International Journal of Data Warehousing and Mining
Volume 20 • Issue 1 • January-December 2024

Arabic Clustering Through Advanced Stemming and WordNet-Based Extraction for Water Cycle Cluster

Deema Mohammed Alsekait
Department of Computer Science and Information Technology, Applied College, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

Jaffar Atwan
https://orcid.org/0000-0003-2322-822X
Department of Computer Information System, Prince Abdullah Bin Ghazi Faculty of ICT, Al~Balqa Applied University, Al-Salt, Jordan

Qusay Bsoul
Cybersecurity Department, College of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan

Sharaf Alzoubi
College of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan

Hanaa Fathi
Applied Science Research Center, Applied Science Private University, Amman, Jordan

Malik Jawarneh
College of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan

Abeer Saber
Benha Univerity, Egypt

Diaa Salama AbdElminaam
https://orcid.org/0000-0003-0881-3164
MEU Research Unit, Middle East University, Amman, Jordan & Jadara Research Center, Jadara University, Irbid, Jordan

ABSTRACT

Natural language processing represents human language in computational technique, which is to achieve the extraction of important words. The verbs and nouns found in the Arabic language are significantly pertinent in the process of differentiating each class label available for the purpose of machine learning, specifically in 'Arabic Clustering'. This paper implemented the extraction of verbs and nouns sourced from the Qur'an and text clustering for further evaluation by using two datasets. The limitations of conventional clusters were identified, such as k-means clustering on the initial centroids. Therefore, the current work incorporated a novel clustering optimisation technique known as the water cycle algorithm; when combined with k-means, the algorithm would select the optimal initial centroids. Consequently, the experiments revealed the proposed extraction technique to outperform other extraction methods when using an actual Qur'an dataset.

KEYWORDS

Arabic Clustering, Noun and Verb Extractions, Qur'an, K-Means, Water Cycle Optimisation, Real Data and TREC

INTRODUCTION

In natural language processing, the extraction method primarily depends on algorithms and considers the textual representation of an entity within a domain. A person with pertinent knowledge of a given term may observe that the process of extracting terms is simple. Regardless, that person might also misidentify a few terms, which can be linked to subjectivity and variances in decision making (Al Zamil & Al-Radaideh, 2014). Within the field of computing, automatic term recognition, also known as term extraction, refers to the methods used to extract a set of actual words that are

DOI: 10.4018/IJDWM.352601

deemed relevant within a given text. According to Cimiano (2006), these algorithms have minimal capacities for sentence interpretation and critical information extraction (see also Cambria & White, 2014). Nevertheless, they can retrieve texts efficiently, segment them into brief excerpts, verify the spelling of words, and determine the quantity of lexical items.

Three steps are typically involved in term extractions: candidate term generation, candidate term scoring, and validation. Typically, text pre-processing techniques like Bag of Words (BOW) or named entity recognition start the first stage of candidate term generation. One filter that might be used in this particular step using statistical techniques is stop word removal. The next step of candidate term scoring yields an analysis of the significance of the candidate and the suitability of the terms generated. However, the third and last stage requires confirming the candidate's accuracy and precision, a process that depends on the availability of resources (Kanan & Fox, 2016). There are a few domains where the public can access the gold standard. Others, on the other hand, such as the Qur'an, are restricted to specific sections or elements or are even utilized with translations instead of the original Arabic text (Beirade et al., 2019).

Text clustering is one of the most widely used methods for identifying events, topics, or text types (Bsoul & Mohd, 2011). Generally speaking, the method uses three main steps to accomplish its objective (Bsoul & Mohd, 2011; Abu-Salem et al., 1999). The first step in text pre-processing involves removing unnecessary words and symbols from any document, including those that describe crimes, finances, or the Qur'an. The second step then attempts to extract key terms—like BOW—from texts that contain information that has been given priority. Ultimately, the Qur'anic text and conventional cluster algorithms must be used to evaluate the final procedure.

Though it can take up all terms after stop and root words are removed in a process known as stemming, a BOW cannot accurately represent the most critical terms in Arabic texts (Al Zamil & Al-Radaideh, 2014; Bsoul & Mohd, 2011). Because Arabic texts developed quickly, it is notable that the textual data contains a variety of vocabulary. Additionally, because BOW is highly dimensional, it is essential for an extraction technique. In addition to enhancing extraction performance and, ideally, processing small data, it should be able to characterize the documentary's theme accurately. In addition, some sentence components—like nouns—have been adopted for use in English in recent years (Fodeh et al., 2011). For improved text clustering quality, the English thesaurus, for example, has been implemented comprehensively [9–18]. Specifically, Romanians have defined “crime” in terms of “actus-reus” (awful or harmful “verbs”) and “mens-rea” (bad or horrible intention “nouns”), emphasizing the importance of verbs and nouns in sentences in distinguishing between various groups (Mishra, 2020).

Thus, this study proposes a novel approach to assessing Arabic extractions using Arabic clustering and Qur'anic texts. It was restricted, though, to introducing novel procedures for verb extraction assessment and extraction. It used the Arabic news data set as a reference and the Qur'an as an actual data set. The Qur'an text should be utilized as a real data set to evaluate the efficacy of this new extraction technique, and Arabic clustering should be employed in computational applications.

The remainder of the paper is organized into sections: In the second section, which will highlight most works related to extraction methods, real and benchmark data sets demonstrate the effect of nouns as extraction and effect verbs as extraction. In the fourth section, the effects of extracted nouns and verbs are illustrated using the following four recommended methods: (a) extraction using the educated text stemmer (ETS), (b) extraction using Arabic WordNet (AWN), and (c) extraction using combinations of the ETS and AWN. The fifth section focuses on two approaches to selecting the initial centroids using the water cycle cluster (WCC), a recently proposed optimization clustering technique. The sixth section discusses the parameter settings the WCC was exposed to and the cluster's results. The last section includes recommendations for more research as well as the conclusion.

MATERIALS AND METHODS

The process of extracting terms from indexed lexical items is often regarded as extremely difficult due to the need to consider multiple or a combination of terms. Researchers have focused a lot of attention on information extraction based on terms or features, employing methods such as entity recognition (Mesmia et al., 2018), BOW (Hmeidi et a., 2008; Alsmearat et al., 2014), *n*-gram (Al-Salemi & Aziz, 2011), and ETS (Al-Shammari & Lin, 2008). To make the phase better and more productive, they use ontology-based extractions and semantic extractions (Al-Zoghby et al., 2018), Arabic word sense disambiguation (Etaiwi & Awajan, 2020; Salloum et al., 2018), semantic word embedding (El Mahdaouy et al., 2018), and semantic relationships (Benabdallah et al., 2017). However, textual data often pose challenges due to their highly dimensional nature and overlapping or ambiguous word senses. Previous work has defined several term extraction methods, including BOW, *n*-gram, and named entity recognition, to tackle the issue of high dimensionality. Although semantic extractions are used to describe overlapping word senses, such as the semantic words displayed in AWN, these methods are also called syntactic extractions. Thus, this paper addresses syntactic extractions and proposes a novel idea for enhancing the extraction procedure.

Two grammatical genders (male and female), three grammatical numbers (plural, dual, and singular), and three grammatical cases (accusative, genitive, and nominative) are generally used to characterize the Arabic language. For instance, an Arabic noun can be either nominative or accusative, depending on whether it is the subject of a sentence or not. It can also be nominative if it is the object of a verb in some sentences. Additionally, if it is the object of a preposition in a sentence, it may be of the genitive type. As a result, it becomes challenging to recognize an Arabic noun form precisely because one must consider its grammatical number, gender, and case.

Furthermore, it is deemed extremely important to consider the necessary degree of analysis when applying the stemming algorithms to a document containing Arabic texts. Strictly speaking, root-based or stem-based algorithms are used depending on the required level of analysis (Al-Shammari & Lin, 2008). Al-Smadi et al. (2019) used supervised machine learning to focus on extraction based on morphological, syntactic, and semantic Arabic terms. They found that syntactic extraction is better than morphological and semantic extraction. Furthermore, Harrag (2014) used the named entity approach to extract the most significant Arabic terms, which has limitations when differentiating between similar sentences. Because of this drawback, the method employed by Helwe and Elbassuoni (2019) only used word embedding from the Arabic name entity to distinguish between two similar sentences. Mustafa (2005) devised a way to search for Arabic text in Arabic information retrieval using two types of information extraction: hybrid *n*-grams and contiguous *n*-grams. The hybrid *n*-grams work better than the contiguous *n*-grams, but both have nouns, making it hard to tell the difference between nouns and verbs when they are turned back to their root stem.

Al-Salemi and Aziz (2011) evaluated the BOW method and the three *n*-gram levels (3, 4, and 5) for extraction. Applying the naïve Bayes classifier indicated that the BOW performed better than all *n*-gram levels examined. In the meantime, Al-Shamari and Lin (2008) stemmed nouns with verbs by employing the ETS, a novel technique, as the stemmer for Arabic root words. The algorithm aims to select nouns with verbs from Arabic documents based on prepositions and other linguistic rules, like the definite article "the."

A review of the literature outlining the extraction procedure (Table 1) indicates that researchers have validated their assessments by disclosing how verbs and nouns are extracted during syntactic extractions. As stated in the first section, the extractions are used directly as a stemmer process, so these approaches have yet to be tested. Furthermore, verbs are significant terms in a text; the Romanian nation's conception of crime conceptually supports the strategy. This theory, previously mentioned in other works, including those by Al-Shamari and Lin (2008), describes crime through the two main aspects of "nouns with verbs." Nonetheless, earlier studies have also demonstrated the application of conventional clustering to evaluate the suggested extraction techniques. Consequently, the Qur'anic

Table 1. Some works describing Arabic extraction

Authors	Data Set Used	Baseline Extraction	Outperform Extraction	Domain
[24]	Data set manual	Khoja and Larkey stemmers	Noun with verbs for stemmer	Arabic clustering
[23]	TREC-2002	N-gram 3, 4, and 5 levels	BOW	Arabic classifiers
[20]	Arabic Wikipedia corpus	Name entity and their our system	Name entity and their our system are equivalent	Arabic classification
[21]	Collection of news articles	BOW	BOW with support vector machines as classifiers better than K-nearest neighbor	Arabic text categorization
[28]	Arabic TREC collection	BOW	Word embedding similarities	Information retrieval
[25]	Data set manual	Ontology 'synonym, antonym, hypernym' and complex extraction	BOW	Marker learning algorithm

Note. BOW = bag of words

text was utilized in this paper for the evaluation, and the Qur'an data set and a manually compiled data set of Arabic news were used to assess the suggested extraction method. In light of this, the new extraction method is evaluated in the next (third) section.

THE PROPOSED EDUCATED TEXT STEMMER AS EXTRACTION

According to the ETS proposed by Al-Shamari and Lin (2008), the stemmers of Arabic text can be depicted according to several prevalent characters. Nevertheless, the main distinction between such stemmers pertains to their grammatical structure. Figure 1 demonstrates the key stages of the Arabic ETS, whereby grammatical structures and knowledge are employed to identify the verbs and nouns. It can be described according to four rules.

Rule 1. The preceding words, list stop verbs and list stop nouns, will flag the words to their category.
Rule 2. The definite articles, such as "ال," that start with words will be flagged as nouns.
Rule 3. The following verb is distinguished as either a noun or a stop word. If recognized as a noun, it will be added to the noun flag.
Rule 4. The flagged noun and verb lists are used as a lookup table, which allows unflagged nouns and unflagged verbs to be identified. These rules have been established by Al-Shamari and Lin (2008).

Table 2 depicts some examples detailing the identification of nouns with verbs according to the four rules of the ETS, which identification is observed from Chapter 2, Verse 25:

وَبَشِّرِ الَّذِينَ آمَنُوا وَعَمِلُوا الصَّالِحَاتِ أَنَّ لَهُمْ جَنَّاتٍ تَجْرِي مِن تَحْتِهَا الْأَنْهَارُ كُلَّمَا رُزِقُوا مِنْهَا مِن ثَمَرَةٍ رِّزْقًا قَالُوا هَٰذَا الَّذِي رُزِقْنَا مِن قَبْلُ وَأُتُوا بِهِ مُتَشَابِهًا وَلَهُمْ فِيهَا أَزْوَاجٌ مُّطَهَّرَةٌ وَهُمْ فِيهَا خَالِدُونَ (25)

Translation in English:

And give good tidings to those who believe and do righteous deeds that they will have gardens [in Paradise] beneath which rivers flow. Whenever they are provided with a provision of fruit therefrom, they will say, "This is what we were provided with before." And it is given to them in likeness. And they will have therein purified spouses, and they will abide therein eternally.

Table 2. Primary steps of the Arabic educated text stemmer by Al-Shamari and Lin (2008)

Educated Text Stemmer					
اقزر	قرمث	نم	اوقزر	املك	راەنألا
therefrom	fruit	of	provided	whenever	Rivers
verb	noun	useful word preceding noun	verb	useful word preceding verb	definite articles
founded by the R4	founded by the R1	founded by the R1	founded by the R1	founded by the R1	founded by the R2

RESULTS AND ANALYSIS FOR QUR'AN

Four sets of experiments were carried out in this study, and each experiment used the Qur'an data set's chosen extraction techniques. Then, using the ETS as the stemmer in all four suggested methods, they were directly compared to the true label.

Noun and Verb Extractions Using Arabic WordNet

The current study identified and used the Arabic data set's verbs and nouns as terms. This study used AWN to determine whether a word could be a verb or noun by looking at whether the word was appropriately provided as a verb or noun in the AWN database. As a result, each document's vector of features included a collection of stemmed verbs and nouns generated during the pre-processing phase. This study looked at how well the AWN-based method for identifying verbs and nouns worked by using examples from two real data sets of the Qur'an that were sourced from the Language Research Group at the University of Leeds. Table 4 compares the extraction performance by utilizing the Qur'an

Figure 1. The educated text stemmer

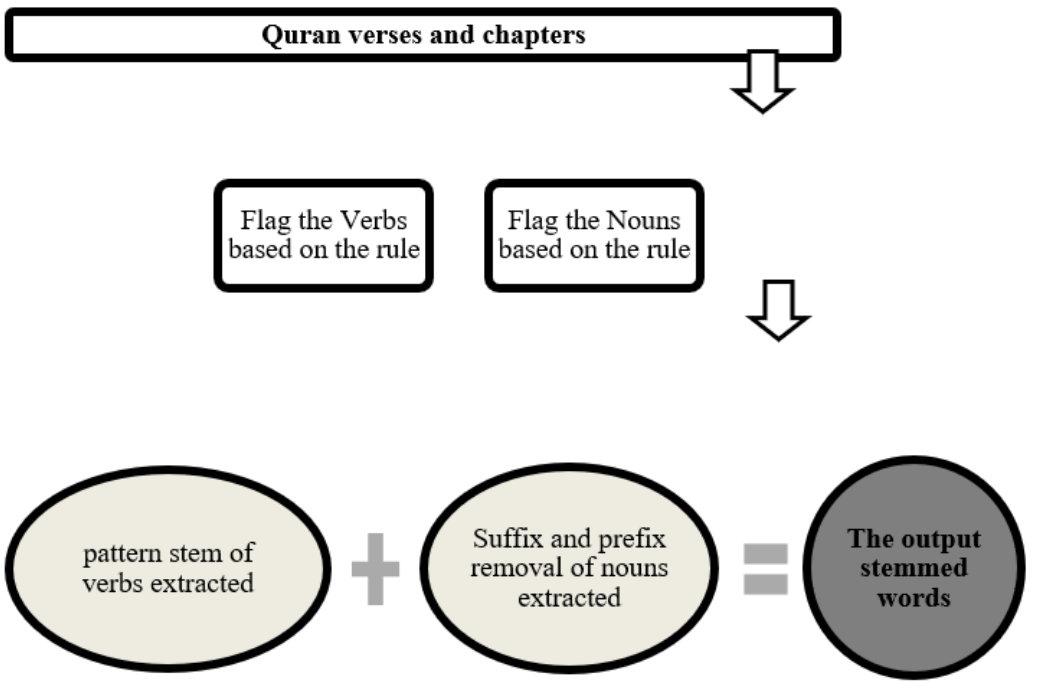


Table 3. The Arabic stemming algorithm of the educated text stemmer (Al-Shamari & Lin, 2008)

Lemmatization
Input: Arabic documents.
Noun Dictionary. Verbs Dictionary.
V: Verb dictionary (one-dimensional array sorted alphabetically).
N: Noun dictionary (one-dimensional array sorted alphabetically).
NSW: Array of stop words proceeding nouns.
VSW: Array of stop words proceeding verbs.
SW: Array of stop words (including both NSW and VSW).
Step 1: Remove useless stop words.
Step 2: Locate words attached to definite articles and proceeded by NSW, and flag them as nouns.
Step 3: Add nouns to the noun dictionary N.
Step 4: Locate verbs proceeded by VSW. Flag verbs in the document.
Step 5: Add identified verbs to the verb dictionary V.
Step 6: Revisit the document searching for existing nouns and verbs.
Step 7: Tokens (words) with missing tags are treated as nouns.
Step 8: Remove the remaining stop words (useful stop words).
Step 9: Apply light stemming algorithm on nouns.
Step 10: Apply Khoja's root-based stemmer on verbs.
Output: Stemmed documents.

data set class labels. The tabled content of BOW demonstrates the performance obtained using only the stemmed nouns (as provided in AWN) and only the stemmed verbs as features. It is evident that fewer verbs were extracted as terms than were labeled; of the 19,356 terms, AWN extracted 3,288.

Furthermore, only 7,553 out of 25,135, or less, of the correct number of words were extracted from the nouns. Likewise, the 10,841 words created by combining nouns with other nouns do not match the 4,441 numbers in the true label. This result was predicted given the weak structure and vast vocabulary of AWN. After AWN identified the stemmed verbs, the procedure only employed those features.

Noun and Verb Extractions Using the Educated Text Stemmer

This work proposed an extraction method for educated text stemmers, using the algorithm of Al-Shamari and Lin (2008) to use the ETS as an extraction tool for a stemmer. According to Section 3's description of stop-word proceeding nouns and verbs, the algorithm extracts the terms and "words" from the text. Table 4 lists the verbs that the ETS was used to extract as terms. AWN and true labeled terms outnumbered true verb terms, with 436 extracted using the ETS, out of 19,356. It was discovered that the nouns extracted by the ETS were 7,553 out of 18,623, which was less than the true label but higher than AWN. Consequently, the performance of this extraction method could have been improved.

In this study, the Arabic ETS was adopted as an extraction method in combination with AWN. Therefore, the words and terms were extracted by using the ETS as extraction; there are some words and terms that are unflagged as nouns or as verbs. These words and terms will be passed to AWN to extract the nouns and verbs from them. Table 5 reveals the different kinds of methods used for extraction accordingly. Similarly, Table 4 shows that the extraction of verbs using the ETS combined with AWN, 681 terms, was better than the ETS with 436. The number of verbs extracted using AWN, 3,288 terms, was better than the extraction of verbs using the ETS combined with AWN, 681 terms, out of 19,356 in true labeled. In contrast, the nouns as extraction yielded 18,627 out of 25,135 words,

Table 4. Results of four extraction methods using Corpus Qur'an as the true label

Methods extraction	AWN	ETS	ETS+AWN	AWN+ETS	True labeled
	# Term	# Term	# Term	# Term	#corpus
BOW	47524	47524	47524	47524	68316
Nouns	7553	18623	18627	24110	25135
Verbs	3288	436	681	3316	19356
Nouns with Verbs	10841	19059	19308	27426	44491

Table 5. The difference between the four methods proposed as extraction

Arabic-WordNet	ETS	ETS with Arabic-WordNet	Arabic-WordNet with ETS
Pre-processing	Pre-processing	Pre-processing	Pre-processing
1- Check each words if the word flagged as verbs moves it to verbs flag.2- If the word flagged as nouns move it as nouns flag.3- Remove non- flag words.	1- Based on the rules of algorithms, flag the noun words and verb words.2- Remove non- flag words.	1- Based on the rules of algorithms, flag the noun words and verb words.2- Non-flagged words use the AWN to flag it as noun or verb.3- Remove non- flag words.	1- Check each words if the word flagged as verbs moves it to verbs flag.2- If the word flagged as nouns move it as nouns flag.3- Non- flagged words applies the rules of ETS to flag it as noun or verb.4- Remove non- flag words.

which was significantly better compared to AWN and the ETS as extraction method. In addition, the combined method between noun and verb words as extraction generated 19,308 out of 44,491 words, worse than the true labeled method but better than the previous two methods proposed.

Arabic WordNet Extraction Combined With the Educated Text Stemmer as Extraction

The last proposed method for detecting verb and noun words was to use AWN first and then the unflagged words and terms to be detected as nouns and verbs using the ETS proposed in Section 4. Table 4 reveals that the extraction of verbs as terms yields 3,316 out of 19,356 words. The verb extraction using AWN with ETS reveals better outcomes than the three previously proposed methods, but it needs to encompass the correct number of verb words. Meanwhile, the noun extraction yields results of 24,110 out of 25,135, which is better in comparison to the previous three proposed methods but remains non-encompassing of all correct numbers of noun words. In addition, the combination of noun and verb words results in 27,426 out of 44,491 words; this result is worse than the true labeled terms but better than when using the ETS first, followed by AWN as extraction. Therefore, the combination of the two extraction methods is not more decadent or yields the correct number for the true labeled terms, but the last proposed AWN combined with the ETS as extraction is better than the other extraction methods. The BOW as extraction is shown in Table 4; it is a popular method of extraction that shows all words are extracted after pre-processing, removal of stop words, and stemming of words.

The main problem addressed in the proposed extraction method is justified in the following example (Chapter 2, Verse 20):

وَلَوۡ ۚ اوُمَاقَ مۡهِيۡلَعَ مَلۡظَاۡ اَذۡاِوۡ هِيۡف اۡوَشۡمَ مۡهَلَ ءَاضَاۡ اَمَّلُكۡ ۖمۡهَرَاصۡبَاۡ فَطۡخَ يۡقُرۡبَلۡا ُدَاكَي
ري ِدَقۡ ءۡيَشۡ ِلُكۡ ىَلَعۡ َهَّللا َنَّاۡ ىَلَعۡ َهَّللا ُلۡهَدۡبَ ۡمِهۡعۡمَسۡب َبَهَذَلۡ ُهَّللا َءَاشَ ۡمِهِرَاصۡبَاۡوَ (20)

Translation in English:

The lightning almost snatches away their sight; whenever it flashes for them, they walk therein, and when darkness covers them, they stand still. And if Allah willed, He could have taken away their hearing and their sight. Certainly, Allah has power over all things.

Table 6. Example of two problems found in proposed extraction method

ٍءْيَش ِّلُك ٰىَلَع َهَّللا َّنِإ ۚ مِهِراَصْبَأَو ْمِهِعْمَسِب َبَهَذَل ُهَّللا َءاَش ْوَلَو ۚ اوُماَق ْمِهْيَلَع َمَلْظَأ اَذِإَو ِهيِف اْوَشَّم مُهَل َءاَضَأ اَمَّلُك ۖ مُهَراَصْبَأ ُفَطْخَي ُقْرَبْلا ُداَكَي ريِدَق (20) The lightning almost snatches away their sight, whenever it flashes for them, they walk therein, and when darkness covers them, they stand still. And if Allah willed, He could have taken away their hearing and their sight. Certainly, Allah has power over all things.							
useful word preceding verb	**verb**	**useful word preceding noun**	**definite articlesال**	**noun**	**flag verb and noun by AWN**	**non-catch**	**useless stop word**
اذاإملك	ملظأءاضأ	ىلع	قربلا	لكقرب	مهراصبأومهراصبأداكي	وماقاوشمفطخي ببهذلهللاءاشا ريدقهللامهعمس	ءيشنإولومهيلعهيفمهل

When we detected "مهراصبأ," "sight" was not extracted by AWN but extracted by the ETS and flagged as a noun, that applied suffix and prefix as steamer on it because it is a noun and became "راصب" as stem, which is not the correct root; the correct root is "رصب." The other problem is that the number of terms or words extracted was less than the correct number in true labeled, as shown in Table 6, which consists of 19 words and six useless stop words. As demonstrated in Table 6, the noncatches/extracted are nine words:'يخطف' يخطف, مشوا, قاموا, شاء, الله,ذهب, ب سمعهم, الله, قدير '; this was the reason why the extraction method proposed did not obtain the correct number as per the true labeled component. Moreover, the following section utilizes Arabic clustering to evaluate the proposed extraction method and detect the proposed machine-learning technique using two data sets. This was done by utilizing a benchmark data set and a manually collected Arabic news data set.

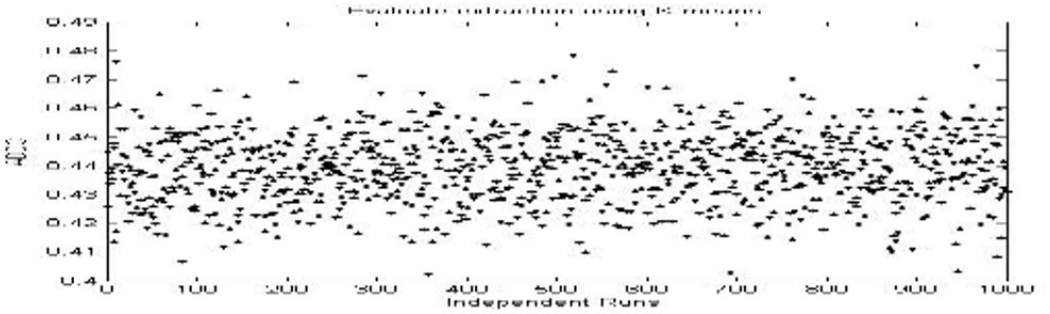
ARABIC CLUSTERING EVALUATION

The clustering process (Wei et al., 2015; Montalvo et al., 2015) generally includes grouping the objects based on resemblance. The *k*-means approach performs the critical role of partitioning during clustering (Wei et al., 2015; Alghamdi & Selamat, 2019) and is often used to undertake partition-based clustering with linear time complications (Anitha & Patil, 2019). Furthermore, Hartigan [37] applauded the primary goal of the *k*-means algorithm, namely, the document mean given to such clusters that is thus utilized to depict each *k* cluster. The mean is called the centroid of the cluster. However, the *k*-means algorithm lacks the sensitivity for the initialization. It requires the number of clusters from the initial itself, that is, the number of clusters needs to be known prior (Wei et al., 2015). Besides, the initial centroids play a vital role in the clustering performance and may cause the algorithm to be stuck in a locally optimum solution (Wei et al., 2015; Wan et al., 2018). The experiment outcomes, as seen in Figure 2, illustrate the problem of local optima, whereby 1,000 independent runs have either good or bad performance each. Note that a locally optimum solution means the clustering algorithm cannot find good clusters during the clustering process.

The *k*-means algorithm must be integrated with specific optimization procedures in order to increase its performance and become less dependent on a given data set and initialization. This makes it possible to find the best clustering centers, refine the clustering centroids further, and obtain good initial clustering centroids (Sahmoudi & Lachkar, 2017). Many fields, including computer science (Senseney & Dickson, 2018), data mining (Jones et al., 2018), industry (Wang & Zhang, 2020), agriculture (Zhang et al., 2016), computer vision (Schmidt et al., 2020), forecasting (Shah et al., 2016), medicine and biology (Colebunders et al., 2014), scheduling (Guo et al., 2009), economy (Malo et al., 2014), and engineering (Girdhar & Bharadwaj, 2019), have adopted the use of nature-inspired metaheuristic algorithms. It served as an Arabic optimization cluster in this work.

This study, however, was restricted to demonstrating the shortcomings of harmony search (HS) with *k*-means, as suggested by Forsati et al. (2013), using their optimal parameter configurations.

Figure 2. The distribution of 1,000 runs of K-means using bag of words



The results of HS clustering and one-step *k*-means (harmony *k*-means) are contrasted in Figures 3 and 4. When compared to HS clustering, the combined method produces superior results. However, a comparison of the outcomes displayed in Figures 2 and 4 indicates that, on occasion, *k*-means clustering produces superior results than harmony *k*-means. Therefore, the HS's limitation suggests that an alternative optimization strategy that can get around this flaw be proposed. As a result, three control parameters—PHMCR, PPAR, and BW—that are referred to as high constraints and trapped in local optima can be used to describe the HS algorithm (Forsati et al., 2013; Yang et al., 2009). This work suggests the use of water cycle optimization as a clustering technique to address the shortcomings of HS.

Water Cycle Clustering

At the start of the put-forward method, an initial population, otherwise described as raindrops, is considered, whereas the best individual selected is the sea. Following this, a number of good raindrops are selected to form a river, while the remaining raindrops are categorized as streams that would flow into the sea and rivers. Therefore, rivers accept water sourced from streams with consideration for their flow level. Moreover, the water in the streams that flow and meet the sea and the rivers is not the same in terms of quantity; it will differ from one stream to another. Furthermore, the rivers flow into the sea, specifically toward the most downhill location. Only one control parameter limit is associated with the water cycle optimization algorithm.

In contrast, three control parameters can be linked with the HS algorithm, namely, PPAR, PHMCR, and BW (Forsati et al., 2013). This has resulted in the development of a new meta-heuristic algorithm, such as water cycle algorithm optimization, to fill some of these knowledge gaps (Eskandar et al., 2012). Nevertheless, water cycle optimization is of little help in text clustering, despite this type of optimization being helpful for the algorithm to avoid rapid convergence (immature convergence). It can also prevent getting trapped in the local optima by employing the evaporation technique.

All suggested algorithms to display the documents use the vector-space model, which represents each term as one dimension of the matrix spaces. As a result, each document di = (wi1, wi2,... win) is seen as a vector with *n* distinct terms in the term space. In the meantime, one potential clustering solution is the vector of centroids. As a result, clustering is considered an optimization task where the goal is to find the optimal cluster centroid rather than the optimal partition. Accordingly, the clustering quality was chosen based on an objective function, and water cycle clustering was used to optimize the objective. In essence, this approach helped to clearly address the clustering goal, which helped to understand better how well the clustering algorithm performed about particular data types; eventually, this will allow task-specific clustering objectives. According to a prior study, the approach also offers the advantage of allowing for the simultaneous consideration of multiple objectives (Handl & Knowles, 2007). Moreover, it is essential to choose several design options when using a general-purpose optimization meta-heuristic for clustering. The dominant options in this

Figure 3. Harmony search as a cluster with 1,000 iterations

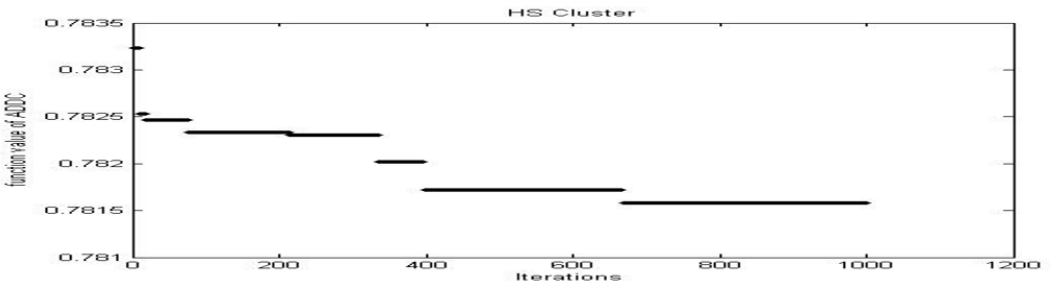
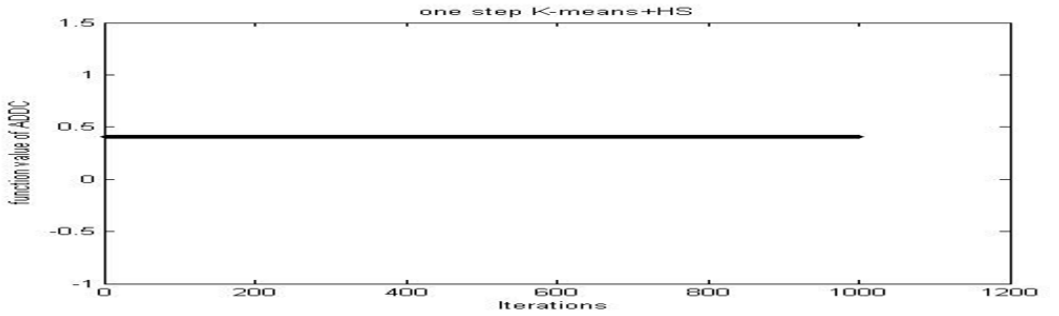


Figure 4. Combination of K-means with harmony search for 1,000 iterations



instance indicate the objective function and problem representation correspondingly; both elements may significantly affect the optimization's performance and the quality of the clustering.

The put-forward algorithm employs certain representations to code the document set's whole partition, P, along with a vector of length, m, which denotes the number of documents presented in Figure 5. Every element acts as the label for this vector that a single document identifies. For instance, if the total number of clusters is represented by K, the solution vectors for each element give an integer value that falls in the range of [K] = {1,..., K}. Furthermore, the assignment that denotes K non-empty clusters acts as a legal assignment, wherein every assignment includes a correspondence associated with a set of K centroids. Similarly, the search space represents the space for all permutations pertaining to size m from the set {1… K}, which meets the constraint enforcing the algorithm. The search space allows the assignment of each document to precisely one cluster such that no cluster remains empty. Thus, the problem is considered NP-hard even when the value for K is 2. Subsequently, a natural method for encoding this type of permutation is by considering each row pertaining to the WCC as an integer vector with m positions, whereby the ith position depicts the cluster assigned to the ith document. Figure 5 provides an example showcasing the solutions. In this example, four documents (2, 3, 7, and 8) originating from the cluster were assigned to Label 3, while the cluster assigned to Label 2 included three documents (4, 6, 9), and so on.

Generation of Initial Clusters

The values pertaining to the problem variables generally incline toward forming an array. In terms of particle swarm optimization and genetic algorithm lexicons, such array is referred to as particle position and chromosome, respectively. Therefore, the label is termed as "raindrop cluster"

Figure 5. Some documents represented by their number of groups from 1 to 5

Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc10	Doc11	Doc12
5	3	3	2	1	4	3	3	2	5	2	5

and defines a single cluster. In the case of an Nvar dimensional cluster problem, a raindrop can be defined as an array of 1 ×Nvar. This array is defined and shown in Equation 1.

$$\text{Raindrop cluster} = [X1, X2, X3 \ldots XN] \tag{1}$$

In Equation 1, at the start of the clustering algorithm, the generation of a candidate (i.e., raindrops of the cluster) that signifies a matrix of raindrops with size Npop×Nvar is performed. Therefore, the arbitrarily generated matrix X can be presented as (columns signify the number of design variables and rows denote the number of clusters) shown in Equation 2.

$$\text{Raindrops of cluster} = \begin{bmatrix} Raindrop_1 \\ Raindrop_2 \\ Raindrop_3 \\ \vdots \\ Raindrop_{Npop} \end{bmatrix} \begin{bmatrix} x_1^1 x_2^1 x_3^1 & \cdots & x_{Nvar}^1 \\ \vdots & \ddots & \vdots \\ x_1^{Npop} x_2^{Npop} x_3^{Npop} & \cdots & x_{Nvar}^{Npop} \end{bmatrix} \tag{2}$$

In Equation 2, the floating-point number (real values) can be employed to represent every decision variable value (X1, X2, X3... XNvar), whereby Npop denotes the number of raindrops (initial cluster) and Nvars represents the number of design variables. First, the creation of Npop raindrops is carried out, after which the cost of a raindrop can be determined by assessing the cost function (Cost), as shown in Equation 3.

$$\text{Costi} = f(x\_1\text{^}i, x\_2\text{^}i, \ldots x\_Nvar\text{^}i) \; i = 1, 2, 3, \ldots, Npop. \tag{3}$$

Cost of Solutions

The calculation of each solution in Npop corresponds to a document cluster, whereby each cell in the solution refers to the cluster number as C = (c_1, c_2…c_k). The C is the set of K centroids that corresponds to a row in Npop. The centroid of the kth cluster is c_k= (c_k1…c_kn), which can be computed as shown in Equation 4.

$$c\_kj = \frac{\sum_{i=1}^{m} aki \; dij}{\sum_{I=1}^{n} aki} \tag{4}$$

The objective function is to confirm the locus of the cluster centroids in a bid to ensure the intracluster similarity is at maximum (keeping the intracluster distance to minimum), while also minimizing the intercluster similarity (keeping the distance between clusters to minimum) concomitantly. The average distance of documents to the cluster centroid (ADDC) is represented by the particular row and defines the fitness value pertaining to each row, which subsequently corresponds to a possible solution. The ADDC is expressed as shown in Equation 5.

$$\text{Costi} = \left[\sum_{i=1}^{UB} \frac{1}{ni} \sum_{j=1}^{mi} D(Ccent, dj)\right] / \text{UB} \quad (5)$$

In Equation 5, the D (.,.) is the cosine similarity, mi is the number of documents in cluster i (e.g., (ni= $\sum_{j=1}^{n}$aij), dij is the jth document of cluster i, and UB is the number of clusters. The newly produced solution can be exchanged by a row in Npop in case of the locally optimized vector yielding better cost value compared to the solutions in Npop.

Numerous Nsr are selected by considering the best individuals (i.e., minimum values) as the rivers and sea. The raindrop with the minimum value in the lot is thus regarded as the sea. In fact, Nsr can be defined as the sum of the number of rivers (i.e., a user parameter) and a single sea, as showcased in Equation 6. Calculation of the remaining initial clusters (i.e., raindrops originating from the streams flowing to the rivers or have a chance to directly meet the sea) can be performed by considering Equation 7.

$$\text{Nsr= Number of Rivers+ (Sea=1)} \quad (6)$$

$$\text{NRaindrops= Npop - Nsr} \quad (7)$$

Equation 8 is then applied for assigning specific raindrops of cluster to the sea and rivers in terms of the intensity.

$$\text{NSn} = \text{round}\left\{\left|\frac{Cost_n}{\sum_{i=1}^{N_r} Cost_i}\right| \times N_{Raindrops}\right\}, \text{n} = 1, 2, \ldots, \text{Nsr} \quad (8)$$

In Equation 8, where NSn can be defined as the number of streams that flow to particular rivers or sea (Eskandar et al., 2012).

Stream Move to the Sea or Rivers

As per Section 5, the streams are produced from the raindrops, which combine to form new rivers. Several streams may also meet the sea directly, while all streams and rivers will finally join the sea (i.e., the best optimal cluster). A stream flows toward the river, which falls under the connecting line between them at an arbitrarily selected distance, as shown in Equation 9.

$$X \in (0, C \times d), C > 1 \quad (9)$$

In Equation 9, where C can be defined as the value between 1 and 2 (near to 2), C may have the best-selected value of 2, and d signifies the current distance between the river and the stream.

In Equation 9, the value of X may be a distributed random number (i.e., either uniform or of any possibly appropriate distribution) between 0 and (C× d). If the value of C is greater than 1, it allows the streams to flow in multiple varying directions towards the rivers. The concept may also be employed in rivers that flow toward the sea. Thus, establishing a new position for the rivers and streams can be performed as shown in Equations 10 and 11.

$$X_{\text{Stream}}^{i+1} = X_{\text{Stream}}^{i} + \text{rand} \times C \times \left(X_{\text{River}}^{i} - X_{\text{Stream}}^{i}\right) \quad (10)$$

$$X_{\text{River}}^{i+1} = X_{\text{River}}^{i} + \text{rand} \times C \times \left(X_{\text{Sea}}^{i} - X_{\text{River}}^{i}\right) \quad (11)$$

In Equations 10 and 11, rand can be defined as a uniformly distributed random number, whereby its value lies between 0 and 1. The positions of the stream and river are alternated (i.e., the river becomes a stream and vice versa) in the case of the solution provided by a stream, which functions better in contrast to its connecting river. Such alternation may also occur for the rivers and sea (Eskandar et al., 2012).

Stream Move to the Sea

Evaporation continues to be among the ultimatum factors that allow the algorithm to remain excluded from rapid convergence, that is, immature convergence (Eskandar et al., 2012). Naturally, rivers and lakes are exposed to water evaporation, while the photosynthesis process in plants allows water to transpire. The evaporated water then travels to the atmosphere to be transformed into clouds, which later condense when exposed to the colder atmosphere thus sending the water back to earth as rain. Subsequently, the rain results in the formation of new streams flowing to the rivers to ultimately meet the sea (Chen et al., 2020). In the proposed method, evaporation was considered for the sea by streams, and rivers end up flowing to the sea. The following pseudo-code would thus aid in determining whether the river would flow into the sea, as shown in Equation 12.

$$\text{If } \left|X_{Sea}^{i} - X_{River}^{i}\right| < d_{max} \quad i = 1, 2, 3, \ldots, N_{sr} - 1 \tag{12}$$

Evaporation and raining process end

In Equation 12, the value of dmax is generally a small number (almost zero). When the distance between the sea and a river is less than dmax, it signifies that the river has joined or met the sea. Therefore, the evaporation process was accounted for; naturally, with an adequate level of evaporation, raining (precipitation) would start. A large dmax value decreases the search intensity, while a small value spurs the search intensity closer to the sea. Thus, the search intensity closer to the sea (i.e., the optimum solution) is governed by dmax. Following this, an adaptive decrease in the dmax value is observed, as shown in Equation 13.

$$d_{max}^{i+1} = d_{max}^{i} - \frac{d_{max}^{i}}{\text{maxiteration}} \tag{13}$$

Raining Process

After the process of evaporation, the process of rain commences. In this process, the new raindrops create streams at several locations (i.e., act similarly to the genetic algorithm mutation operator). To determine the new sites of the newly generated streams, Equation 14 is employed accordingly.

$$X_{Stream}^{new} = LB + rand \times \left(UB - LB\right) \tag{14}$$

In Equation 14, where LB and UB represent the lower bound and upper bound, respectively; they are described by the specified problem.

Here, the superior newly created raindrop is considered a river flowing toward the sea, whereas the remaining clusters of newly created raindrops are supposed to form certain new streams that flow toward the river or directly to the sea. With the aim of improving the rate of convergence and the algorithm's computational performance for controlled problems, Equation 15 is employed specifically for the streams that flow directly toward the sea. This equation serves to spur the production of streams that flow toward the sea directly so as to develop sea exploration (i.e., the optimal cluster) in the viable area for controlled problems (Eskandar et al., 2012).

Table 7. Some scenarios of the parameters for water cycle as clustering

Scenario	Npop	Nsr
1	8	2
2	8	4
3	8	8
4	16	2
5	16	4
6	16	8
7	24	2
8	24	4
9	24	8

$$X_{Stream}^{new} = Xsea + \sqrt{\mu} \times rand\left(1, N_{var}\right) \quad (15)$$

In Equation 15, where μ represents the coefficient demonstrating the search region close-ranged to the sea and rand represents the random number that is normally distributed. The larger the μ value, the more probability is present for exiting the viable region. Conversely, the smaller the μ value, the smaller the search region in the sea. An appropriate μ value is set to 0.1. Scientifically, the term $\sqrt{\mu}$ found in Equation 15 represents the standard deviation; thus, μ describes the variance notion. By adopting such notions, the individuals that produce variance μ are distributed close to the most optimal cluster achieved, that is, sea (Eskandar et al., 2012), the pseudo-code of *k*-means combines with water cycle as unsupervised clustering shown in Figure 6.

Stop Criteria

The WCC ends upon two conditions: either when no change is seen in the fitness average by a predetermined value ε= dmax following certain iterations, or when the maximum amount of generation is achieved.

EXPERIMENTAL SETTING

Parameter Setting of Water Cycle as Clustering

This section seeks to examine the development of a solution for the technique's settings of two significant variables in the water cycle cluster. These variables are Npop and Nsr, where Nsr is the total number of rivers (a user variable) and one sea, and Npop is the number of raindrops (preliminary population). In this portion, the effects of the changes in a single variable are highlighted by examining three distinct situations, as displayed in Table 7. Moreover, the experiments demonstrate that the number of clusters yields the best outcomes if a linear relationship is found between Npop and Nsr. Every condition is examined using ten runs, and the maximum iteration number is fixed at 100 for each run. The ADDC value of the solution represents the fitness function value. The evaluation technique used is the WCC, described in Section 5, where the dmax is 1E03. The best condition is the scenario fifth condition, whereby Npop = 16 and Nsr = 4.

Performance Measurement and Data Sets

This research utilized the general F-measure to evaluate the external condition, which was prevalent among the Arabic clustering measures present (Larsen & Aone, 1999; Jardine & van

Rijsbergen, 1971). A greater overall F-measure offers the best cluster, whereby the F-measure metric for cluster validation integrates the notions of precision and recall for information retrieval purposes. Each cluster is regarded as the outcome of the classes, and every class is supposed to be the required document set for that class. The value of F-measure appears at the interval (0, 1), and the higher values of F-measure signify a greater quality of clustering.

The tests conducted in this research employed modern, unmarked, and unedited Arabic text, which consisted of a sample containing approximately 1,680 documents obtained from several Arabic online sources. The data set used for testing comprised four categories: politics, economics, sports, and art articles; each contained documents obtained from Bsoul and Mohd (2011) and Bsoul et al. (2014). The alternative collection of samples consisted of 383,872 Arabic documents, primarily made up of newswire dispatches issued by Agency France Press from 1994 to 2000. Standard Text Retrieval Conference (TREC; Atwan et al., 2015; Khorsheed & Al-Thubaity, 2013; El Mahdaouy et al., 2019) classes and base truth were also produced for this compilation; for TREC 2001, 10 classes were defined.

Results of the Water Cycle Cluster Algorithm

For this particular portion, *k*-means, single-step *k*-means having harmony search (WCC and KHS), and harmony search as clustering (HSCLUST) were utilized alongside actual and standard data sets. The measure of cosine correlation can be noted in the similarity measure in all of the techniques. At this point, it must be emphasized that the outcome given in the remaining portion is an average of more than 20 runs for the methods to ensure fairness. Furthermore, the techniques included 1,000 iterations for each run to simplify the comparison. No variable must be established in the case of the *k*-means technique. In the case of the WCC, every data set requires the Npop to be fixed at two times the number of classes present in the data set, while the Nsr is fixed at half of the number of classes in every data set. The same variables that Forsati et al. (2013) looked at are used for the combined *k*-means and HS method. The HMS is set to two times the number of clusters in the data set, HMCR is set to 0.6, PARmin is set at 0.45, and PARmax is set at 0.9.

To improve the technique, a single-step *k*-means technique was presented, whereby a new solution for clustering was produced by using the operations of the water cycle. The following process was also utilized to obtain a new solution. In this technique, the WCC's exploratory power and the *k*-means technique's fine-tuning power were interspersed in all iterations to ensure they yielded clusters of high quality; this was termed *k* with the WCC.

The performance of the algorithms in the collected documents for justifying the F-measure is displayed in Table 8. They made use of AWN with the ETS as the proposed method for extraction. In comparing and evaluating the outcomes for all techniques, the WCC with *k*-means reveals the most significantly excellent F-measure, whereas HS as clustering is obviously the worst cluster. The recommended WCC in this study is superior to HS, while the integration of the water cycle and *k*-means cluster surpasses the cluster of the water cycle alone.

On the basis of the outcomes given in Table 8, the integration of the water cycle and *k*-means cluster is used to assess all four suggested Arabic extraction methods. Table 9 subsequently demonstrates the integration of the ETS with AWN as the technique for extraction that surpasses other techniques (i.e., AWN and the ETS). Furthermore, verbs and nouns as extraction techniques are similarly superior compared to the method of just verbs or just nouns and BOW extraction. Therefore, the outcomes shown in Table 9 demonstrate AWN and BOW techniques as superior, as seen in the second column. At the same time, the nouns are better than verbs in all extraction methods. In certain cases, the nouns are also better than BOW.

Additionally, a statistical analysis was conducted to determine the optimal techniques for extracting and hybridizing the WCC and whether the outcomes of these hybridizations differ significantly from one another. The results of the proposed extract of nouns and verbs using AWN+ETS and evaluation by hybridizing *k*-means with the WCC are examined. According to the Friedman test,

Figure 6. Pseudo-code of water cycle as clustering combined with K-means

1 Set user parameter of the WCA: N_{pop}, N_{sr}, d_{max}, and Maximum_ Iteration.

2: **Input: Based on number of** N_{pop} a set of N unit-length document vectors $X = \{X_1,..., X_N\}$in IR^d and the number of groups$K$.

3: **Output:** a partition of the document vectors given by the cluster identity vector $Y = \{y_1, ...y_N\}$, $y_n \epsilon \{1, ..., K\}$

4: **Steps:**

1. **Initialization:** initialize the *unit-length* cluster centroid vectors $\{\mu 1,..., \mu_K\}$;
2. **Data assignment:** for each document vector x_n, sety_n = arg min $x_n^T \mu_K$;
3. **Centroid estimation:** for cluster k, let$X_k = \{x_n | y_n = k\}$, the centroid is estimated as$\mu_K = \sum_{x \epsilon x_k} x / || \sum_{x \epsilon x_k} x ||$;

5. ***Stop*** *if* y *does not change, otherwise go back to Step 2.*

6 Determine the number of streams (individuals) which flow to the rivers and sea using Eqs. (6) and (7).

7 **Use the Centroid estimated by K-means in step 4 sub-step** 3.

8 Define the intensity of flow (How many streams flow to their corresponding rivers and sea) using Eq. (8).

```
while (t < Maximum_Iteration) or (any stopping condition)
    for i = 1 : Population Size (Npop)
        Stream flows to its corresponding rivers and sea using Eqs. (10) and (11)
        Calculate the objective function of the generated stream using Eq. (4) and (5)
            if F_New_Stream < F_river
                River = New_ Stream;
                if F_New_ Strea m < F_Sea
                    Sea = New_ Stream;
                end if
            end if
        River flows to the sea using Eq. (12)
        Calculate the objective function of the generated river using Eq. (4) and (5)
            if F_New_ River < F_Sea
                Sea = New_ River;
            end if
    end for
    for i = 1: number of rivers (Nsr)
        if (distance (Sea and River) < dmax) or (rand < 0.1)
            New streams are created using Eq. (13)
        end if
    end for
    Reduce the dmax using Eq. (14)
end while
    Post-process results and visualization
```

Table 8. Results of three cluster algorithms using F-measure evaluation and proposed Arabic WordNet with educated text stemmer as extraction method

REF	Data Sets Name	K-WCC	HS	WCC	K with HS
DSM	Arabic news	**0.791**	0.602	0.681	0.788
DSR	TREC 2001	**0.615**	0.541	0.572	0.594

**Best result: underline and bold

Table 9. Results of four extraction methods using benchmark data set and manually collected Arabic news data set

Methods extraction	AWN		ETS		ETS+AWN		AWN+ETS	
	DSM	DSR	DSM	DSR	DSM	DSR	DSM	DSR
BOW	0.609	0.568	0.609	0.568	0.609	0.568	0.609	0.568
Nouns	0.511	0.476	0.602	0.555	0.642	0.562	0.664	0.571
Verbs	0.495	0.458	0.486	0.449	0.49	0.461	0.511	0.47
Nouns with Verbs	0.59	0.51	0.74	0.583	0.754	0.596	0.791	0.615

Table 10. Ranking of the proposed algorithms using Friedman Test

Algorithms	Ranking
K-means with WCC and AWN	10.8
K-means with WCC and ETS	10.15
K-means with WCC and ETS +AWN	10.01
HS with AWN+ETS	9.39
WCC with AWN+ETS	9.09
K-means with HS and AWN+ ETS	9.05
K-means with WCC and AWN+ETS	8.53
Friedman test (p-value) 0.00	
man-Davenport (p-value)	0.00

*The best in font bold

Table 10 ranks the suggested extraction techniques and the hybrid WCC with *k*-means and other algorithms as clustering (the lower the value, the higher the rank). The p-values for the Friedman and Iman-Davenport statistical tests are indicated in the final two rows of Table 10. The hybrid WCC with *k*-means and the suggested extract of noun and verb has the lowest value, as noted in the results tabulated in 10, and is therefore ranked first. The WCC with AWN+ETS, HS with AWN+ETS, hybrid *k*-means with HS and ETS +AWN, *k*-means with the WCC and ETS +AWN, *k*-means with the WCC and ETS, and *k*-means with the WCC and AWN are ranked second, third, fourth, fifth, sixth, and seventh, respectively.

The Wilcoxon test with 0.05 critical levels was conducted to verify whether these results are statistically different. The p-value of these different comparisons is presented in Table 11, where the "+" symbol means that the hybrid algorithm is statistically better than the nonhybrid algorithm (p-value < 0.05), while "-" means otherwise (i.e., p-value > 0.05), and "=" means there are no significant differences between the hybrid scheme and the individual algorithm (p-value $= 0.05$). These values prove that the hybrid algorithms are significantly better than the *k*-means and global search of the standard WCC using our proposed extraction methods.

Table 11. Evaluation of proposed on local and global searches using P-value of Wilcoxon Test

Methods extraction	Hybridize K-means with WCC VS K-means WCC	
BOW	+	+
Nouns	+	+
Verbs	+	+
Nouns with Verbs	+	+

Note. The "+" sign indicates that the p-value is less than or equal to the critical level (p-value ≤0.05).

CONCLUSION

The current paper aims to enhance Arabic text extractions via the proposed extraction of nouns with verbs only. Therefore, the verbs are important for differentiating from one group to another. To achieve such a method of extraction, the extraction of Arabic verbs was evaluated using two data sets, namely, the Qur'an text data set, the TREC 2001 data set, and the Arabic news clustering data set. This study combined the ETS and AWN to extract words using two different methods. The results clearly show that the combined method improves the extraction of nouns and verbs from the Qur'an and Arabic clustering while lowering the number of unnecessary words. It also shows in the experiments that clustering changes the way the data are extracted, with the *k*-means being sensitive to clusters for any initial centers.

Furthermore, a new optimization cluster called the WCC is proposed, whereby this work combines the powerful components of the WCC with the benefits of *k*-means. The results reveal that the combination outperforms the WCC and other clustering algorithms. Similarly, AWN and the ETS method together are better than other proposed extractions, whereas the worst extraction method is AWN.

Based on this limitation found during this work, we detect "مرأصاب," "sight" is not seen by AWN but flagged as a noun by ETS, so that applied suffix and prefix on it because it is a noun and becomes "راصاب," but the stem is "رصاب" Another example is that "نوعامس" is matched as a noun word and converted to "عامس." It is suggested that future works opt for an expansion pattern of the word, which has 39 kinds of words, such as فعال,فعيل,افعل, تفتعل. That is, patterns will convert such as راصاب and عامس to رصاب and عمس. Second, meta-heuristic feature selection should be employed, as it can reduce the number of features. In the third place, the suggested nouns with verbs as extraction should be tested using the Arabic classifier (Paci et al., 2013) domain since these algorithms depend less on the centers of each cluster. Fourth, the combined extraction should be evaluated from a semantic perspective, as it may generate a new extraction method concerning the problem shown and explained in Table 6. The proposed method in this work needs to be extended to retrieve the correct number of words from the Qur'anic text. It is hoped that the technique of nouns with verbs as extraction proposed in this work enhances the performance and effectiveness of Arabic clustering.

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

PROCESS DATES

07, 2024

This manuscript was initially received for consideration for the journal on 05/15/2024, revisions were received for the manuscript following the double-anonymized peer review on 07/31/2024, the manuscript was formally accepted on 07/18/2024, and the manuscript was finalized for publication on 07/31/2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Diaa Salama AbdElminaa; diaa.salama@miuegypt.edu.eg

REFERENCES

Abdulameer, A. S., Tiun, S., Sani, N. S., Ayob, M., & Taha, A. Y. (2020). Enhanced clustering models with wiki-k-nearest neighbors based representation for web search result clustering. *Journal of King Saud University. Computer and Information Sciences*.

Abu-Salem, H., Al-Omari, M., & Evens, M. W. (1999). Stemming methodologies over individual query words for an Arabic information retrieval system. *Journal of the American Society for Information Science*, *50*(6), 524–529. 10.1002/(SICI)1097-4571(1999)50:6<524::AID-ASI7>3.0.CO;2-M

Al-Salemi, B., & Aziz, M. J. A. (2011). Statistical Bayesian learning for automatic Arabic text categorization. *Journal of Computational Science*, *7*(1), 39–45. 10.3844/jcssp.2011.39.45

Al-Shammari, E. T., & Lin, J. (2008, October). Towards an error-free Arabic stemming. In *Proceedings of the 2nd ACM workshop on improving non-English web searching* (pp. 9-16). 10.1145/1460027.1460030

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. 10.1016/j.ipm.2018.01.006

Al Zamil, M. G., & Al-Radaideh, Q. (2014). Automatic extraction of ontological relations from Arabic text. *Journal of King Saud University. Computer and Information Sciences*, *26*(4), 462–472. 10.1016/j.jksuci.2014.06.007

Al-Zoghby, A. M., Elshiwi, A., & Atwan, A. (2018). Semantic relations extraction and ontology learning from Arabic texts: A survey. In *Intelligent natural language processing: Trends and applications* (pp. 199–225). Springer. 10.1007/978-3-319-67056-0_11

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. 10.1016/j.jksuci.2017.06.002

Alsmearat, K., Al-Ayyoub, M., & Al-Shalabi, R. (2014, November). An extensive study of the bag-of-words approach for gender identification of Arabic articles. In *2014 IEEE/ACS 11th international conference on computer systems and applications (AICCSA)* (pp. 601-608). 10.1109/AICCSA.2014.7073254

Anitha, P., & Patil, M. M. (2019). RFM model for customer purchase behavior using k-means algorithm. *Journal of King Saud University. Computer and Information Sciences*.

Atwan, J., Mohd, M., Kanaan, G., & Bsoul, Q. (2014, December). Impact of stemmer on Arabic text retrieval. In *Asia information retrieval symposium* (pp. 314–326). Springer. 10.1007/978-3-319-12844-3_27

Beirade, F., Azzoune, H., & Zegour, D. E. (2019). Semantic query for Quranic ontology. *Journal of King Saud University. Computer and Information Sciences*.

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. 10.1007/s10772-017-9405-5

Bsoul, Q., Al-Shamari, E., Mohd, M., & Atwan, J. (2014, December). Distance measures and stemming impact on Arabic document clustering. In *Asia information retrieval symposium* (pp. 327–339). Springer.

Bsoul, Q. W., & Mohd, M. (2011, December). Effect of ISRI stemming on similarity measure for Arabic document clustering. In *Asia information retrieval symposium* (pp. 584–593). Springer. 10.1007/978-3-642-25631-8_53

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. 10.1109/MCI.2014.2307227

Chae, G., Park, J., Park, J., Yeo, W. S., & Shi, C. (2016). Linking and clustering artworks using social tags: Revitalizing crowd-sourced information on cultural collections. *Journal of the Association for Information Science and Technology*, *67*(4), 885–899. 10.1002/asi.23442

Chen, C., Wang, P., Dong, H., & Wang, X. (2020). Hierarchical learning water cycle algorithm. *Applied Soft Computing*, *86*, 105935. 10.1016/j.asoc.2019.105935

Chen, T. T. (2016). The congruity between linkage-based factors and content-based clusters: An experimental study using multiple document corpora. *Journal of the Association for Information Science and Technology*, *67*(3), 610–619. 10.1002/asi.23413

Cimiano, P. (2006). *Ontology learning and population from text: Algorithms, evaluation and applications* (pp. 19-34).

Colebunders, R., Kenyon, C., & Rousseau, R. (2014). Increase in numbers and proportions of review articles in tropical medicine, infectious diseases, and oncology. *Journal of the Association for Information Science and Technology*, *65*(1), 201–205. 10.1002/asi.23026

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2018). Improving Arabic information retrieval using word embedding similarities. *International Journal of Speech Technology*, *21*(1), 121–136. 10.1007/s10772-018-9492-y

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2019). Word-embedding-based pseudo-relevance feedback for Arabic information retrieval. *Journal of Information Science*, *45*(4), 429–442. 10.1177/0165551518792210

Eskandar, H., Sadollah, A., Bahreininejad, A., & Hamdi, M. (2012). Water cycle algorithm: A novel metaheuristic optimization method for solving constrained engineering optimization problems. *Computers & Structures*, *110*, 151–166. 10.1016/j.compstruc.2012.07.010

Etaiwi, W., & Awajan, A. (2020). Graph-based Arabic text semantic representation. *Information Processing & Management*, *57*(3), 102183. 10.1016/j.ipm.2019.102183

Fodeh, S., Punch, B., & Tan, P. N. (2011). On ontology-driven document clustering using core semantic features. *Knowledge and Information Systems*, *28*(2), 395–421. 10.1007/s10115-010-0370-4

Forsati, R., Mahdavi, M., Shamsfard, M., & Meybodi, M. R. (2013). Efficient stochastic algorithms for document clustering. *Information Sciences*, *220*, 269–291. 10.1016/j.ins.2012.07.025

Gbadoubissa, J. E. Z., Ari, A. A. A., & Gueroui, A. M. (2018). Efficient k-means based clustering scheme for mobile networks cell sites management. *Journal of King Saud University. Computer and Information Sciences*.

Girdhar, N., & Bharadwaj, K. K. (2019). Community detection in signed social networks using multiobjective genetic algorithm. *Journal of the Association for Information Science and Technology*, *70*(8), 788–804. 10.1002/asi.24164

Gu, X., & Blackmore, K. L. (2019). Developing a scholar classification scheme from publication patterns in academic science: A cluster analysis approach. *Journal of the Association for Information Science and Technology*, *70*(11), 1262–1276. 10.1002/asi.24195

Guo, Y. W., Li, W. D., Mileham, A. R., & Owen, G. W. (2009). Applications of particle swarm optimization in integrated process planning and scheduling. *Robotics and Computer-integrated Manufacturing*, *25*(2), 280–288. 10.1016/j.rcim.2007.12.002

Handl, J., & Knowles, J. (2007). An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, *11*(1), 56–76. 10.1109/TEVC.2006.877146

Harrag, F. (2014). Text mining approach for knowledge extraction in Sahîh Al-Bukhari. *Computers in Human Behavior*, *30*, 558–566. 10.1016/j.chb.2013.06.035

Helwe, C., & Elbassuoni, S. (2019). Arabic named entity recognition via deep co-learning. *Artificial Intelligence Review*, *52*(1), 197–215. 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. 10.1016/j.aei.2007.12.001

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, *7*(5), 217–240. 10.1016/0020-0271(71)90051-9

Jones, K. M., McCoy, C., Crooks, R., & VanScoy, A. (2018). Contexts, critiques, and consequences: A discussion about educational data mining and learning analytics. *Proceedings of the Association for Information Science and Technology*, *55*(1), 697–700. 10.1002/pra2.2018.14505501085

Kanan, T., & Fox, E. A. (2016). Automated Arabic text classification with p-stemmer, machine learning, and a tailored news article taxonomy. *Journal of the Association for Information Science and Technology*, *67*(11), 2667–2683. 10.1002/asi.23609

Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language Resources and Evaluation*, *47*(2), 513–538. 10.1007/s10579-013-9221-8

Larsen, B., & Aone, C. (1999, August). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 16-22). 10.1145/312129.312186

Ma, J., & Lund, B. (2020). A cluster analysis of data mining studies in library and information science from 2006 to 2018. *Proceedings of the Association for Information Science and Technology*, *57*(1), e413. 10.1002/pra2.413

Ma, S., & Zhang, C. (2017). Document representation and clustering models for bilingual documents clustering. *Proceedings of the Association for Information Science and Technology*, *54*(1), 499–502. 10.1002/pra2.2017.14505401056

Ma, S., Zhang, C., & He, D. (2016). Document representation methods for clustering bilingual documents. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–10. 10.1002/pra2.2016.14505301065

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796. 10.1002/asi.23062

Mesmia, F. B., Haddar, K., Friburger, N., & Maurel, D. (2018). CasANER: Arabic named entity recognition tool. In *Intelligent natural language processing: Trends and applications* (pp. 173–198). Springer. 10.1007/978-3-319-67056-0_10

Mishra, P. (2020). Big data digital forensic and cybersecurity. In *Big data analytics and computing for digital forensic investigations* (p.183). 10.1201/9781003024743-9

Montalvo, S., Martínez, R., Fresno, V., & Delgado, A. (2015). Exploiting named entities for bilingual news clustering. *Journal of the Association for Information Science and Technology*, *66*(2), 363–376. 10.1002/asi.23175

Mu, T., Goulermas, J. Y., Korkontzelos, I., & Ananiadou, S. (2016). Descriptive document clustering via discriminant learning in a co-embedded space of multilevel similarities. *Journal of the Association for Information Science and Technology*, *67*(1), 106–133. 10.1002/asi.23374

Mustafa, S. H. (2005). Character contiguity in n-gram-based word matching: The case for Arabic text searching. *Information Processing & Management*, *41*(4), 819–827. 10.1016/j.ipm.2004.02.003

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. 10.1016/j.jksus.2012.12.001

Sahmoudi, I., & Lachkar, A. (2017). Formal concept analysis for Arabic web search results clustering. *Journal of King Saud University. Computer and Information Sciences*, *29*(2), 196–203. 10.1016/j.jksuci.2016.09.004

Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A survey of Arabic text mining. In *Intelligent natural language processing: Trends and applications* (pp. 417–431). Springer. 10.1007/978-3-319-67056-0_20

Savoy, J. (2015). Text clustering: An application with the state of the union addresses. *Journal of the Association for Information Science and Technology*, *66*(8), 1645–1654. 10.1002/asi.23283

Schmidt, T., Mosiienko, A., Faber, R., Herzog, J., & Wolff, C. (2020). Utilizing HTML-analysis and computer vision on a corpus of website screenshots to investigate design developments on the web. *Proceedings of the Association for Information Science and Technology*, *57*(1), e392. 10.1002/pra2.392

Senseney, M., & Dickson Koehl, E. (2018). Text data mining beyond the open data paradigm: Perspectives at the intersection of intellectual property and ethics. *Proceedings of the Association for Information Science and Technology*, *55*(1), 890–891. 10.1002/pra2.2018.14505501162

Shah, C., Hendahewa, C., & González-Ibáñez, R. (2016). Rain or shine? Forecasting search process performance in exploratory search tasks. *Journal of the Association for Information Science and Technology*, *67*(7), 1607–1623. 10.1002/asi.23484

Wan, Y., Zhong, Y., & Ma, A. (2018). Fully automatic spectral–spatial fuzzy clustering using an adaptive multiobjective memetic algorithm for multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 2324–2340. 10.1109/TGRS.2018.2872875

Wang, L., & Zhang, L. (2020). A quantitative text analysis of artificial intelligence industry policy in China. *Proceedings of the Association for Information Science and Technology*, *57*(1), e358. 10.1002/pra2.358

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, *42*(4), 2264–2275. 10.1016/j.eswa.2014.10.023

Yang, F., Sun, T., & Zhang, C. (2009). An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization. *Expert Systems with Applications*, *36*(6), 9847–9852. 10.1016/j.eswa.2009.02.003

Zhang, J., Zhai, S., Stevenson, J. A., & Xia, L. (2016). Optimization of the subject directory in a government agriculture department web portal. *Journal of the Association for Information Science and Technology*, *67*(9), 2166–2180. 10.1002/asi.23550

Jaffar Atwan is assistant professor in computer and information science at AL-Balqa Applied University, Jordan, his current job since February 2001. His knowledge and skills are grounded by computer and information science, university teaching and research, technician experience, computer development, and computer center teaching. His research interests are in information retrieval and natural language processing, and his major strengths include teaching, team and program building, and scientific research. He has a PhD in information science and natural language processing from UKM with a demonstrated history of working in the education management industry. He is skilled in search engine technology, data analysis, and programming and has expertise in information retrieval, information science, syntax, morphology, indexing, word sense, disambiguation, ontologies, knowledge representation, and text analysis.

Deema AlSekait is a talented assistant professor with a strong information technology (IT) background. She received her PhD in information technology from Towson University, United States. She has an extensive research portfolio that covers a wide range of topics, from machine learning to health informatics, artificial intelligence, and cloud computing. Additionally, she advocates for improving access to science, technology, engineering, and math careers. She is a shining example of an outstanding IT professional in an ever-changing field, as she perfectly combines her research expertise, pedagogical knowledge, and uncompromising commitment to social advancement. Her tireless efforts continue to inspire innovation, ensuring that the future of information technology remains bright and inclusive.

Hanaa Fathi received her PhD in computer science from the Faculty Of Science, Menufia University, Egypt, in 2022. She is an assistant professor on the Faculty of Information Technology, Applied Science Private University. She has worked on several research topics and has published more than 10 technical papers in feature selection, classification, optimization, machine learning, and web service in international journals. She attended the 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), in December 2019 in Taza, Morocco, and the fourth edition of the International Conference on Intelligent Systems and Computer Vision (ISCV 2020). She majors in machine learning, optimization, and medical application.

Qusay Bsoul received his MS and PhD degrees in information science from the National University of Malaysia. He is an associate professor at Applied Science Private University, Jordan. His research interests are cyber-security/ cyber-crime and optimization problems, wireless sensor networks, text mining, data mining, information extraction, supervised/unsupervised feature selection, and recommendation systems. His disciplines are artificial neural networks, artificial intelligence, unsupervised learning, clustering algorithms, machine learning, classification supervised learning neural networks, artificial intelligence, feature extraction, data clustering, knowledge discovery, cluster analysis, applied artificial intelligence, prediction, high-dimensional data analysis, knn predictive analytics, pattern recognition, and feature selection.

Sharaf Khaled Mahmoud Alzoubi is a distinguished assistant professor with a specialization in mobile computing. He completed his PhD in mobile computing from Northern Malaysia University in 2016 and holds a master's degree in information technology from the same university. Additionally, he earned his bachelor's degree in computer information systems from Jordan University of Science and Technology. He has extensive teaching experience, having served as an assistant professor at Amman Arab University and as a lecturer at Jordan University of Science and Technology. His research interests lie in the areas of mobile computing, smart devices, and virtual reality learning. He has published several research papers in esteemed international journals, including topics on the success factors of mobile learning and the impact of demographic factors on virtual reality learning in developing countries. He is also actively involved in various academic committees and professional associations, contributing significantly to the field of computing and information technology.

Dr. Malik is an industry and academe expert. He started his professional career as analyst programmer in Jordan then later became an IT lecturer in Malaysia. He was also a certified trainer at Canadian Training Center of Human Development and a quality assurance coordinator. He is now an IT lecturer at Gulf College, Sultanate of Oman. He completed his Doctor of Information Science at UKM University, Malaysia. He has wide range of IT skills, project management and research data analysis. His main objective is to secure a challenging position in a reputable organization to expand my learnings, knowledge, and skills and secure a responsible career opportunity to fully utilize my training and skills in addition to make a significant contribution to the success of my institution.my Skills and expertise Research And Development Academic Writing Research Paper Writing Research Methodology Data Collection Research Papers Research Analysis Article Writing Report Writing Research Proposal Writing Qualitative Analysis Quantitative Analysis Team Working Methodology Interviewing Survey Methodology and Data Analysis Data Analysis Quantitative Data Analysis Research Project Management Quantitative Methodology Qualitative Inquiry Power Point Presentation Content Analysis Survey Analysis Project Proposal Writing Research Management Research Coordination Quantitative Methods Mixed Methods Qualitative Methods His passion for research can be seen in his numerous publications and on-going researches. He published papers on mobile marketing, vibratory haptic interface model, e-learning, animation and virtual reality learning. He looks on the topics such as virtual environment, visual informatics, 3D environment, artificial intelligence, internet of things and smart city in a multicultural perspective. Thus, his researches involve different industries and countries.

Abeer Saber was born in Damietta, Dumyat, Egypt, in 1992. She received the BSc degree in computer science and the MSc degree in computer science from Mansoura University, Egypt, in 2013 and 2018, respectively. She received the PhD in computer science from Menoufia University, Egypt, in 2022. She is currently a lecturer in information technology with the Faculty of Computers and Artificial Intelligence, Damietta University, Egypt. She has published many research articles in prestigious international conferences and reputable journals. She is also a reviewer for many journals. Her current research interests include big data analysis, semantic web, linked open data, optimization, machine learning, deep learning, bioinformatics, and IoT.

Diaa Salama AbdElminaam was born in 1982 in Kafr Saqr, Sharqia, Egypt. He received a BSc from the Faculty of Computers and Informatics, Zagazig University, Egypt, in 2004, graduating with honors. He obtained a master's degree in information systems from the Faculty of Computers and Information, Menoufia University, Egypt, in 2009, specializing in cryptography and network security. He obtained his PhD in information systems from the Faculty of Computers and Information, Menoufia University, Egypt 2015. He has been an associate professor in the Information Systems Department, Faculty of Computers and Information, Benha University, Egypt, since 2011. He has worked on several research topics and contributed to more than 120 technical papers in the areas of wireless networks, wireless network security, information security and internet applications, cloud computing, mobile cloud computing, the Internet of Things, and machine learning in international journals, international conferences, local journals, and local conferences. His major interests are cryptography, network security, the IoT, big data, cloud computing, and deep learning.