Classical Vector Space-based Semantic Representation and Dimensionality Reduction for Arabic Text Clustering: a survey

Ebtihal Nooreldeen and Anna M. Zanaboni

College of Computer Science and Information Technology, Sudan University of Science and Technology, Khartoum, Sudan

Received: 07/10/2025
Accepted: 03/02/2026

***Abstract*-- The Arabic language poses many challenges for automatic text processing due to its linguistic complexity, including rich morphology, the use of diacritics, and the frequent omission of vowels in written text. In general, two key factors significantly influence performance in text processing: the incorporation of semantic information in document representation and the control of dimensionality. This paper presents a survey of Arabic document clustering employing classical vector space models for text representation (as opposed to dense embedding models), with a particular focus on semantic representation and dimensionality reduction. The survey explores several applications of clustering, such as text summarization and information extraction, and reviews existing comparisons between clustering algorithms. Although classical vector space methods provide a somewhat limited perspective, the body of work found, covering the period from 2001 to 2025, is quite diverse. Key challenges, major findings, and the Arabic document corpora used in the reviewed literature are concisely presented in comprehensive synoptic tables.**

***Index Terms*-- Arabic text clustering, dimensionality reduction, document representation, literature review, machine learning, unsupervised learning.**

***المستخلص*-- تُشكِّل اللغة العربية العديد من التحديات أمام المعالجة الآلية للنصوص بسبب تعقيدها اللغوي، بما في ذلك غنى بنيتها الصرفية، واستخدام التشكيل، وكثرة حذف الحركات في النصوص المكتوبة. وبوجه عام، يؤثر عاملان رئيسيان تأثيرًا كبيرًا في أداء معالجة النصوص، وهما: دمج المعلومات الدلالية في تمثيل الوثائق، والتحكم في الأبعاد. تقدم هذه الورقة دراسةً مسحيةً حول تجميع الوثائق العربية باستخدام نماذج فضاء المتجهات التقليدية لتمثيل النصوص، وذلك بدلًا من نماذج التضمينات الكثيفة، مع التركيز بصورة خاصة على التمثيل الدلالي وتقليل الأبعاد. كما تستعرض الدراسة عددًا من تطبيقات التجميع، مثل تلخيص النصوص واستخراج المعلومات، وتراجع المقارنات المنشورة بين خوارزميات التجميع المختلفة .وعلى الرغم من أن نماذج فضاء المتجهات التقليدية توفر منظورًا محدودًا نسبيًا مقارنةً بالاتجاهات الحديثة، فإن الأدبيات التي جُمعت، والتي تغطي الفترة من عام 2001 إلى عام 2025، تُظهر تنوعًا ملحوظًا. وتعرض الدراسة بإيجاز أبرز التحديات، وأهم النتائج، ومجموعات بيانات الوثائق العربية (Corpora) المستخدمة في الدراسات التي تمت مراجعتها، وذلك في جداول تلخيصية شاملة.**

I. Introduction

The Arabic language consists of 28 letters, each of which changes shape depending on its position within a word. It also features complex elements such as diacritics, a lack of vowel letters, and no capitalization. Articles, prepositions, and pronouns attach directly to verbs, nouns, and adjectives. A single Arabic word can often correspond to an entire phrase in English, as shown in Table 1.

There is no standardized benchmark for Arabic datasets, which poses challenges for its application in the semantic web. The language requires adherence to specific morphological rules, making it a more complex area of research than Latin languages. One example is the stemming process, which involves returning a word to its lexical root by removing affixes. This process is particularly intricate in Arabic due to the presence of four types of affixes: antefixes, prefixes, suffixes, and postfixes. These complications make research in Arabic even more difficult, especially in tasks such as automatic classification, like categorization and clustering. [1][2][3].

The text clustering process consists of three main phases: *data representation*, in which documents are transformed into an appropriate form that captures their relevant linguistic and possibly semantic features; *similarity evaluation*, which involves the definition and computation of suitable similarity or distance measures between document representations; and *grouping*.[4][5][6]In which documents are organised into coherent clusters using specific algorithms based on the results of the previous two phases. The first two phases depend on the type of data and the characteristics of the corpus. [5][7] Careful selection of appropriate representation and similarity methods is crucial for effective clustering.

TABLE 1
Arabic Words and their English Equivalents

Arabic Word	English Translation
ليحدثونهم	to speak with them
يستعدون	to get ready for
يلتقون	to meet with someone
يناقشون	to discuss with others
يعتنون	to take care of

Documents can be represented using a set of index terms or keywords. This is known as the logical view of the document, in contrast to the full-text view, which includes all words in the document. Full-text

representation in large collections leads to high dimensionality, which can result in high computational costs, increased storage requirements, and degraded clustering quality. The logical view is preferable as it reduces dimensionality by removing noisy words [7][8]. In the context of data representation, the emphasis is on identifying the key characteristics that effectively describe the content. Recognizing the relationships between words and extracting their semantic meanings are critical for successful text clustering[9][9][10][11]. Additionally, it is important to maintain a low-dimensional representation while preserving essential information. As stated in [12], the representation models are divided into two main categories, namely Classical Representation Models and Distributed Representation Models, as shown in Fig. 1.

Classical Representation Models

According to these models, terms constituting a document are treated as independent and discrete entities, and the representation of a term is based on its frequency or presence in the document. A commonly used method belonging to this category is the classical Vector Space Model (see Section II), in which a document is represented as a vector in a continuous vector space where documents with similar content occupy nearby positions. One of the earliest examples of a vector space model is the bag-of-words representation, in which each dimension corresponds to a word in the vocabulary. Classical representation models suffer from several well-known limitations: they are expensive in terms of dimensionality, they don't preserve the order of terms in the document, and they do not consider relationships between terms. [13].

Distributed Representation Models

These models overcome the challenges of high dimensionality and show the ability to capture semantic information. They are further divided into continuous word representations and contextual word representations, as follows.

Continuous Word Representations: These are methods for creating word embeddings by mapping words into dense, real-valued, low-dimensional vector spaces learned from large text corpora, typically using neural network-based architectures. These representations capture both semantic and syntactic regularities by exploiting word co-occurrence patterns, such that words appearing in similar contexts are associated with similar vectors. As a result, continuous word representations enable the modelling of linguistic relationships such as similarity and analogy. [13]. They constitute the foundation of widely used approaches like Word2Vec, GloVe, and FastText. In some studies, [14]Arabic documents are represented as AraVec. [15] Word embedding.

Contextual Word Representations: these are techniques that exploit multi-layer neural networks to compute dynamic representations of words depending on the context in which they appear. Embeddings from Language Model (ELMo) is an example. Some research applied this model to Arabic, such as AraBERT [14], QARiB[17], and AraGPT2[16].

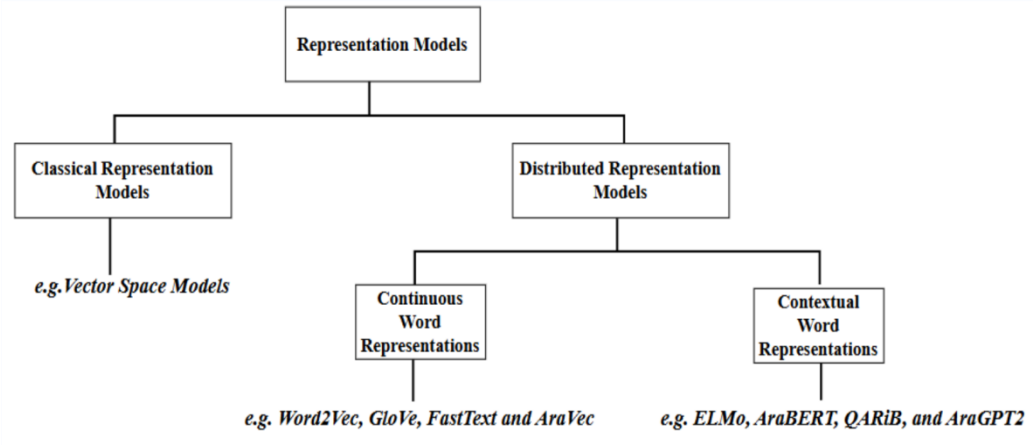

Fig. 1. Classification of Representation Models as in [12] and [19]

This taxonomy is compatible with the categorization proposed in [19]. Another categorization is presented in [5] and shown in Fig. 2, identifying the three most used document representation methods, specifically Vector Space Models, Probabilistic Topic Models and Statistical Language Models.

Vector Space Models

In these models (see Section II), a document is represented as a vector in a continuous space, where similarity between documents is captured by the geometry of their representations.

Probabilistic Topic Models

In these models, documents are represented as a mixture of topics, where a topic is a probability distribution over words or terms. For example, the Latent Dirichlet Allocation is a Bayesian network, where topics are considered as latent variables inferred from observed data, which are documents and words.

Statistical Language Models

According to these models, documents are represented through probability distributions over

sequences of words. For example, *n-gram-based* models assume that the probability of a word depends on the *n* words preceding it in the text.

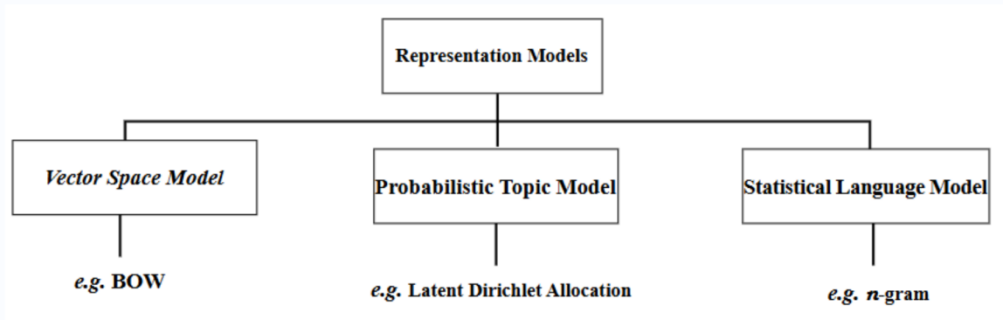

Fig. 2. Classification of Representation Models as in [5]

Document representations are used to calculate the semantic similarities between documents. Semantic similarity can be analysed at three levels: word level, sentence level, and document level. Methods for measuring semantic similarity can be categorized into four main groups. [20][21], as follows.

Co-occurrence-based methods: these methods use the bag-of-words model for text representation, measuring similarity through techniques like Cosine similarity.

Statistical corpus-based methods: these methods represent words based on their context within a corpus, capturing meaning from their distribution patterns.

Lexical database-based methods: these methods rely on resources like WordNet to determine semantic relationships between words.

Word embedding-based methods: these methods derive meaning from the context of words in large datasets, utilizing techniques such as Word2Vec, GloVe, or FastText.

In the present work, a characterisation of Vector Space Models is provided in Section. II and a review of the state of the art in text clustering for Arabic is presented in Section III, focusing on three main aspects: issues in clustering Arabic text, such as preprocessing techniques, algorithm comparison, domains of application, and topic modelling; the use of dimensionality reduction techniques for high-dimensional data; the employment of vector space models and semantically enriched representations. Conclusive remarks are provided in the section. IV.

II. Vector Space Models

The vector space–based approach is a framework for representing items as vectors in a shared geometric space, where similarity between items is reflected by geometric closeness in that space. Given this definition, vector space-based models differ in how the vector space is constructed and, consequently, how semantic information is encoded, namely explicitly in sparse dimensions (sparse vector representations), implicitly through latent projections (low-dimensional latent representations), or learned directly as dense embeddings. Fig. 3 shows a schema of this categorisation.

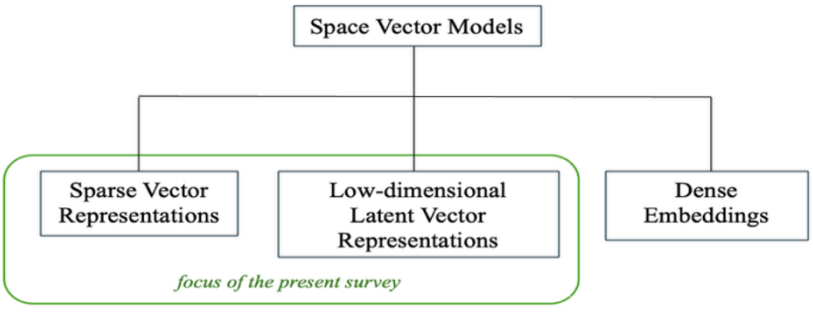

Fig. 3. Classification of Space Vector Model

Sparse Vector Representations

Sparse vector representations are based on the frequency of terms in the document. The vector components are usually the weights of the terms found in the document. A term can be a single word or a phrase, and the dimensionality of the representation is the number of different terms occurring in the corpus of documents. Similarity between documents is typically measured using cosine similarity or other distance metrics.

Bag-of-Words (BoW) [22] is a canonical and widespread model in which documents are represented as high-dimensional vectors over a shared vocabulary. Each vector component encodes the frequency of a word in the document. Term weighting schemes, such as term

frequency–inverse document frequency (tf-idf), improve discriminative power by down-weighting common words.

BoW and tf-idf are simple and interpretable; however, they fail to capture semantic similarity between terms not explicitly shared, and vectors representing documents tend to be high-dimensional and sparse.

Prior studies [22][23]show that tf-idf vectors achieve competitive retrieval performance, particularly in domains with small or domain-specific corpora, but their effectiveness decreases in tasks requiring semantic generalization.

Low-dimensional Latent Vector Representations

Latent semantic models derive vector representations by applying dimensionality reduction techniques to sparse co-occurrence matrices. In these models, the original explicit feature space is projected onto a lower-dimensional latent space, where dimensions no longer correspond to individual terms but capture underlying semantic factors inferred from global co-occurrence patterns.

Latent Semantic Analysis [24] extends BoW representations by performing dimensionality reduction (usually through singular value decomposition) on the term-document matrix. The resulting vectors representing documents are low-dimensional and capture latent semantic structure.

Prior studies [25] show that Latent Semantic Analysis outperforms BoW in tasks requiring semantic similarity, such as information retrieval and document clustering, particularly when synonymy and polysemy are present.

Dense Embeddings

Dense embeddings represent documents as low-dimensional vectors learned directly through predictive or optimization-based objectives. The dimensions of these vectors are latent and non-interpretable, and semantic relationships are encoded implicitly through the geometry of the learned space rather than through explicit feature dimensions or matrix factorization.

Predictive neural models, such as Word2Vec [26][27] and GloVe [28], represent words as dense vectors in a low-dimensional continuous space. These vectors are learned so that words appearing in similar contexts occupy nearby positions. Similarity is measured through cosine similarity or dot product, and linear relationships between vectors capture meaningful semantic analogies.

Prior studies [28] show that embeddings consistently outperform sparse models on word similarity, analogy, and semantic clustering tasks. However, in [29] and [30]cases are reported where sparse vectors and latent methods remain competitive depending on text length and task demands.

III. Literature Review

For the review, 53 papers were considered. Inclusion criteria were the publication date between 2001 and 2025 and the use of sparse or low-dimensional latent document representation (as described in the Section II). The distribution over the years of the papers mentioned in this work is shown in Fig. 4.

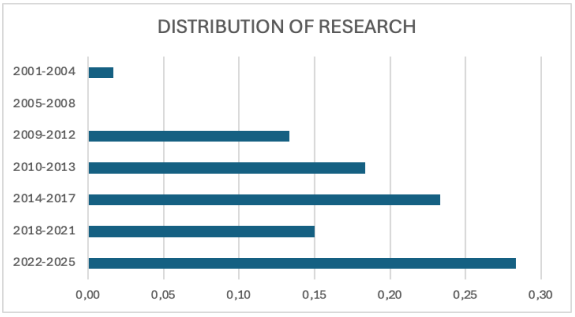

Fig. 4. Distribution over the years of the papers mentioned in this survey

Although each considered work faces different aspects of Arabic document processing, we categorized the reviewed research into three main areas: clustering in Arabic texts, traditional techniques for addressing high dimensionality, and semantic representations of documents. Certain studies address multiple categories simultaneously.

A. Issues in Clustering Arabic Text

Research on Arabic text clustering can be examined from multiple perspectives. Among others, we focused on the following: the preprocessing techniques applied prior to clustering, the application of clustering across different tasks, the range of clustering algorithms employed, the result of comparative evaluations of these algorithms, the use of topic modelling as an alternative or complementary approach to clustering, and methods for addressing the problem of initializing cluster centroids.

1) Preprocessing

Research on clustering in the Arabic language is less extensive compared to similar studies in other languages. Much of the existing research has concentrated on the morphological aspects of Arabic, specifically by examining preprocessing techniques for Arabic documents and their effects on information retrieval and the clustering of these documents.

The preprocessing techniques include tokenization, removal of stop-words, normalization, stemming, vectorization, feature selection, term pruning, lemmatization, handling of synonyms, and semantic processing. [1][3] [31][33][34]Error! Reference source not found.[35][36].

Applying stemming speeds up the clustering process and reduces the size of documents. However, the effect of stemming on clustering quality varies: in some cases. [3][32][34][35]it has a negative impact, while in other cases [1], in some instances,[37][36][37], it has a positive effect. Summary of the mentioned research is shown in Table 2.

2) Applications

Clustering is widely used to enhance tasks such as summarization [38][39], information retrieval [40], and web search results [41][42], and has also been applied to domain-specific Arabic texts, including crime [43], Islamic documents[37][37], hate speech tweets [44] and Quran text as classical Arabic [45].

Summary of these studies in different tasks is shown in Table 3.

Hierarchical clustering typically yields more effective results compared to flat clustering, like K-means. However, flat clustering algorithms have linear complexity, while hierarchical clustering algorithms exhibit quadratic complexity. To exploit the advantages of both, hybrid clustering methods, combining K-means and hierarchical clustering, are used. In [38], hybrid clustering is used for summarization purposes. From each cluster, key phrases are extracted to identify important sentences, which are then used to represent the documents as their summaries. Fuzzy C-Means and Latent Dirichlet Allocation are employed for multi-document summarization in [39].

TABLE 2
Stemming Effect on Clustering

Research	Impact	Algorithm	Measure	Distance	Dataset
[1]	Positive and Negative	-	-	-	-
[3]	Negative	K-means Latent Diriclet Allocation	Rand index, Jaccard, F-measure, Entropy	-	CCA, Alwatan newspaper, OSAC
[32]	Negative	K-means	Accuracy	-	1445 documents
[34]	Negative	K-means	Purity, Entropy	Euclidean, Cosine, Jaccard, Pearson, Kullback-Leibler, Divergence	CCA
[35]	Negative	Agglomerative Hierarchical	Purity, Entropy	Euclidean, Cosine, Jaccard Pearson	CCA
[33]	Positive	K-means	Purity, F-measure	Cosine, Jaccard, Pearson, Euclidean, Kullback-Leibler, Divergence	1680 documents
[36]	Positive	K-means	Recall, Precision, F-measure	-	BBC
[37]	Positive	K-means	F-measure	Euclidean, Cosine, Jaccard	In-house dataset (Islamic documents)

First, the TAC-2011 dataset is clustered after reducing its high dimensionality using Singular Value Decomposition. Next, Latent Dirichlet Allocation is applied to the resulting clusters to extract latent topics, which are then used to generate summaries by selecting the top-ranked sentences associated with each topic. The proposed method is evaluated using precision, recall, and F-measure, and the results demonstrate promising performance compared with existing studies.

The study [40] explores the use of clustering in information retrieval. Two types of clustering algorithms were employed to cluster the retrieved documents, namely K-means as flat clustering, and Ward and group-average as agglomerative clustering. The dataset consisted of 1000 documents from the Al-Sabaah newspaper, represented using the Vector Space Model with tf-idf weighting.

The performance of information retrieval without clustering was measured using precision, recall, and F-measure, while the clustering of returned documents was assessed using a new method based on accuracy. The average accuracy for K-means and Ward was 0.50, with the average agglomerative clustering performing the worst. For clustering web search results, authors in [42] proposed an approach based on Formal Concept Analysis. Fig. 5 illustrates an example of the resulting clusters, where each rectangle represents a cluster containing similar documents or objects (denoted as G1 through G8), grouped based on shared terms among the documents.

Three approaches to clustering were employed in [46]for the classification of Arabic texts: unsupervised, semi-supervised, and semi- supervised with dimensionality reduction. Documents were initially pre-processed by removing stop words, stemming words to their roots, applying tf-idf weighting, and calculating document similarity. In the unsupervised approach, clustering was performed by K-means and incremental K-means. The semi-supervised approach required some labelled data to improve the creation of initial centroids through a threshold function. The clustering results were then applied to a classification model, which was tested using a labelled test file, and dimensionality reduction was applied to the centroids to accelerate the classification process. A proposed score function is used to weigh each term in the class; dimensionality reduction is achieved by selecting a ratio of reduction from the top (highest scores) and bottom (lowest scores). Online dynamic datasets were used, with evaluation metrics including F-measure and entropy. The best F-measure achieved was 0.86, with a dimensionality reduction of 5% from the top and 30% from the bottom, with 5% of the training data labelled.

In [41] the Enhanced K-Means algorithm is introduced for clustering search results. The resulting clusters are labelled with the most significant terms within each cluster, using tf-idf as a term-weighting method. Enhanced K-Means improved execution time compared to standard K-Means by reducing the number of distance calculations through the reuse of previously computed distances. However, the differences in purity and entropy

between the two algorithms were minimal. The dataset used in this study consisted of 6,500 documents generated from searches of 50 terms, with three clusters created for each term.

Arabic clustering has been applied to specific domains, for example, to crime, like in [43], where the type of crime is determined for each document. Documents are represented by keywords extracted from them, and clustering is done by Self-Organizing Maps. Hate speech tweets from the L-HSAB dataset were clustered using the K-means algorithm in [44], where tf-idf was employed for text representation and PCA for dimensionality reduction. However, the study does not report any clustering evaluation metrics or corresponding performance values.

TABLE 3
Applications of Clustering

FCM: Fuzzy C-Means, LDA: Latent Dirichlet Allocation, FCA: Formal Concept Analysis, NMI: Normalized Mutual Information, NCE: Normalized Complementary Entropy, SOM: Self-Organizing Maps, UMAP: Uniform Manifold Approximation and Projection, SVD: Singular Value Decomposition

Research	Finding	Algorithm	Measure	Distance	Dataset
[38]	Summarization of documents using a keyphrase extraction module from clusters	K-means hierarchical clustering	Accuracy	Cosine Jaccard	Essex Arabic Summaries Corpus
[39]	Summarization of documents by FCM and LDA	FCM	Precision Recall F-measure		TAC-2011
[41]	Enhanced K-means improved execution time compared to K-means	enhanced K-means	Purity Entropy	-	6500 documents
[42]	FCA was integrated in web search results	FCA	NMI NCE	-	ODP
[40]	K-means with Ward distance better than average agglomerative clustering	K-means Ward's Average agglomerative	Recall Precision F-measure Accuracy	-	1,000 news documents
[43]	The type of crime extracted from documents	SOM	-	-	50 news documents
[37]	The best performance is achieved by K-means and Cosine	K-means	F-measure	Cosine Euclidean Jaccard	Islamic documents
[46]	The classifier produced better results.	K-means Incremental K-means	F-measure Entropy	Cosine	Online dynamic
[47]	Dimensionality reduction and semantic extraction from Classical Arabic based on tf-idf with SVD and UMAP	K-means	Silhouette Score, Calinski-Harabasz Index Davies-Bouldin Index		All verses from Surah Al-Baqarah
[44]	Clustering hate speech tweets.	K-means	-	-	L-HSAB

In [37] Islamic documents, written both in Modern Standard Arabic and Classical Arabic, are clustered via K-means. They are represented using the bag-of-words model, and similarity/distances are calculated using three different metrics, namely Cosine similarity, Euclidean distance, and Jaccard distance. Results show that clustering with stemming outperforms clustering without stemming. The best results are obtained using the Cosine distance metric and two clusters, achieving an F-measure of 0.85. The study is conducted on an in-house dataset. In [47]Classical Arabic texts are represented using tf-idf and clustered by K-means. In this research, dimensionality reduction is obtained by Truncated Singular Value Decomposition along with Uniform Manifold Approximation and Projection. A 10-cluster solution for all the verses of Surah Al-Baqarah was identified as the optimal one and was interpreted by examining the top tf-idf terms in each cluster so that a different theme was assigned to each cluster.

3) Algorithm comparisons

Some research (as shown in Table 4) investigates algorithm comparisons, such as Bisect K-means versus K-means [48], K-means versus K-medoids [49], K-means hierarchical agglomerative clustering, and fuzzy c-means[14], and Latent Dirichlet Allocation versus K-means[3][50][51].

In [48] a hybrid algorithm (bisect K-Means) is compared to the standard K-means algorithm for the analysis of Arabic documents. Five similarity/distance measures are explored: Pearson correlation coefficient, Cosine similarity, Jaccard distance, Euclidean distance, and averaged Kullback-Leibler divergence.

A comparison between the K-means and K-medoids algorithms to enhance the clustering of Arabic documents for better information retrieval is found in [49]. The study utilized an inverted file approach for document indexing and representation, using a dataset of 242 documents. Results showed that K-medoids outperformed K-means in both recall and precision. Similarly, the study in [14] presented a comparative study of three clustering algorithms for Arabic text: K-means, Hierarchical Agglomerative Clustering, and Fuzzy C-Means. The comparison was based on six criteria: dataset size, number of clusters, stability, complexity, priority, and noisy data. The findings of this comparison are summarized in

Table 5. K-means and Fuzzy C-Means outperform Hierarchical Agglomerative Clustering in large datasets. Furthermore, K-means is the best option when considering the complexity of the algorithm and the large number of

clusters. In contrast, Hierarchical Agglomerative Clustering excels in terms of stability and its ability to handle noisy data. Each algorithm needs prior knowledge to function effectively: minimal prior knowledge includes the set of documents to cluster, which occurs in Hierarchical Agglomerative Clustering; K-means also requires K; and Fuzzy C-Means has more knowledge than both K-means and Hierarchical Agglomerative Clustering.

A dataset comprising 3,404 Arabic comments was used. The grouping of comments into clusters was similar across all three clustering algorithms.

An example of using topic modelling for clustering is presented in [3][50][51], using the Latent Dirichlet Allocation and the K- means algorithms on Arabic texts. In [3] the effectiveness of the two algorithms was assessed through a comparison on three versions of an Arabic corpus: raw data, cleaned data without stop words, and stemmed data. Results show that Latent Dirichlet Allocation Outperformed K-means in clustering quality. Additionally, removing stop words improved the results, while stemming had a negative effect.

In [50] and [51], K-Means and Latent Dirichlet Allocation were combined in two steps: in the first step, Latent Dirichlet Allocation was applied to uncover topics within the documents, and in the second step, K-Means was used for clustering. The documents were pre-processed and represented using normalized tf-idf. In [50], the dataset used contained 2,700 Modern Standard Arabic news documents across 9 categories. K- Means was run 25 times with different initial centroids, and the average results were taken to address the issue of poor starting centroids. Compared to previous studies on the same dataset, this method outperformed others, yielding better results than K-Means alone. The evaluation was based on various metrics, including purity, F-measure, entropy, and accuracy. Similarly, the study in [51] used Arabic security news with 4004 observations. The combination of Latent Dirichlet Allocation and K-means outperforms K-means alone in different evaluation metrics: purity, within-cluster sum of squares, Dunn index, normalized mutual information, normalized variation of information, and Rand index.

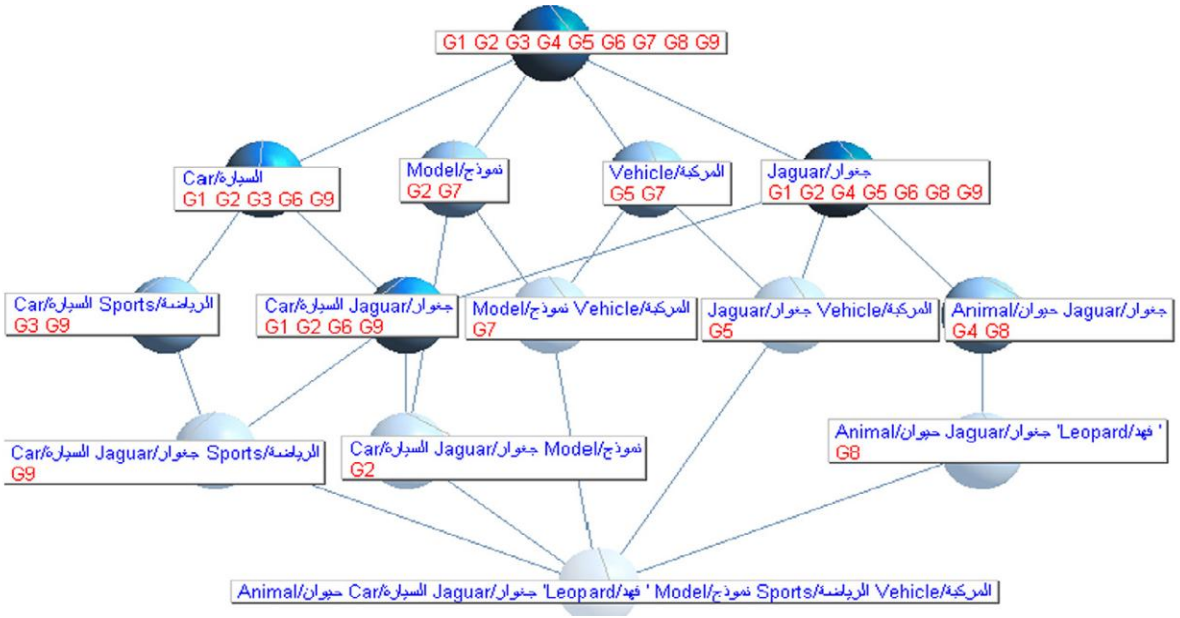

Fig. 5. Example of a concept lattice[42]

4) Choice of initial centroids

Optimization methods are combined with clustering to enhance the problem of initial centroids in K-means. [52][53]. A summary of those methods is shown in Table 6. Particle Swarm Optimization is employed in [52] to address the challenge of selecting initial centroids in the K-means algorithm. In this approach, each particle represents a clustering solution with randomly chosen centroids. The fitness of each particle is assessed using a function that minimizes the Sum of Squared Errors, helping to identify the best global and local solutions. Three datasets are used: BBC, CNN, and OSAC. The hybrid approach combining K-means with Particle Swarm Optimization demonstrates superior performance compared to standard K-means, particularly in terms of recall, precision, and F-measure. Authors in [53] integrate the Water Cycle Optimization algorithm with K-means to address the sensitivity of K-means to the selection of initial centroids. The proposed hybrid clustering approach demonstrates superior performance compared to the standalone Water Cycle Optimization algorithm and other clustering methods. Experiments are conducted using the TREC 2001 dataset and an Arabic news dataset, with cosine similarity employed as the similarity measure and the F-measure used for performance evaluation.

B. VSM-based techniques for addressing high dimensionality

HIGH DIMENSIONALITY IS HANDLED IN THREE MAIN WAYS: BY ENHANCING DOCUMENT REPRESENTATION, AS ILLUSTRATED IN Table 7(*latent vectors*)and TABLE 8(*sparse vectors*)*;* or by improving clustering efficiency in high-dimensional spaces,

as shown in TABLE 9; or by addressing both objectives simultaneously, as demonstrated in [57]. In [34], documents were summarized using Latent Semantic Analysis and represented as bag-of-words based on the summaries. Clustering accuracy was assessed by applying the K-means algorithm both with and without stemming. The results showed a purity score of 0.32 when stemming was used, compared to 0.77 for full representation. This study indicates that stemming has a negative effect on clustering performance. In [58], a modification of the bag-of-words method was proposed to reduce high dimensionality by incorporating semantics. This approach groups similar words together, achieving a reduction in dimensionality by 40%. However, the proposed method has not yet been evaluated for clustering.

In [59]documents were represented through keyphrases, and agglomerative hierarchical clustering was applied. The authors proposed a similar measure based on a common lemma form of keyphrases between the documents. This keyphrase-based representation achieved better results for clustering compared to the full text-based representation.

In [60]the quality of clustering was enhanced by employing agglomerative hierarchical clustering and representing documents using a suffix tree to extract key phrases instead of all the words in the documents.

TABLE 4
Algorithms Comparison

FCM: Fuzzy C-Means, LDA: Latent Dirichlet Allocation, HAC: Hierarchical Agglomerative Clustering, NMI: Normalised Mutual Information, NVI: Normalised Variation of Information, RI: Rand Index

Research	Finding	Algorithm	Measure	Distance	Dataset
[48]	Bisect K-means outperformed K-means	Bisect K-means K-means	Purity	Pearson Cosine Jaccard Euclidean Kullback-Leibler divergence	300 news documents
[49]	K-medoids outperformed K-means	K-medoids K-means	Recall Precision	Proposed one	242 documents
[14]	Comparison based on 6 criteria	K-means HAC FCM	-	-	3404 Arabic comments
[3]	LDA outperformed K-means	K-means LDA	Rand index Jaccard F-measure Entropy	-	CCA Alwatan newspaper OSAC
[50]	LDA with K-means outperformed K-means	K-means LDA	F-measure Entropy Accuracy	-	MSA
[51]	LDA with K-means outperformed K-means	LDA K-means	Purity Within-Cluster-Sum-of-Squares Dunn index NMI NVI RI	-	Security news

TABLE 5
Comparison of Clustering Algorithms in [14]
k=number of clusters, t=number of iterations, n=number of documents, d=number of dimensions

Criteria	K-Means	Hierarchical Agglomerative Clustering	Fuzzy C-Means
Quality for large data size	Good	Bad	Good
Performance for large k	Excellent	Average	Average
Stability	No	Yes	No
Complexity	$O(tkn)$	$O(n^3)$	$O(ndk^2t)$
A-priori knowledge	The set of documents The number of clusters	The set of documents	The set of documents The number of clusters The threshold The degree of blurring
Noisy data	No	Yes	No

TABLE 6
Optimization for Initial Centroids
PSO: Particle Swarm Optimization, WCO: Water Cycle Optimization

Research	Finding	Algorithm	Measure	Distance	Dataset
[52]	K-means with PSO outperformed K-means	K-means PSO	F-measure	-	BBC, CNN OSAC
[53]	K-means with WCO outperformed K-means and WCO	K-means WCO	F-measure	Cosine	TREC Arabic news

TABLE 7
Representation in High Dimensionality (Latent vectors)

PLSA: Probabilistic Latent Semantic Analysis, LSI: Latent Semantic Indexing, EM: Expectation Maximization, SOM: Self-Organising Maps

Research	Finding	Algorithm	Measure	Distance	Dataset
[34]	Dimensionality reduction done by summarization	K-means	Purity Entropy	Euclidean Cosine Jaccard Pearson Kullback-Leibler Divergence	CCA
[54]	The topic model outperforms PLSA.	-	Purity MICD DBI	-	News
[55]	Latent Semantic Indexing with EM outperformed SOM and K-means	EM SOM K-means	Accuracy	-	Alanba newspaper
[56]	LSI with cosine improves accuracy by 15%	EM	Accuracy	-	In-house (1000 documents)

The method was utilized on the Corpus of Contemporary Arabic (CCA), which includes 278 documents. Data mining techniques were employed to reduce high dimensionality, as in [61][62][63]. A novel approach for clustering Arabic documents is presented in [62] .The method employs the FPMax algorithm, a data mining technique used to extract Maximal Frequent Wordsets, which capture recurring word combinations within the dataset. Rather than representing documents using individual terms, the approach relies on word combinations, thereby significantly reducing the high dimensionality of the feature space. These word combinations are subsequently used in the clustering process by calculating their similarity with the documents to form clusters. The method was applied to the CNN and OSAC datasets, achieving an F-measure of 0.80. In [61]the challenge of high dimensionality is tackled by utilizing frequent item sets, a data mining technique for defining clusters. Hierarchical clustering is then used to organize documents based on their relevance to specific topics or clusters, facilitating easier browsing. Each cluster or topic is represented by shared words (frequent item sets). The methodology also incorporates the application of tf-idf following the stemming process. The proposed approach was compared to Frequent Itemset Hierarchical Clustering, bisecting K-means, and Frequent Term-based Text Clustering on multiple datasets. The study in [63] employs the Frequent Itemset Hierarchical Clustering method proposed in [61] and integrates N-gram analysis for the Arabic language. It achieves an F-measure of 0.7, outperforming the score observed by 0.6 for European languages.

Traditional dimensionality reduction techniques, namely Principal Component Analysis, Nonnegative Matrix Factorization, and Singular Value Decomposition, were used in [64]. Documents were represented using tf-idf, and the K-means clustering algorithm was applied. Principal Component Analysis produced the best results, particularly in terms of accuracy and normalized mutual information. The dataset used included 7,232 Arabic news documents across 10 manually curated topics, as well as an English corpus from Reuters consisting of 7,293 documents covering 10 topics.

The topic model is an alternative approach to address the issues associated with the bag-of-words; Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation are examples of this approach.[11]. The topic model is used in document representation to reduce dimensionality. [54]. Instead of representing a document by words or terms, it is represented as a set of topics; and this method can be directly used for document clustering. In [55], Arabic documents are categorized using a latent semantic indexing representation based on Singular Value Decomposition in conjunction with clustering techniques. This approach is particularly effective for labelling data in the absence of training samples. Expectation-Maximization, Self-Organizing Maps, and K-means are employed as clustering algorithms.

The study in [56] proposed an automatic text clustering approach based on Latent Semantic Indexing and the cosine similarity measure. Latent Semantic Indexing combined with singular value decomposition was applied twice to address high dimensionality: first, documents were represented in the Latent Semantic Indexing space and pairwise similarities were computed using cosine similarity; second, Latent Semantic Indexing was reapplied prior to clustering, which was then performed using the Expectation-Maximization algorithm. Experiments were conducted on an in-house corpus consisting of 1,000 documents, and the proposed

approach achieved a 15% improvement in accuracy compared to using Latent Semantic Indexing alone.

Feature selection methods such as Principal Component Analysis, Singular Value Decomposition, Information Gain, Chi-square, and Relief are commonly employed to reduce high-dimensional feature spaces. In [65], the authors proposed a method called OCATC, in which documents are represented using tf-idf. High dimensionality is addressed by selecting an optimal feature selection strategy, where Particle Swarm Optimization is used to determine the optimal solution for Arabic document classification. This optimal solution includes the number of selected features, the feature selection method, and the classifier.

The study in [65] compared microword (character bi-gram) representation with full-word representation for clustering Arabic documents. Principal Component Analysis was applied to reduce dimensionality, and experiments were conducted on a dataset comprising 250 documents. The results showed that both representation methods achieved comparable accuracy.

Paper [53] introduces a method for term extraction based on verbs and nouns. The method utilizes the Educated Text Stemmer and Arabic WordNet and achieves dimensionality reduction through the proposed extraction process. The extracted terms are evaluated individually and in conjunction with clustering methods. Clustering performance, measured using the F-measure, is compared across different term extraction techniques and the bag-of-words representation, with the combination of Arabic WordNet and Educated Text Stemmer yielding the best results.

Some studies have proposed methods for detection in text corpora, such as [67][68]. These methods can potentially be utilized to enhance Arabic text clustering by representing documents using extracted keywords. In [67], the authors introduced the Safar Keyword Extraction Tool for automatic keyword extraction. The final keyword list is constructed by integrating three lists generated across three phases. In the first phase, a frequency-based list is produced from the top 50 most frequent tokens in the corpus. The second phase generates a vector space–based list derived from the most frequent tokens appearing across all paragraphs in the corpus. In the third phase, logistic regression is employed to construct a third list representing domain-specific expertise.

Traditional clustering algorithms such as K-means often suffer from convergence to local optima and exhibit a low convergence rate, particularly when dealing with high-dimensional data. To address these limitations, a hybrid optimization approach, namely GWOCS-mini[69], which integrates Grey Wolf Optimization and Cuckoo Search, is proposed and combined with minibatch K-means. To mitigate the challenges associated with high dimensionality, Principal Component Analysis is applied. Additionally, the well-known initial centroid sensitivity problem of K-means is alleviated by employing the minibatch K-means strategy, which provides more stable and reliable initial centroids. Experimental results demonstrate that the proposed approach outperforms several metaheuristic methods, traditional clustering algorithms, and existing studies on English text datasets, as measured by the silhouette coefficient, Adjusted Rand Index, and lowest Root Mean Square Error. However, the runtime performance of the proposed method is higher compared to some other approaches, indicating a trade-off between clustering accuracy and computational efficiency. The effectiveness of the proposed method is validated using multiple benchmark datasets, including NADA, CNN, ARPD, and BBC datasets. In [70], traditional data representations were used, specifically CountVectorizer and tf-idf, and the performance limitations of K- means in high-dimensional data were overcome by an approach combining the Grey Wolf Optimization algorithm with Self-Organizing Maps. A significant improvement over conventional Self-Organizing Maps and K-means methods was achieved. The datasets used were MSA and NADA.

A method for clustering Arabic documents based on rough set theory is proposed in [71]. Document preprocessing and representation are performed using tf-idf or the Count Vector model, followed by the extraction of a feature table. Weighted distances between documents are then computed according to rough set theory and used as input for the clustering process. The method is evaluated on the CNN and OSAC datasets, achieving an F-measure of 0.85.

In [57] document representation was based on the categories documents belong to, rather than on the individual words they are composed of. A hybrid approach that integrates Bayesian-Vectoriser and K-means was introduced, and it achieved an 85% improvement in clustering performance and a 95% reduction in runtime.

TABLE 8
Representation in High Dimensionality (sparse vectors)

HAC: Hierarchical Agglomerative Clustering, MFW: Maximal Frequent Wordsets, FIHC: Frequent-Itemset-based Hierarchical Clustering, NMI: Normalized Mutual Information, PCA: Principal Component Analysis, SVD: Singular Value Decomposition, PSO: Particle Swarm Optimization, WCO: Water Cycle Optimization, NMF: Nonnegative Matrix Factorisation

Research	Finding	Algorithm	Measure	Distance	Dataset
[58]	The semantic vector space model resulted in a 40% reduction in dimensionality.	-	Dimension Reduction Ratio	-	BBC's
[59]	Keyphrase-based representation better than full text–based representation	Hierarchical Clustering	Purity Entropy	Euclidean Cosine Jaccard Dice	345 news Arabic documents

Research	Finding	Algorithm	Measure	Distance	Dataset
				Overlap Pearson Proposed	
[60]	Suffix tree representation improved clustering.	HAC	Purity Entropy	Euclidean Cosine Jaccard Pearson Kullback-Leibler Divergence	CCA
[62]	MFW-based representation returned promising results compared to other research.	Proposed algorithm based on MFW	F-measure Precision Recall	Euclidean Cosine Jaccard Overlap Manhattan	CNN OSAC
[61]	Frequent Itemsets representation returned promising results compared to other research.	FIHC	F-measure	Proposed one	Classic4 Hitech Wap Reuters Re0
[63]	Frequent Itemsets with n-grams returned good results for Arabic compared to other languages.	FIHC	F-measure	Proposed one	Built-in house
[64]	PCA produced the best results compared to NMF and SVD	K-means	Accuracy NMI	-	Arabic news documents Reuters-English dataset
[66]	Microword (character bi-gram) representation and full-word representation achieved comparable accuracy.	K-means	Accuracy	-	250 documents
[65]	PSO optimizes high-dimensionality	PSO	F-measure	-	CNN BBC
[53]	Clustering performance is good for the proposed term extraction(Arabic WordNet+Educated Text Stemmer)	K-means WCO	F-measure	Cosine	TREC Arabic news

C. Semantic Representation

Sometimes tools like WordNet are used to calculate semantic similarity between words, addressing polysemy and synonymy. [72]. For example, in [73]a comparative study was conducted. Three representation modes (bag-of-words, n-gram, concept-based) were compared for classification problems. The concept-based representation relies on external knowledge from WordNet to define synonyms but does not consider polysemy, where the same word can have different meanings depending on the context within the text.

The Resource Description Framework offers semantic representation for text documents. In [73] Arabic documents are represented by integrating Resource Description Framework and DBpedia to facilitate easy access and processing. Resource Description Framework triples (subject, predicate, and object) are extracted using predefined rules. Named Entity Recognition (via GATE [75]) is employed to identify entities, which are then linked to DBpedia URLs by searching for the entities on Wikipedia. A dataset comprising 60 different Wikipedia articles is utilized, resulting in 250 sentences, 2,219 words, 160 entities, 108 triples, and 145 URLs. However, this representation method is not applied for information retrieval or clustering.

A model for document representation in the clustering of Arabic web pages is proposed in [76]. The process begins by extracting semantic class features, where verbs within the documents are identified and assigned to semantic classes based on VerbNet. Next, the documents are vectorized by representing them as vectors of semantic classes instead of individual terms, using either semantic class density or probability distribution for weighting. Finally, the documents are clustered, and the clusters are annotated with the most prominent topics. A dataset of 753 documents from online Arabic newspapers is used for the model evaluation. This method outperforms the standard K-means algorithm in terms of purity, mean intra-cluster distance, and Davies–Bouldin index.

Some studies address both high dimensionality and semantic representation issues, such as [77], which proposes a hybrid document representation that combines DBMS-document distances derived from tf-idf features with word embeddings (FastText). This hybrid representation is applied to classification tasks to improve efficiency by reducing the dimensionality and sparsity associated with tf-idf, while simultaneously incorporating semantic information through word embeddings. The study evaluates the proposed approach using seven news-based datasets.

TABLE 10 shows a summary of related work in semantic

representation.

TABLE 9
Clustering Efficiency in High Dimensionality

GWOCS: Grey Wolf Optimization and Cuckoo Search, GWO: Grey Wolf Optimization, SOM: Self-Organizing Maps

Research	Finding	Algorithm	Measure	Distance	Dataset
[69]	GWOCS solves local optimal and initial centroid effectively	minibatch K-means	Silhouette coefficient, Adjusted Rand Index	Cosine	NADA CNN ARPD BBC
[70]	Integration of GWO with SOM outperformed SOM and K-means in high-dimensional data.	GWO SOM	F-measure Precision Recall Accuracy	-	MSA NADA
[71]	New method for clustering	Rough set theory	F-measure	Weighted distances	CNN OSAC

TABLE 10
Related work in semantic representation

BOW: Bag Of Words, RDF: Resource Description Framework, MICD: More Intra-Class Diversity, DBI: Davies-Bouldin Index

Research	Finding	Algorithm	Measure	Distance	Dataset
[43]	Concept-based representation outperformed BOW and n-grams.	K-NN	F-measure	Jaccard Cosines Inner Dice	Online newspaper
[76]	Semantic class features clustering outperformed K-means	K-means	Purity MICD DBI	Squared Euclidean	Online Arabic newspaper
[74]	RDF and DBpedia integrated to represent Wikipedia articles	-	-	-	60 Wikipedia articles
[77]	Classification efficiency is improved by hybrid representation(DBMS, FastText)	-	-	-	News datasets

IV. CONCLUSIONS

A survey of the literature on Arabic document clustering has been conducted, with particular emphasis on incorporation of semantic information and on the strategies for handling the high dimensionality inherent to sparse and latent Vector Space Methods. Although sparse and latent Vector Space Methods provide a somewhat limited perspective, the body of work reviewed shows considerable methodological diversity, and it reveals an upward trend in research activity over the years, with the interval 2021 to 2025 emerging as the most productive in terms of research output. This growing interest reflects the expanding availability of Arabic digital content and the corresponding need for scalable, unsupervised methods to organize and explore large document collections.

Clustering techniques have been employed across a wide range of applications such as information retrieval, text summarization, topic discovery, and as a preprocessing step for classification. In many practical scenarios, clustering plays a fundamental role by enabling the discovery of latent thematic structures in unlabelled data and by reducing the dependence on manually annotated resources, which remain scarce for Arabic. As modern applications increasingly rely on large-scale text analytics, effective document clustering is becoming an essential component for managing information overload and supporting semantic tasks.

The impact of stemming on clustering performance was investigated in many works, yet no clear consensus has emerged. Some studies reported performance gains attributed to dimensionality reduction and increased term overlap, while others observed degradation due to semantic loss or over-stemming effects. These mixed findings underscore the sensitivity of clustering outcomes to preprocessing choices, especially in Arabic, where morphological variation plays a central role in meaning representation.

Semantic information was explicitly considered in only a limited subset of studies, most commonly through named entity recognition techniques. Despite the recognized importance of semantics for effective clustering, most approaches still rely on representations based on lexical similarity. Recent advances in embedding-based representations, such as word embeddings, sentence embeddings, and transformer-based models, offer a promising alternative.

With respect to evaluation, news datasets were by far the most frequently used benchmarks. While such datasets provide convenient and standardized test beds, this restricted focus limits the generalizability of results to other domains such as biographical texts, religious texts, or legal documents, where linguistic variability and noise are often more pronounced.

The morphological richness and syntactic complexity of Arabic, together with orthographic variations, complicate core preprocessing tasks, which directly affect clustering quality. Compared to languages that have been extensively studied in this area, Arabic remains underrepresented in clustering-focused natural language processing research, which could support a wide range of real-world Arabic-

language applications.

REFERENCES

[1] F. S. Al-Anzi, D. Abuzeina. (2015, Dec) Stemming Impact on Arabic Text Categorization Performance: a Survey. Presented at the 5th international conference on information and communication technology and accessibility (ICTA). IEEE.[Online]. Available:https://ieeexplore.ieee.org/abstract/document/7426875/.

[2] H. Froud, R. Benslimane, A. Lachkar, S. A. Ouatik. (2010, Sep). Stemming and similarity measures for Arabic documents clustering. Presented at the 2010 5th International Symposium on I/V Communications and Mobile Network. IEEE. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5656417/.

[3] A. Kelaiaia, H. Merouani. (2016, Mar). Clustering with Probabilistic Topic Modeling on Arabic Texts: A Comparative Study of LDA and K-Means. The International Arab Journal of Information Technology. [Online]. 13(2), pp.332–338. Available: http://www.ccis2k.org/iajit/PDF/Vol.13,%20No.2/6146.pdf.

[4] A. K. Jain, M. N. Murty, P. J. Flynn. (1999). Data Clustering: A Review. ACM Computing Surveys (CSUR). [Online], 31(3), pp. 264–323. Available: https://dl.acm.org/doi/abs/10.1145/331499.331504.

[5] K. N. Singh, H. M. Devi, A. K. Mahanta. (2017, Jun). Document representation techniques and their effect on document Clustering and Classification: A Review.International Journal of Advanced Research in Computer Science. [Online]. 8(5). Available: https://www.ijarcs.info/index.php/Ijarcs/article/view/3822.

[6] Q. Bsoul, R.A. Salam, J. Atwan, M. Jawarneh. (2021, Dec). Arabic text clustering methods and suggested solutions for theme-based quran clustering: analysis of literature. Journal of Information Science Theory and Practice. [Online]. 9(4), pp: 15-34.Available:https://koreascience.kr/article/JAKO202100551544940.page

[7] R. Maronna, C. C. Aggarwal, C. K. Reddy. Data Clustering Algorithms and Applications. 2016. ISBN: 9781466558212.

[8] R. Baeza-Yates, B. Ribeiro-Neto. (1999, Jan). Modern information retrieval. [Online]. 463. Available: https://people.ischool.berkeley.edu/~hearst/irbook/print/chap10.ps.gz.

[9] M. Afzali, S. Kumar. (2019, Feb). Text Document Clustering: Issues and Challenges. Presented at the 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8862247.

[10] H. M. Alghamdi, A. Selamat. (2019, Jan). Arabic Web page clustering: A review.Journal of King Saud University– Computer and Information Sciences. [Online]. 31(1), pp. 1-14. Available: https://www.sciencedirect.com/science/article/pii/S1319157817300290.

[11] N. Shah, S. Mahajan. (2012, Aug). Semantic-based Document Clustering: A Detailed Review.International Journal of Computer Applications. [Online]. 52(5), pp. 42–52. Available: https://www.academia.edu/download/74479640/pxc3881598.pdf.

[12] R. Dounia, K. Yasssine, E. N. Noureddine. (2022, May). Exploring text representation impact on K-means based Arabic text documents clustering. Presented at 2022 International Conference on Intelligent Systems and Computer Vision, ISCV2022. Institute of Electrical and Electronics Engineers Inc.[Online], pp. 1-5. Available: https://ieeexplore.ieee.org/abstract/document/9806067/.

[13] C. D.Manning, P. Raghavan, H. Schütze. (2008, May 27). Introduction to Information Retrieval. [Online]. 39. Available: https://www.cis.unimuenchen.de/~hs/teach/14s/ir/pdf/19web.pdf.

[14] L. Ajallouda, F. Z. Fagroud, A. Zellou, E. Benlahmar. (2020, Nov). K-means HAC and fcm which clustering approach for Arabic text? Presented at ACM International Conference Proceedings Series. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3419604.3419779.

[15] Soliman, Abu Bakr; Eissa, Kareem; El-Beltagy, Samhaa R. (2017, Nov). Aravec: A set of Arabic word embedding models for use in Arabic NLP. Presented at: 3rd International Conference on Arabic Computational Linguistics.[Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917321749.

[16] W. Antoun, F. Baly, H. Hajj. (2020). Arabert: Transformer-based model for Arabic language understanding. arXiv preprint arXiv:2003.00104.[Online]. Available: https://arxiv.org/abs/2003.00104.

[17] S. A. Chowdhury, A. Abdelali, K. Darwish, J. Soon-Gyo, J. Salminen, B. J.Jansen. Improving Arabic text categorization using transformer training diversification. In: Proceedings of the fifth Arabic natural language processing workshop, Barcelona, 2020, pp. 226–236.

[18] W. Antoun, F. Baly, H. Hajj. AraGPT2: Pre-trained transformer for Arabic language generation. In Proceedings of the sixth Arabic natural language processing workshop, Kyiv, 2021, pp: 196-207.

[19] V. Mehta, M. Agarwal, R. K. Kaliyar. (2024, Apr). A comprehensive and analytical review of text clustering techniques. International Journal of Data Science and Analytics. [Online]. 18(3), pp. 239–258. Available:https://link.springer.com/article/10.1007/s41060-024-00540-x.

[20] M. Alian, A. Awajan. (2020, Apr). Semantic similarity for English and Arabic texts: a review. Journal of Information and Knowledge Management. [Online]. 19(4), pp. 2050033. Available: https://www.worldscientific.com/doi/abs/10.1142/S0219649220500331.

[21] N. M. Salih, K. Jacksi. (2020, May). State of the art document clustering algorithms based on semantic similarity.Informatika. [Online]. 14(2), pp. 58–75. Available: https://pdfs.semanticscholar.org/3894/0587856efae6550b54a133ca4ebe141d1b18.pdf.

[22] *G. Salton, A. Wong, C. S. Yang. (1975, Nov).* A vector space model for automatic indexing. Communications of the ACM. [Online].*18(11), pp: 613–620. Available: https://dl.acm.org/doi/abs/10.1145/361219.361220.*

[23] G. Salton, C. Buckley. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management. [Online].24(5), pp: 513–523. Available: https://www.sciencedirect.com/science/article/pii/0306457388900210

[24] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. (1990, Sep). Indexing by latent semantic analysis. Journal of the American Society for Information Science. [Online]. 41(6), pp: 391–407. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9.

[25] M. W. Berry, P. G. Young. (1995, Dec). Using latent semantic indexing for multilanguage information retrieval.Computers and the Humanities. [Online]. 29(6), pp: 413–429. Available: https://link.springer.com/article/10.1007/BF01829874.

[26] T. Mikolov, K. Chen, G. Corrado, J. Dean. (2013, Sep). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. [Online]. Available: https://arxiv.org/abs/1301.3781.

[27] T.Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (NIPS). 2013, pp: 1532-1543.

[28] J. Pennington, R. Socher, C. D. Manning. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014, pp: 1532–1543.

[29] O. Shahmirzadi, A. Lugowski, K. Younge. (2019, Dec). Text Similarity in Vector Space Models: A Comparative Study. Presented at: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). [Online]. Available:https://ieeexplore.ieee.org/abstract/document/8999168/

[30] P. Sitikhu, K.Pahi, P. Thapa, S.Shakya. (2019, Nov). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. Presented at: 2019 Artificial Intelligence for Transforming Business and Society (AITB). [Online]. Available:https://ieeexplore.ieee.org/abstract/document/8947433/

[31] A. Tahani, S., Shireen, A., Reem, S., Shahendah, N., Sarah. (2024). Preprocessing Techniques for Clustering Arabic Text: Challenges

and Future Directions. International Journal of Advanced Computer Science and Applications. [Online]. 15(8), pp. 1301-1314. Available: https://doi.org/10.14569/ijacsa.2024.01508126.

[32] O.M. Al-Omari. Evaluating the Effect of Stemming in Clustering of Arabic documents.(2011, Jul). Academic Research International1. [Online]. 1(1), pp. 284–291. Available: http://www.savap.org.pk/journals/ARInt./Vol.1(1)/2011(1.1-28).pdf.

[33] B. Q. Walid, M. Mohd. Effect of ISRI stemming on similarity measure for Arabic document clustering.7th Asia Information Retrieval Symposium, pp: 584-593. Berlin, Heidelberg: Springer Berlin Heidelberg. Dec, 2011.

[34] H. Froud, A. Lachkar,S. A. Ouatik. (2013, Feb). Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. In: arXiv preprint arXiv: 1302.1612. [Online]. Available: https://arxiv.org/abs/1302.1612.

[35] H. Froud, A. Lachkar. Agglomerative hierarchical clustering techniques for Arabic documents. In: Advances in Computational Science, Engineering and Information Technology: Proceedings of the Third International Conference on Computational Science, Engineering and Information Technology (CCSEIT-2013), KTO Karatay University,2013, pp. 255–267.

[36] O. A Ghanem, W. M. Ashour. (2012, Jul). Stemming Effectiveness in Clustering of Arabic Documents.International Journal of Computer Applications. [Online]. 49(5), pp.1–6. Available: https://www.academia.edu/download/63671801/pxc3880674.pdf.

[37] M. H. Ahmed, S. Tiun. (2014, Sep). K-means based algorithm for Islamic document clustering. International Journal on Islamic Applications in Computer Science and Technology. [Online]. 2(3), pp. 13-21. Available: https://www.academia.edu/download/101744316/836.pdf.

[38] H. N. Fejer, N. Omar. Automatic Arabic text summarization using clustering and keyphrase extraction. In: Proceedings of the 6th International Conference on Information Technology and Multimedia. IEEE, Putrajaya, 2014, pp. 293-298.

[39] A. T. Al-Taani, S. H. Al-Sayadimailto:robertas.damasevicius@vdu.lt. (2022, Dec). Extractive text summarization of Arabic multi-document using fuzzy C-means and Latent Dirichlet Allocation. International Journal of System Assurance Engineering and Management. [Online]. 15(2), pp: 713-726. Available: https://link.springer.com/article/10.1007/s13198-022-01783-2.

[40] A.H. Aliwy, K. Aljanabi, H.A. Alameen. (2022, Jan). Arabic text clustering technique to improve information retrieval. Presented at AIP Conference Proceedings. [Online]. 2386(1). Available: https://pubs.aip.org/aip/acp/article-abstract/2386/1/050021/2820614.

[41] A. F.Alsuhaim, A. M.Azmi, M. Hussain. (2021, Apr). Improving the retrieval of Arabic web search results using enhanced K-Means clustering algorithm. Entropy[Online]. 23(4), pp.1–19. Available:https://doi.org/10.3390/e23040449.

[42] I. Sahmoudi, A. Lachkar. (2017, Apr). Formal concept analysis for Arabic web search results clustering.Journal of King Saud University-Computer and Information Sciences. [Online]. 29(2), pp. 196–203. Available:https://www.sciencedirect.com/science/article/pii/S1319157816300696.

[43] M. Alruily, A. Ayesh, A. Al-Marghilani. “Using Self-Organizing Map to Cluster Arabic Crime Documents”. In Proceedings of the International Multi-Conference on Computer Science and Information Technology, Wisla, 2010, pp.357–363.

[44] S. Salloum, K. Tahat, A. Mansoori, R. Alfaisal, D. Tahat. (2024, Dec). Leveraging K-Means Clustering for Analysis of Arabic Hate Speech Tweets. Presented at: Global Congress on Emerging Technologies (GCET-2024). IEEE. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10934641/.

[45] A. A. Badawy, A. Salah, M. A. Mahdi. (2025, Jun). Exploring Thematic Structures in Surah Al-Baqarah using TF-IDF, Dimensionality Reduction, and K-means Clustering. Journal of Computing and Communication. [Online]. 4(2), pp.1-12. Available: https://jocc.journals.ekb.eg/article_446634.html.

[46] A. K. Sangaiah1, A.E. Fakhry, M. Abdel-Basset, I. El-henawy.(2018, Feb). Arabic text clustering using improved clustering algorithms with dimensionality reduction. Cluster Computing. [Online]. 22, pp.4535-4549. Available: https://link.springer.com/article/10.1007/s10586-018-2084-4.

[47] A. A. Badawy, A. Salah, M. A. Mahdi. (2025, Jun). Exploring Thematic Structures in Surah Al-Baqarah using TF-IDF, Dimensionality Reduction, and K-means Clustering. Journal of Computing and Communication. [Online]. 4(2), pp.1-12. Available: https://jocc.journals.ekb.eg/article_446634.html.

[48] D. Abuaiadah. (2016, Dec). Using bisect k-means clustering technique in the analysis of Arabic documents. ACM Transactions on Asian and Low-Resource Language Information Processing. [Online]. 15(3), pp. 1-13. Availabe:https://dl.acm.org/doi/abs/10.1145/2812809

[49] M.S. Alkoffash. (2012, Aug). Comparing Arabic Text Clustering using K-means and K-medoids. International Journal of Computer Applications. [Online]. 51(2), pp. 5-8. Available: https://www.researchgate.net/publication/258651793_Automatic_Arabic_Text_Clustering_using_K-means_and_K-mediods.

[50] M. Alhawarat, M. Hegazi. (2018, July), Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. IEEE Access. [Online]. 6, pp. 42740–42749. Available:https://ieeexplore.ieee.org/abstract/document/8402221

[51] A. R. Alharbi, M. Hijji,A. Aljaedi. (2021, Nov). Enhancing topic clustering for Arabic security news based on k-means and topic modelling. IetNetworks. [Online]. 10(6), pp. 278–294. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ntw2.12017.

[52] A. S. Daoud, A. Sallam, M. E. Wheed. (2017, Sep). Improving Arabic Document Clustering using K-Means Algorithm and Particle Swarm Optimization. Presented at 2017 Intelligent Systems Conference (IntelliSys). IEEE. [Online], pp.879–885. Available: https://ieeexplore.ieee.org/abstract/document/8324233/.

[53] D. M. Alsekait, H. Fathi, J. Atwan, M. Jawarneh, Q. Bsoul, D. S. AbdElminaam, S. Alzoubi. (2024, Dec). Arabic Clustering Through Advanced Stemming and WordNet-Based Extraction for Water Cycle Cluster. International Journal of Data Warehousing and Mining (IJDWM). [Online]. 20(1), pp: 1-25. Available: https://dl.acm.org/doi/10.4018/IJDWM.352601.

[54] H. Alghamdi, A. Selamat. (2015, Jul). Topic Modelling used to improve Arabic Web pages Clustering. Presented at 2015 International Conference on Cloud Computing(ICCC). IEEE. [Online]. Available:https://ieeexplore.ieee.org/abstract/document/7149662.

[55] F. S. Al-anzi, D. Abuzeina. (2016, Jan). Big Data Categorization for Arabic Text Using Latent Semantic Indexing and Clustering. Presented at International Conference on Engineering Technologies and Big Data Analytics (ETBDA’2016). [Online]. Available: https://www.sciencegate.app/document/10.15242/iie.e0116001

[56] F. S. Al-Anzi, D. AbuZeina. (2016). Reinforcing Arabic Language Text Clustering. Presented at: International Arab Conference on Information Technology(ACIT’2016). [Online]. Available:https://acit2k.org/ACIT/images/stories/year2014/month1/10_.pdf.

[57] H. M. Alghamdi, A. Selamat, N. S. Abdul Karim. (2014, Sep). Improved text clustering using k-means Bayesian vectoriser. Journal of Information and Knowledge Management. [Online]. 13(3), pp. 1450026. Available: https://www.worldscientific.com/doi/abs/10.1142/s0219649214500269.

[58] A . Awajan. Semantic vector space model for reducing Arabic text dimensionality. (2015, Apr). Presented at 2015 Fifth International Conference on Digital Information and Communication Technology and its Applications (DICTAP). IEEE.[Online]. Available: https://ieeexplore.ieee.org/abstract/document/7113185/.

[59] M. Hussein, A. Alsammak, T. Elshishtawy. Keyphrase-Based Hierarchical Clustering for Arabic Documents. In: Proceedings of the 10th International Conference on Informatics and Systems - INFOS ’16. Association for Computing Machinery, 2016, pp. 61–67.

[60] D. M. Alsekait, H. Fathi, J. Atwan, M. Jawarneh, Q. Bsoul, D. S. AbdElminaam, S. Alzoubi. (2024, Dec). Arabic Clustering Through

Advanced Stemming and WordNet-Based Extraction for Water Cycle Cluster. International Journal of Data Warehousing and Mining (IJDWM). [Online]. 20(1), pp: 1-25. Available: https://dl.acm.org/doi/10.4018/IJDWM.352601.

[61] B. C.M. Fung, K. Wang, M. Ester. Hierarchical Document Clustering Using Frequent Itemsets. In: Proceedings of the 2003 SIAM international conference on data mining. Society for Industrial and Applied Mathematics (SIAM), 2003, pp. 59–70.

[62] K.A. Salman, H. K. Khafaji. (2024, Jun). A New Algorithm for Arabic Document Clustering Utilizing Maximal Wordsets. Revue Intelligence Artificielle. [Online]. 38(3), pp. 805-813. Available: https://iieta.org/journals/ria/paper/10.18280/ria.380307.

[63] H. S. Al-Sarrayrih, R. AlShalabi. Clustering Arabic documents using frequent itemset-based hierarchical clustering with N-grams. In Proceedings: Int. Conf Inf. Technol.(ICIT). 2009, p. 113.

[64] A. A. Mohamed. An effective dimension reduction algorithm for clustering Arabic text. (2020, Mar). Egyptian Informatics Journal. [Online]. 21(1), pp.1–5. Available: https://www.sciencedirect.com/science/article/pii/S1110866518301579.

[65] Y. A. Alhajmailto:yalhag@gmail.com, A. Dahoumailto:dahou.abdghani@univ-adrar.edu.dz, M. A. A. Al-qanessmailto:alqaness@whu.edu.cn, L. Abualigahmailto:aligah.2020@gmail.com, A. A. Abbasi, N. A. O. Almaweri, M. Abd Elaziz, R. Damaševičius. (2022, Jun). A novel text classification technique using improved particle swarm optimization: A case study of Arabic language. Future Internet. [Online]. 14(7), pp: 194.

[66] D.Abuzeina. (2019, Jul). Exploring bigram character features for Arabic text clustering. Turkish Journal of Electrical Engineering and Computer Sciences. [Online]. 27(4), pp: 3165-3179. Available: https://journals.tubitak.gov.tr/elektrik/vol27/iss4/56/.

[67] D. Namly, K. Bouzoubaa, R. Tachicart. (2025, Jan). A three-phase model for keyword detection in Arabic corpora. Indonesian journal of electrical engineering and computer science. [Online]. 37(1), pp: 206.Available:https://ijeecs.iaescore.com/index.php/IJEECS/article/view/38741.

[68] M. Muhammed, S. Azab, N. Ali1, M. Gheith. (2025, Nov). Automatic term extraction for Arabic text: Approaches, techniques, and challenges. Journal of Artificial Intelligence in Engineering Practice. [Online]. 2(2), pp: 51-64. Available: https://journals.ekb.eg/article_459839.html.

[69] T. Almutairi, R. Alotaibi, S. Saifuddin, S. Sarhan. (2025, Jul). A Hybrid Metaheuristic Approach for Arabic Text Clustering. Arabian Journal for Science and Engineering. [Online]. pp: 1-19. Available: https://link.springer.com/article/10.1007/s13369-025-10452-y.

[70] S. Larabi-Marie-Sainte, M. B. Alamir, A. Alameer.(2023, Sep). Arabic Text Clustering Using Self-Organizing Maps and Grey Wolf Optimization. Applied Sciences. [Online]. 13(18), p. 10168. Available: https://www.mdpi.com/2076-3417/13/18/10168.

[71] K. A. Salman, H. K. Khafaji. (2024, Dec). Unprecedented Rough Sets Model for Arabic Document Clustering. Ingénierie des Systèmes d'Information. [Online]. 29(6), pp: 2503. Available: https://www.iieta.org/journals/isi/paper/10.18280/isi.290635.

[72] R. K. Ibrahim, S. R. M. Zeebaree, K. F. S. Jacksi. (2019). Survey on semantic similarity based on document clustering. Advances in Science, Technology and Engineering Systems. [Online]. 4(5), pp. 115–122. Available: https://www.researchgate.net/profile/Subhi-Zeebaree/publication/335700827_Survey_on_Semantic_Similarity_Based_on_Document_Clustering/links/5e1b708d299bf10bc3a9056a/Survey-on-Semantic-Similarity-Based-on-Document-Clustering.pdf.

[73] A. Karima, E. Zakaria, T. G. Yamina. (2012, Apr). Arabic text categorization: a comparative study of different representation modes.Journal of Theoretical and Applied Information Technology. [Online]. 38(1), pp.1–5. Available: https://jatit.org/volumes/Vol38No1/thirtyeighth_volume_1_2012.php

[74] G. Zakria, M. Farouk, K. Fathy, M. N. Makar. Semantic Representation Extraction from Unstructured Arabic Text. In: Proceedings of the 8th International Conference on Software and Information Engineering. 2019, pp. 222–226.

[75] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters. (2014, May). Developing language processing components with GATE. [Online]. 8. Available: https://gate.ac.uk/releases/gate-8.0-build4825-ALL/doc/tao/tao.pdf.

[76] H. M. Alghamdi, A. Selamat, N. S. Abdul Karim. (2014, Dec). Arabic web pages clustering and annotation using semantic class features. Journal of King Saud University- Computer and Information Sciences [Online]. 26(4), pp. 388–397. Available: https://www.sciencedirect.com/science/article/pii/S1319157814000263.

[77] M. Louail, C. K. Hamdi-Cherifhttps://orcid.org/0000-0002-3773-5233, A. Hamdi-Cherif. (2024, Aug). Tasneef: a fast and effective hybrid representation approach for Arabic text classification. IEEE Access. [Online]. 12, pp. 120804 - 120826. Available: https://ieeexplore.ieee.org/abstract/document/10654296/.