneously solve two problems: feature selection and clustering.

As stated before, there are three processes involved in Arabic clustering, from the low-level process of clustering up to the high-level process of extracting terms from the Quran themes. As demonstrated in our experiments in Section 4.2, we managed to identify the weakness of extracting information from the Quran themes when addressing the problem of the *k*-means clustering algorithm. In addition, it is very important to extract words such as

nouns and verbs with their semantic meaning related to the themes within the Quran, whereby themes or types in the Quran have nouns to describe information such as names, themes, dates, and locations, while verbs can be used to describe information such as themes, types, and reasons.

Based on the limitations discussed, we can conclude that existing methods for detecting and identifying Arabic text clusters have weaknesses, either with respect to the method of extraction, unsupervised feature selection, and clustering algorithm, or the method of internal evaluation. These are such as ADDC for clustering and MAD for unsupervised feature selection. The problems are related to clustering algorithms and the need to verify the effectiveness of a clustering algorithm by evaluating the extraction process. Therefore, results of previous methods should often be suboptimal (Aouf et al., 2008) and the problem is related to extracting the most important terms for Quran themes (El Mahdaouy et al., 2018). In addition, experiments on Arabic clustering need to be conducted to show the effect of optimization methods such as harmony search on extraction method performance for both Arabic as well as English.

In this study, all the weaknesses mentioned are shown through experiments in the next section; furthermore, we propose solutions to assist the detection and identification of the groups of themes of the Quran using Arabic text optimization for feature selection and clustering.

4. PROPOSED QURAN THEME CLUSTERING APPROACH

This section consists of three subsections, each containing a future direction to be employed for Quran theme clustering. The first section suggests the establishment of new real-text Dataset. The dataset for this study were taken from Al-Salemi and Aziz (2011). The second section recommends extracting the most important terms, as discussed in Sections 2 and 3.1, with respect to the seman-

Table 7. Details of dataset for cluster and classifier domains taken from Al-Salemi and Aziz (2011)

Categories	# DOCUMENTS	# features
Art	420	2,951
Economics	420	3,859
Politics	420	3,172
Sport	420	3,482

tic meaning of nouns and verbs. Section 4.2 describes the selection of the most important features using the black hole (BH) method. In addition, we use the BH method to cluster the data to assign the optimal initial centroids for Quran themes. Using this concept, we propose multi-objective optimization for text clustering, especially for Arabic text clustering.

4.1. Data Collection and Performance Measure

In the conducted experiments, unedited, modern, and unmarked Arabic text was utilized, consisting of a sample of almost 1,680 documents gathered from a number of Arabic online resources. The initial dataset is comprised of four categories, namely, economics, art, sport, and politics articles. Each includes documents collected from Al-Salemi and Aziz (2011). Table 7 provides a summary of the dataset along with features number of every category, as well as the total of 13,464 words. However, the second dataset includes a set specifically designed to assess the extraction of Arabic text for three domains, namely, Arabic classifiers, Arabic clustering, and Arabic IR created as part of TREC 2001. The set has 383,872 Arabic documents, mostly newswire dispatches issued by Agence France Press between the years 1994 and 2000. Ground truth and standard TREC queries have been created for such collection: 25 queries were considered as part of TREC 2001, and the collection of queries has matching relevance judgments produced utilizing the technique of pooling. Based on this, part of TREC 2001 is defined for classifiers and clustering as revealed in Table 8, which include ten classes along with the number of words (a total of 19,508 words)

Table 8. Description of document dataset categories taken from TREC 2001

Classes	# documents	# features
1	383	1,457
2	315	2,013
3	246	1,277
4	222	2,607
5	174	1,402
6	179	1,683
7	393	2,733
8	321	2,169
9	242	1,086
10	556	3,081

as well as the number of related documents.

4.2. Proposed Term Extraction Method

As mentioned in Section 3.1 regarding the weakness of extraction methods, we suggest using Word-Net for noun and verb extraction to extract information that provides the most important features for describing specific Quran themes and their semantic meaning. This can be done by identifying whether a term could be a verb or noun by examination of the stemmed feature that exists in the Arabic Word-Net verb or noun database, as shown in Fig. 11. Quran themes could be identified and detected by looking at nouns as well as verbs along with their semantic meaning. In addition, the redundant terms extracted as nouns and verbs would be removed. In contrast, the problems related to the clusters used to evaluate the extraction methods, as mentioned before, lead us to verify this new problem experimentally.

As noted before, Table 1 shows a summary of the research for extractions. The main gap in this process is the evaluation, which uses clustering search that depends on the initial cluster centroids of each cluster's group. Previous researchers have used independent runs of Arabic text clustering and then taken the average; however, that is not sufficient. In fact, they did not consider the effect of clustering algorithms on extraction to make the comparisons impartial and fair to prove this. BOW has been applied as extraction method using spherical k-means, AP, and k-means. To evaluate the four clusters, we used ADDC as the evaluation metric. Figs. 7, 8, and 9 show the results of 1,000 runs for spherical k-means, AP and k-means, and clustering, which have different results, either positive or

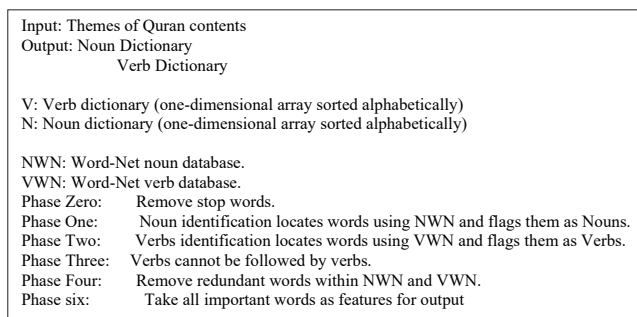


Fig. 11. Proposed method to extract nouns and verbs.

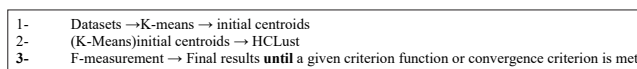


Fig. 12. One step k-means with harmony search.

negative. Good results are close to zero and this happened to all of them. However, the results of k-means showed that k-means are very dependent on initial cluster centroids. The dependency is less with spherical k-means and followed by AP. It can be seen that the worst result of k-means was close to the best result of spherical k-means. Moreover, AP's results were better than the others, but on pairs of dataset there were "two classes of the dataset." However, on big dataset, this is not possible. The AP method needs an impractically large memory to normalize the dataset, and this experiment is shown in Fig. 10.

4.3. Black Hole Optimization as Unsupervised Feature Selection and Clustering Method

In the last two decades, many meta-heuristic methods have been used for data mining, including simulated annealing (Güngör & Ünler, 2007), Tabu search (Kharroush-eh et al., 2011), genetic algorithms (Liu et al., 2012), ant colony optimization (Niknam & Amiri, 2010), the neural gas algorithm (Qin & Suganthan, 2004), honey bee mating optimization (Fathian et al., 2007), differential evolution (Das et al., 2007), PSO (Izakian & Abraham, 2011), artificial bee colony optimization (Karaboga & Ozturk, 2011), gravitational search (Hatamlou et al., 2012), binary search (Hatamlou, 2012), firefly optimization (Senthilnath et al., 2011), big bang-big crunch (Hatamlou et al., 2011), harmony search k-means (Forsati et al., 2013), and BH optimization (Hatamlou, 2013). However, in this study we conducted an experiment to show the weaknesses of harmony search with k-means proposed for English clustering by Forsati et al. (2013). We used the best parameter settings obtained by Forsati et al. (2013), and the pseudo code of harmony search as a cluster with k-means shown

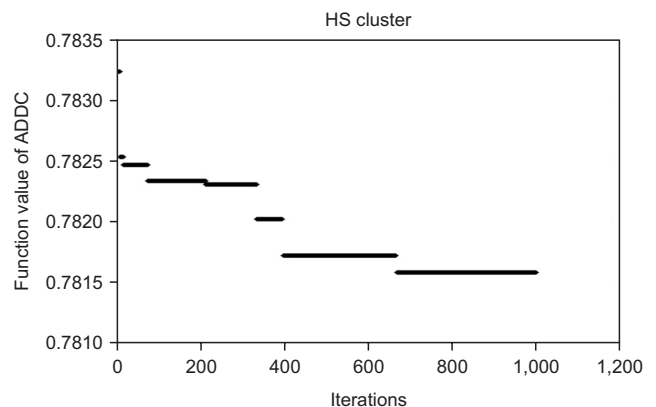


Fig. 13. Harmony search for 1,000 iterations using BOW extraction. BOW, Bag of Words; ADDC, average distance of documents to the cluster centroid.

in Figs. 3, 4, and 12. Comparisons of the results of harmony search clustering (Fig. 13) and one-step k-means (harmony k-means) are shown in Fig. 14, where the combined method was better than harmony search clustering. This can be seen in both Fig. 13 and Fig. 14. However, results shown in Figs. 7, 8, 9, and 13 showed that k-means clustering at times can produce better results than harmony k-means. Therefore, another optimization method can be proposed to overcome this weakness.

To address the weakness of harmony search, we suggest employing BH optimization as a cluster (Hatamlou, 2013) for Quran themes clustering. Owing to its few common features with other population-based methods, the BH optimization algorithm is in fact a population-based method. A population of candidate solutions to a given problem is generated and placed randomly within the search space, as in other population-based algorithms that utilize certain mechanisms to gradually improve the population to obtain the optimal solution. In genetic algorithms, for instance, mutation and crossover operations help to achieve gradual improvements. By moving the candidate solutions around in the search space, this progress can be achieved in PSO using the best locations found so far, which are updated when the candidates reach better locations. However, by shifting all the candidates towards the best candidate at each iteration, namely, as in BH, and considering newly generated candidates in the search space instead of those candidates included within the current set of the BH, the population in the BH algorithm can grow. First, to solve the BH problem, researchers tend to utilize benchmark functions (Zhang et al., 2008). However, a new mechanism, known as the BH, has been

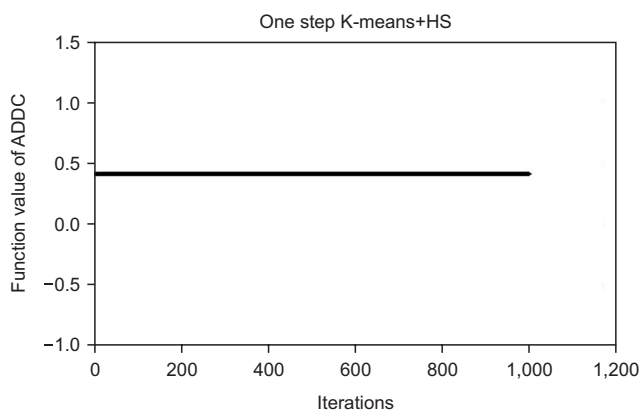


Fig. 14. One step k-means+harmony search for 1,000 iterations using BOW extraction. BOW, Bag of Words; ADDC, average distance of documents to the cluster centroid.

introduced into PSO because of the method suggested in Zhang et al. (2008). A new particle is created randomly near the best particle at each iteration in this method. After this, the algorithm updates the locations of the particles either using PSO or a new mechanism that is mainly based on two randomly created numbers. To summarize, PSO has paved the way for this method. BH is a new particle used to augment the convergence speed of the PSO and limit the convergence to local optima. Moreover, a BH can attract other particles under certain conditions. The theme horizon of the BH and the destruction of stars (candidates) has not been tackled in this approach, but such optimization has been treated in Hatamlou (2013). The best candidate at each iteration in the BH algorithm is considered to be a BH; however, all the remaining candidates are considered to be normal stars. A BH is not randomly created; it is created by one of the real candidates of the population. After this, depending on their current location and a random number, all the candidates are shifted towards the BH. The suggested BH algorithm is beneficial mainly because of its simple structure and easy implementation. Moreover, it is free from parameter tuning issues (Hatamlou, 2013).

The main notion of utilizing the BH for feature selection is basically to generate an area of space features which have a big amount of concentrated mass. Thus, the potential of a nearby object feature escaping its gravitational pull for significant features is reduced. Therefore, it is believed that anything falls into a BH, among them light, which is eternally gone from the universe. The recommended BH algorithm begins with employing a primary population of candidate solutions to an objective function which is estimated for them as well as to an optimization problem. Therefore, the best candidate is selected in each iteration as the BH whereas the others comprise the normal features or normal stars. Subsequently, the process of initialization is complete, and the BH begins with pulling the stars around its feature. If the resemblance between the BH centroid and star (feature) is high, it will be swallowed by the significant features of BH and is gone everlastingly. In this case, a fresh star (candidate solution) is created randomly and put in the search space, followed by a new search. The steps involved are as follows:

- Loop based on the number of BHs that will be assigned.
- For each star “feature,” evaluate the objective function for each BH using MAD.
- Select the best stars “features” that have the best fit-

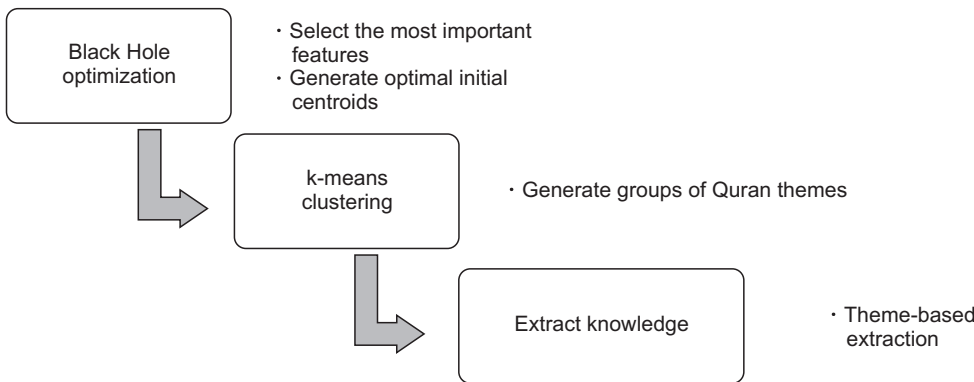


Fig. 15. The proposed method.

- ness value as the BH important features for all BHs.
- Change the location of each star “features” according to $x_i(t+1)=x_i(t)+rand \times (x_{BH}-x_i(t)) \dots i=1, 2, 3, \dots, N$
 - If a star “feature” reaches a location with higher similarity than the BH, exchange their locations.
 - If a star “feature” crosses the theme horizon of the BH, replace it with a new star “feature” in a random location in the search space.
 - When a termination criterion (a maximum number of iterations or a sufficiently good fitness) is met, exit the loop.

In the above equation, $x_i(t)$ and $x_i(t+1)$ are the i -th star of the locations at iterations t and $t+1$, respectively. In addition, x_{BH} is the location of the BH in the search space. Variable *Rand* is a random number in the interval [0, 1], and N is the number of stars (candidate solutions).

BH has been proposed as feature selection and currently we are introducing BH as clustering. The BH as cluster started with a fresh star (candidate solution). The candidate solution is created randomly and put in the search space, followed by a new search. The steps involved are as follows:

- Loop based on the number of BHs that will be assigned.
- For each star “number of group”, evaluate the objective function for each BH using ADDC.
- Select the best stars “number of group” that have the best fitness value as the BH of each documents/versus belong to theme/group for all BHs.
- Change the location of each star “number of group” according to $x_i(t+1)=x_i(t)+rand \times (x_{BH}-x_i(t)) \dots i=1, 2, 3, \dots, \text{Number of groups}$
- If a star “number of group” reaches a location with higher similarity than the BH, exchange their loca-

```

Input: objective function
Output: optimal solution
Initialize a population of black holes with random locations in the search space (Big Bang)
While
  do
    For each black hole,
      Evaluate the objective function using ADDC
      Select the global best black hole that has the best fitness value
      Change the location of each black holes
    End of while
    Repeat
      Choose C = Best black hole           generated from BH optimizations
      Initialize A as zero
      for all  $d_i$  in D do:
        let  $j = \text{argmin}_{k \in \{1,2,\dots,K\}} D(d_i, c_k)$ ;
        assign  $d_i$  to cluster  $j$ , i.e.,  $A[i][j] = 1$ 
      end for
    Update the cluster means as  $c_k = \frac{\sum_{i=1}^m (\sum_{j=1}^k A[i][k] d_i)}{\sum_{i=1}^m (\sum_{j=1}^k A[i][k])}$  for  $k = 1, 2, \dots, K$ 
  until a given criterion function or convergence criterion is met.
  
```

Fig. 16. The proposed Black Hole and k-means clustering method.

- If a star “number of group” crosses the theme horizon of the BH, replace it with a new star “number of group” in a random location in the search space.
- When a termination criterion (a maximum number of iterations or a sufficiently good fitness) is met, exit the loop.

Fig. 15 shows the proposed method using BH as feature selection, which is then followed by employing BH as a cluster for Quran themes clustering. The possibility of a nearby object that is the Quran theme document escaping its gravitational pull (i.e., that of the clusters) is minimized. The proposed BH with the combination of k-means are shown in Fig. 16. The proposed BH as a feature selection reduced the unimportant features and the proposed BH as a cluster will select the optimal initial centroid of each theme.

5. CONCLUSION

The ultimate aim of this paper was to employ Arabic text clustering to cluster Quran themes into clusters.

Hence, this study reviewed the necessary improvements to Arabic text clustering, and suggested possible research directions for improving Arabic text clustering with respect to extraction, feature selection, and clustering. In this review paper, the limitations related to Arabic text clustering were discussed and the limitations of existing works were demonstrated and presented through our experiments. For instance, the k-means algorithm finds the best clusters compared to other algorithms as shown in Figs. 7, 8, 9, and 10. But the k-means weakness is in the initial centroids. Therefore, we proposed to use BH as a cluster to create the optimal Initial centroids. Next, we examined the weakness of term extraction methods and conducted a number of experiments to demonstrate the problems identified in existing approaches. For this problem we suggested using nouns and verbs with their semantic meaning as a new extraction method for Quran theme clustering. This can provide better performance compared to the current approach.

We identified a new problem related to extraction depending on the clustering algorithm used, and for that we suggest the use of BH as feature selection and as clustering to evaluate the proposed extraction. We also noted the limitations of using AP clustering on a big dataset, and our suggestion is to use BH as feature selection. For the dataset we have generated and published real Quran theme dataset and Arabic text dataset, which are freely available online.¹

ACKNOWLEDGMENTS

This paper presents a work that is supported by the Ministry of Higher Education (MOHE), under the Transdisciplinary Research Grant Scheme (TRGS-MOHE) (TRGS/1/2015/USIM/01/1).

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

- Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *The Journal of Supercomputing*, 73(11), 4773-4795. <https://doi.org/10.1007/s11227-017-2046-2>.
- Abualigah, L. M., Khader, A. T., & Al-Betar, M. A. (2016, July 13-14). Unsupervised feature selection technique based on genetic algorithm for improving the text clustering. In A. Altamimi, & F. Almasalha (Eds.), *Proceedings of the 7th International Conference on Computer Science and Information Technology* (pp. 1-6). Institute of Electrical and Electronics Engineers.
- Abualkishik, A. M., Omar, K., & Odiebat, G. A. (2015). QEFISM model and Markov Algorithm for translating Quran reciting rules into Braille code. *Journal of King Saud University - Computer and Information Sciences*, 27(3), 238-247. <https://doi.org/10.1016/j.jksuci.2015.01.001>.
- Ahmad, R., Wazirali, R., Bsoul, Q., Abu-Ain, T., & Abu-Ain, W. (2021). Feature-selection and mutual-clustering approaches to improve DoS detection and maintain WSNs' lifetime. *Sensors (Basel, Switzerland)*, 21(14), 4821. <https://doi.org/10.3390/s21144821>.
- Alghamdi, H. M., & Selamat, A. (2019). Arabic web page clustering: A review. *Journal of King Saud University - Computer and Information Sciences*, 31(1), 1-14. <https://doi.org/10.1016/j.jksuci.2017.06.002>.
- Al-Salemi, B., & Aziz, M. J. A. (2011). Statistical Bayesian learning for automatic Arabic text categorization. *Journal of Computer Science*, 7(1), 39-45. <https://doi.org/10.3844/jcssp.2011.39.45>.
- Al-Shammari, E., & Lin, J. (2008, July 24). A novel Arabic lemmatization algorithm. In D. Lopresti, S. Roy, K. Schulz, & L. V. Subramaniam (Eds.), *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data* (pp. 113-118). Association for Computing Machinery.
- Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., & Qawasmeh, O. (2019). Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management*, 56(2), 308-319. <https://doi.org/10.1016/j.ipm.2018.01.006>.
- Alweshah, M., Alkhalailah, S., Albashish, D., Mafarja, M., Bsoul, Q., & Dorgham, O. (2021). A hybrid mine blast algorithm for feature selection problems. *Soft Computing*, 25(1), 517-534. <https://doi.org/10.1007/s00500-020-05164-4>.
- Al-Zoghby, A. M., Elshawi, A., & Atwan, A. (2018). Semantic relations extraction and ontology learning from Arabic texts—a survey. In K. Shaalan, A. Hassani, & F. Tolba (Eds.), *Intelligent natural language processing: Trends and applications* (pp. 199-225). Springer.
- Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review*, 27(4), 509-523. <https://doi.org/10.1177/0894439309332293>.

¹<https://www.dropbox.com/s/yolg2s3qfj8zn7f/Arabic%20dataset.rar?dl=0>

- Aouf, M., Liyanage, L., & Hansen, S. (2008, June 30-July 2). Review of data mining clustering techniques to analyze data with high dimensionality as applied in gene expression data. In V. C. S. Lee, J. Chen, W.-K. Ng, K.-L. Ong, & T. Y. Tan (Eds.), *Proceedings of the 2008 International Conference on Service Systems and Service Management* (pp. 1-5). Institute of Electrical and Electronics Engineers.
- Arle, J. E., & Carlson, K. W. (2021). Medical diagnosis and treatment is NP-complete. *Journal of Experimental & Theoretical Artificial Intelligence*, 33(2), 297-312. <https://doi.org/10.1080/0952813X.2020.1737581>.
- Atwan, J., Mohd, M., Kanaan, G., & Bsoul, Q. (2014, December 3-5). Impact of stemmer on Arabic text retrieval. In A. Jaafar, N. M. Ali, S. A. M. Noah, A. F. Smeaton, P. Bruza, Z. A. Bakar, N. Jamil, T. Mohd, & T. Sembok (Eds.), *Proceedings of the 10th Asia Information Retrieval Societies Conference* (pp. 314-326). Springer.
- Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, 56(5), 1698-1735. <https://doi.org/10.1016/j.ipm.2019.05.009>.
- Baço, F., Lobo, V., & Painho, M. (2005, May 22-25). Self-organizing maps as substitutes for K-means clustering. In V. S. Sunderam, G. D. van Albada, P. M. A. Sloot, & J. Dongarra (Eds.), *Proceedings of the 5th International Conference on Computational Science* (pp. 476-483). Springer.
- Beirade, F., Azzoune, H., & Zegour, D. E. (2021). Semantic query for Quranic ontology. *Journal of King Saud University - Computer and Information Sciences*, 33(6), 753-760. <https://doi.org/10.1016/j.jksuci.2019.04.005>.
- Benabdallah, A., Abderrahim, M. A., & Abderrahim, M. E. A. (2017). Extraction of terms and semantic relationships from Arabic texts for automatic construction of an ontology. *International Journal of Speech Technology*, 20(2), 289-296. <https://doi.org/10.1007/s10772-017-9405-5>.
- Bharti, K. K., & Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science*, 5(2), 156-169. <https://doi.org/10.1016/j.jocs.2013.11.007>.
- Bouras, C., & Tsogkas, V. (2010, May 9-15). Assigning web news to clusters. In G. O. Bellot, H. Sasaki, M. Ehmann, & C. Dini (Eds.), *Proceedings of the 5th International Conference on Internet and Web Applications and Services* (pp. 1-6). Institute of Electrical and Electronics Engineers.
- Bsoul, Q., Al-Shamari, E., Mohd, M., & Atwan, J. (2014, December 3-5). Distance measures and stemming impact on Arabic document clustering. In A. Jaafar, N. M. Ali, S. A. M. Noah, A. F. Smeaton, P. Bruza, Z. A. Bakar, N. Jamil, T. Mohd, & T. Sembok (Eds.), *Proceedings of the 10th Asia Information Retrieval Societies Conference* (pp. 327-339). Springer.
- Bsoul, Q., Salim, J., & Zakaria, L. Q. (2016). Document clustering approach to detect crime. *World Applied Sciences Journal*, 34(8), 1026-1036. <https://doi.org/10.5829/idosi.wasj.2016.34.8.109>.
- Cagnina, L., Errecalde, M., Ingaramo, D., & Rosso, P. (2014). An efficient particle swarm optimization approach to cluster short texts. *Information Sciences*, 265, 36-49. <https://doi.org/10.1016/j.ins.2013.12.010>.
- Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48-57. <https://doi.org/10.1109/MCI.2014.2307227>.
- Campello, R. J. G. B., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157(21), 2858-2875. <https://doi.org/10.1016/j.fss.2006.07.006>.
- Cisty, M. (2010). Application of the harmony search optimization in irrigation. In Z. W. Geem (Ed.), *Recent advances in harmony search algorithm* (pp. 123-134). Springer.
- Connolly, J.-F., Granger, E., & Sabourin, R. (2012). An adaptive classification system for video-based face recognition. *Information Sciences*, 192, 50-70. <https://doi.org/10.1016/j.ins.2010.02.026>.
- Dai, X., He, Y., & Sun, Y. (2010a, October 23-24). A two-layer text clustering approach for retrospective news event detection. In F. L. Wang, & T. Jin (Eds.), *Proceedings of the Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence* (pp. 364-368). Institute of Electrical and Electronics Engineers.
- Dai, X.-Y., Chen, Q.-C., Wang, X.-L., & Xu, J. (2010b, July 11-14). Online topic detection and tracking of financial news based on hierarchical clustering. *Proceedings of the 2010 International Conference on Machine Learning and Cybernetics* (pp. 3341-3346). Institute of Electrical and Electronics Engineers.
- Das, S., Abraham, A., & Konar, A. (2007). Automatic clustering using an improved differential evolution algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(1), 218-237. <https://doi.org/10.1109/TSMCA.2007.909595>.
- Dueck, D., & Frey, B. J. (2007, October 14-21). Non-metric affinity propagation for unsupervised image categorization. *Proceedings of the 11th International Conference on Computer Vision* (pp. 1-8). Institute of Electrical and Electronics Engineers.
- El Mahdaouy, A., El Alaoui Ouatik, S., & Gaussier, E. (2018). Improving Arabic information retrieval using word embed-

- ding similarities. *International Journal of Speech Technology*, 21(1), 121-136. <https://hal.archives-ouvertes.fr/hal-01706531>.
- Elayeb, B. (2019). Arabic word sense disambiguation: A review. *Artificial Intelligence Review*, 52(4), 2475-2532. <https://doi.org/10.1007/s10462-018-9622-6>.
- Farhan, Y. H., Mohd, M., & Noah, S. A. M. (2020). Survey of automatic query expansion for Arabic text retrieval. *Journal of Information Science Theory and Practice*, 8(4), 67-86. <https://doi.org/10.1633/JISTaP2020.8.4.6>.
- Fathian, M., Amiri, B., & Maroosi, A. (2007). Application of honey-bee mating optimization algorithm on clustering. *Applied Mathematics and Computation*, 190(2), 1502-1513. <https://doi.org/10.1016/j.amc.2007.02.029>.
- Forsati, R., Mahdavi, M., Shamsfard, M., & Reza Meybodi, M. (2013). Efficient stochastic algorithms for document clustering. *Information Sciences*, 220, 269-291. <https://doi.org/10.1016/j.ins.2012.07.025>.
- Gbadoubissa, J. E. Z., Ari, A. A. A., & Gueroui, A. M. (2020). Efficient K-means based clustering scheme for mobile networks cell sites management. *Journal of King Saud University - Computer and Information Sciences*, 32(9), 1063-1070. <https://doi.org/10.1016/j.jksuci.2018.10.015>.
- Grislin-Le Strugeon, E., Marcal de Oliveira, K., Thilliez, M., & Petit, D. (2021). A systematic mapping study on agent mining. *Journal of Experimental & Theoretical Artificial Intelligence*. <https://doi.org/10.1080/0952813X.2020.1864784>.
- Güngör, Z., & Ünler, A. (2007). K-harmonic means data clustering with simulated annealing heuristic. *Applied Mathematics and Computation*, 184(2), 199-209. <https://doi.org/10.1016/j.amc.2006.05.166>.
- Guo, Y. W., Li, W. D., Mileham, A. R., & Owen, G. W. (2009). Applications of particle swarm optimisation in integrated process planning and scheduling. *Robotics and Computer-Integrated Manufacturing*, 25(2), 280-288. <https://doi.org/10.1016/j.rcim.2007.12.002>.
- Harrag, F. (2014). Text mining approach for knowledge extraction in Sahih Al-Bukhari. *Computers in Human Behavior*, 30, 558-566. <https://doi.org/10.1016/j.chb.2013.06.035>.
- Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374), 388-394. <https://doi.org/10.1080/01621459.1981.10477658>.
- Hatamlou, A. (2012). In search of optimal centroids on data clustering using a binary search algorithm. *Pattern Recognition Letters*, 33(13), 1756-1760. <https://doi.org/10.1016/j.patrec.2012.06.008>.
- Hatamlou, A. (2013). Black hole: A new heuristic optimization approach for data clustering. *Information Sciences*, 222, 175-184. <https://doi.org/10.1016/j.ins.2012.08.023>.
- Hatamlou, A., Abdullah, S., & Hatamlou, M. (2011, December 13-15). Data clustering using big bang-big crunch algorithm. In P. Pichappan, H. Ahmadi, & E. Ariwa (Eds.), *Proceedings of the 1st International Conference on Innovative Computing Technology* (pp. 383-388). Springer.
- Hatamlou, A., Abdullah, S., & Nezamabadi-pour, H. (2012). A combined approach for clustering based on K-means and gravitational search algorithms. *Swarm and Evolutionary Computation*, 6, 47-52. <https://doi.org/10.1016/j.swevo.2012.02.003>.
- Hearst, M. A. (1999, June 20-26). Untangling text data mining. In R. Dale, & K. Church (Eds.), *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3-10). Association for Computational Linguistics.
- Helwe, C., & Elbassuoni, S. (2019). Arabic named entity recognition via deep co-learning. *Artificial Intelligence Review*, 52(1), 197-215. <https://doi.org/10.1007/s10462-019-09688-6>.
- Hmeidi, I., Hawashin, B., & El-Qawasmeh, E. (2008). Performance of KNN and SVM classifiers on full word Arabic articles. *Advanced Engineering Informatics*, 22(1), 106-111. <https://doi.org/10.1016/j.aei.2007.12.001>.
- Izakian, H., & Abraham, A. (2011). Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. *Expert Systems with Applications: An International Journal*, 38(3), 1835-1838. <https://doi.org/10.1016/j.eswa.2010.07.112>.
- Jain, A. K. (2010). Data clustering: 50 Years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. <https://doi.org/10.1016/j.patrec.2009.09.011>.
- Jo, T. (2009, July 28-31). Clustering news groups using inverted index based NTSO. *Proceedings of the 1st International Conference on Networked Digital Technologies* (pp. 1-7). Institute of Electrical and Electronics Engineers.
- Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computing*, 11(1), 652-657. <https://doi.org/10.1016/j.asoc.2009.12.025>.
- Kaur, A., & Sood, S. K. (2020). Cloud-Fog based framework for drought prediction and forecasting using artificial neural network and genetic algorithm. *Journal of Experimental & Theoretical Artificial Intelligence*, 32(2), 273-289. <https://doi.org/10.1080/0952813X.2019.1647563>.
- Kharrousheh, A., Abdullah, S., & Nazri, M. Z. A. (2011). A modified Tabu search approach for the clustering problem. *Journal of Applied Sciences*, 11(19), 3447-3453. <https://doi.org/10.3923/jas.2011.3447.3453>.
- Kushwaha, N., & Pant, M. (2021). Fuzzy electromagnetic op-

- timisation clustering algorithm for collaborative filtering. *Journal of Experimental & Theoretical Artificial Intelligence*, 33(4), 601-616. <https://doi.org/10.1080/0952813X.2019.1647557>.
- Liu, R., Jiao, L., Zhang, X., & Li, Y. (2012). Gene transposon based clone selection algorithm for automatic clustering. *Information Sciences: An International Journal*, 204, 1-22. <https://doi.org/10.1016/j.ins.2012.03.021>.
- Mafarja, M. M., & Mirjalili, S. (2017). Hybrid whale optimization algorithm with simulated annealing for feature selection. *Neurocomputing*, 260, 302-312. <https://doi.org/10.1016/j.neucom.2017.04.053>.
- Mahdavi, M., & Abolhassani, H. (2009). Harmony K-means algorithm for document clustering. *Data Mining and Knowledge Discovery*, 18(3), 370-391. <https://doi.org/10.1007/s10618-008-0123-0>.
- Mehra, P. S., Doja, M. N., & Alam, B. (2020). Fuzzy based enhanced cluster head selection (FBECs) for WSN. *Journal of King Saud University - Science*, 32(1), 390-401. <https://doi.org/10.1016/j.jksus.2018.04.031>.
- Menai, M. E. B. (2014). Word sense disambiguation using evolutionary algorithms - application to Arabic language. *Computers in Human Behavior*, 41(C), 92-103. <https://doi.org/10.1016/j.chb.2014.06.021>.
- Mesmia, F. B., Haddar, K., Friburger, N., & Maurel, D. (2018). CasANER: Arabic named entity recognition tool. In K. Shaalan, A. Hassanien, & F. Tolba (Eds.), *Intelligent natural language processing: Trends and applications* (pp. 173-198). Springer.
- Mottaghinia, Z., Feizi-Derakhshi, M.-R., Farzinvas, L., & Salehpour, P. (2021). A review of approaches for topic detection in Twitter. *Journal of Experimental & Theoretical Artificial Intelligence*, 33(5), 747-773. <https://doi.org/10.1080/0952813X.2020.1785019>.
- Mustafa, S. H. (2005). Character contiguity in N-gram-based word matching: The case for Arabic text searching. *Information Processing and Management: An International Journal*, 41(4), 819-827. <https://doi.org/10.1016/j.ipm.2004.02.003>.
- Niknam, T., & Amiri, B. (2010). An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. *Applied Soft Computing*, 10(1), 183-197. <https://doi.org/10.1016/j.asoc.2009.07.001>.
- Paci, M., Nanni, L., & Severi, S. (2013). An ensemble of classifiers based on different texture descriptors for texture classification. *Journal of King Saud University - Science*, 25(3), 235-244. <https://doi.org/10.1016/j.jksus.2012.12.001>.
- Qasim, I., Jeong, J.-W., Heu, J.-U., & Lee, D.-H. (2013). Concept map construction from text documents using affinity propagation. *Journal of Information Science*, 39(6), 719-736. <https://doi.org/10.1177%2F0165551513494645>.
- Qin, A. K., & Suganthan, P. N. (2004). Robust growing neural gas algorithm with application in cluster analysis. *Neural Networks: The Official Journal of the International Neural Network Society*, 17(8-9), 1135-1148. <https://doi.org/10.1016/j.neunet.2004.06.013>.
- Raharjo, S., Wardoyo, R., & Putra, A. E. (2020). Detecting proper nouns in Indonesian-language translation of the Quran using a guided method. *Journal of King Saud University - Computer and Information Sciences*, 32(5), 583-591. <https://doi.org/10.1016/j.jksuci.2018.06.009>.
- Rostam, N. A. P., & Malim, N. H. A. H. (2021). Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting. *Journal of King Saud University - Computer and Information Sciences*, 33(6), 658-667. <https://doi.org/10.1016/j.jksuci.2019.03.007>.
- Safee, M. A. M., Saudi, M. M., Pitchay, S. A., & Ridzuan, F. (2016). A systematic review analysis for Quran verses retrieval. *Journal of Engineering and Applied Sciences*, 11(3), 629-634. <https://doi.org/10.36478/jeasci.2016.629.634>.
- Salloum, S. A., Al Hamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A survey of Arabic text mining. In K. Shaalan, A. E. Hassanien, & F. Tolba (Eds.), *Intelligent natural language processing: Trends and applications* (pp. 417-431). Springer.
- Saloot, M. A., Idris, N., Mahmud, R., Jaafar, S., Thorleuchter, D., & Gani, A. (2016). Hadith data mining and classification: A comparative analysis. *Artificial Intelligence Review*, 46(1), 113-128. <https://doi.org/10.1007/s10462-016-9458-x>.
- Sayed, G. I., Darwish, A., & Hassanien, A. E. (2018). A new chaotic multi-verse optimization algorithm for solving engineering optimization problems. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2), 293-317. <https://doi.org/10.1080/0952813X.2018.1430858>.
- Selim, S. Z., & Ismail, M. A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(1), 81-87. <https://doi.org/10.1109/TPAMI.1984.4767478>.
- Senthilnath, J., Omkar, S. N., & Mani, V. (2011). Clustering using firefly algorithm: Performance study. *Swarm and Evolutionary Computation*, 1(3), 164-171. <https://doi.org/10.1016/j.swevo.2011.06.003>.
- Shahrivari, S., & Jalili, S. (2016). Single-pass and linear-time k-means clustering based on MapReduce. *Information Systems*, 60, 1-12. <https://doi.org/10.1016/j.is.2016.02.007>.
- Tabakhi, S., Moradi, P., & Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony opti-

- mization. *Engineering Applications of Artificial Intelligence*, 32, 112-123. <https://doi.org/10.1016/j.engappai.2014.03.007>.
- Touahri, I., & Mazroui, A. (2021). Studying the effect of characteristic vector alteration on Arabic sentiment classification. *Journal of King Saud University - Computer and Information Sciences*, 33(7), 890-898. <https://doi.org/10.1016/j.jksuci.2019.04.011>.
- Velmurugan, T., & Santhanam, T. (2011). A survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal*, 10(3), 478-484. <https://doi.org/10.3923/itj.2011.478.484>.
- Wu, J., Dong, M., Ota, K., Li, J., & Guan, Z. (2018). Big data analysis-based secure cluster management for optimized control plane in software-defined networks. *IEEE Transactions on Network and Service Management*, 15(1), 27-38. <https://doi.org/10.1109/TNSM.2018.2799000>.
- Yahya, A, A. (2018). Centroid particle swarm optimisation for high-dimensional data classification. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(6), 857-886. <https://doi.org/10.1080/0952813X.2018.1509378>.
- Yauri, A. R., Kadir, R. A., Azman, A., & Murad, M. A. A. (2013). Quranic verse extraction base on concepts using OWL-DL ontology. *Research Journal of Applied Sciences, Engineering and Technology*, 6(23), 4492-4498. <https://doi.org/10.19026/rjaset.6.3457>.
- Zhang, J., Liu, K., Tan, Y., & He, X. (2008, June 7-11). Random black hole particle swarm optimization and its application. *Proceedings of the 2008 International Conference on Neural Networks and Signal Processing* (pp. 359-365). Institute of Electrical and Electronics Engineers.
- Zhang, Y., Li, H.-G., Wang, Q., & Peng, C. (2019). A filter-based bare-bone particle swarm optimization algorithm for unsupervised feature selection. *Applied Intelligence*, 49(8), 2889-2898. <https://doi.org/10.1007/s10489-019-01420-9>.
- Zheng, Z., Li, J., & Han, Y. (2020). An improved invasive weed optimization algorithm for solving dynamic economic dispatch problems with valve-point effects. *Journal of Experimental & Theoretical Artificial Intelligence*, 32(5), 805-829. <https://doi.org/10.1080/0952813X.2019.1673488>.