IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received May 22, 2018, accepted June 25, 2018, date of publication July 3, 2018, date of current version August 20, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2852648

Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents

M. ALHAWARAT AND M. HEGAZI
Department of Computer Science, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
Corresponding author: M. Alhawarat (m.alhawarat@psau.edu.sa)

This work was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University under Grant 2015/01/3855.

ABSTRACT Clustering Arabic text documents is of high importance for many natural language technologies. This paper uses a combined method to cluster Arabic text documents. Mainly, we use generative models and clustering techniques. The study uses latent Dirichlet allocation and k-means clustering algorithm and applies them to a news data set used in previous similar studies. The aim of this paper is twofold: it first shows that normalizing the weights in the vector space, for the document-term matrix of the text documents, dramatically improves the quality of clusters and hence the accuracy of clustering when using k-means algorithm. The results are compared to a recent study on clustering Arabic text documents. Second, it shows that the combined method is superior in terms of clustering quality for Arabic text documents according to external measures, such as purity, F-measure, entropy, accuracy, and other measures. It is shown in this paper that the purity of the combined method is 0.933 compared to 0.82 for k-means algorithm, and these figures are higher in comparison to a recent similar study. This is also confirmed by the other used validation measures. The correctness of the combined method is then confirmed using different Arabic data sets.

INDEX TERMS Clustering text documents, K-means, Arabic language, topic modeling, latent Dirichlet allocation (LDA).

I. INTRODUCTION

Clustering text documents is of high importance in the era of information explosion. Data on the internet is dramatically increasing every single day. Large part of that data is in text format and in most cases exist with no labels. Notwithstanding, manually annotating text documents is usually a tedious human task; although automatic annotation techniques exist, still they are not accurate. For this and other reasons, clustering is considered an important data mining technique in categorizing, summarizing, organizing and classifying text documents. Having said so does not mean that clustering gives better results than classification when labels are available for data.

Extracting information from text sources comprises one important task that is used nowadays for several purposes, especially in natural language processing (NLP). Some language technologies need information about text documents to accomplish certain tasks with high performance. Dealing with natural languages is not an easy task, especially for some languages including Arabic. This is due to many reasons such as the lack of benchmark data sets and related resources, absence of standard normalization methods, inadequacy of accurate stemming algorithms, the highly derivative nature of Arabic words and ambiguity imposed by diacritic marks [1], [2].

Topic modeling is an important field of study that gained great attention in last years. It has important applications in many fields like Information Retrieval (IR) and NLP. Topic modeling aims at extracting a pre-specified number of topics from a set of text documents based on statistical concepts. This process is considered as an unsupervised task where no prior knowledge about the text is required. Topic modeling and clustering are much alike; they are both: unsupervised learning techniques, need a number of categories to be specified beforehand and require no labels to operate.

Topic modeling has many benefits in the context of our study; it serves as a mechanism for both feature reduction and feature selection. First, we use topic modeling techniques to reduce the vector space model (VSM) to a simpler, and ultimately a representative one. This can be considered a good solution to the very well-known problem of high-dimensionality in data and text mining. Second, the proposed methodology in this study considers topic modeling as a feature selector by uncovering latent semantic variables in text documents.

Our work is inspired by the recent work [3] in integrating topic modeling and clustering. The aim of that work was to achieve better results for recognizing local topics within one document, and a group of global topics across a set of text documents using LDA and clustering techniques. They also used Bernoulli distribution to decide between local and global topics. Their work may be viewed as a method of linking the results of one technique to be the input to the other in order to extract better topics and to achieve better clustering. We use a similar methodology and apply it to Arabic text documents.

This study considers a news dataset [4] composed of 2700 documents of 9 categories. In this study we use external measures to evaluate the resulted clusters, such as purity, F-measure, entropy, accuracy, and others.

To validate the correctness of the combined method, it is applied to several Arabic datasets; these are freely available on the internet and used to verify and confirm the correctness of the combined method.

This study utilizes a combined method of clustering algorithms and topic modeling techniques to cluster Arabic text documents. The performance of this methodology gives better results than regular clustering algorithms. Different Arabic news datasets are used in the study to validate the methodology. Also, different external performance measures are calculated for both combined and regular clustering methods. The results for the combined method is superior in terms of the used external measures.

The rest of the document is organized as follows: section II presents literature review, section III discusses the clustering algorithms and validation plans that are used in this study, section IV describes the data preparation and methodology, section V illustrates experiments and results, section VI includes discussion and section VII concludes the paper.

II. LITERATURE REVIEW

There exist few research works that integrate topic modelling techniques with clustering algorithms and apply them to English text documents [3]. To the best of our knowledge; this study is the first that integrates and applies topic modeling techniques and clustering algorithms together to Arabic text documents.

A. CLUSTERING ARABIC DOCUMENTS

Some studies applied clustering algorithms to Arabic text [5]–[9]. For example, recent work by Abuaiadah [5] used bisect k-means clustering algorithm to analyze and cluster Arabic text documents. They use an in-house 2700 news documents classified into 9 categories. The author showed that such an algorithm gives better results compared to standard k-means algorithm, he used different distance and similarity measures. Al-Sarrayrih and Al-Shalabi [6] have clustered Arabic text documents based on Frequent Itemset Hierarchical Clustering algorithm (FIHC). They applied their algorithm on an in-house 600 documents classified in 6 classes. They obtained promising accuracy compared to clustering European languages. Froud and Lachkar [7] have applied hierarchical clustering algorithm to Arabic text documents with different distance measures including: Euclidean distance, Cosine Similarity, Jaccard Coefficient, and Pearson Correlation Coefficient. They report that Ward function outperforms other linkage techniques and that using stemming algorithms will not improve accuracy of clustering results but makes clustering faster. Ghwanmeh [8] showed that using clustering algorithms enhance retrieval of information compared to IR systems without clustering; where they used hierarchal k-means algorithm. El-Haj *et al.* [10] used k-means clustering algorithm in multi-document extractive summarization process. Their results compared well to top systems at Document Understanding Conference (DUC) 2002. Hussein *et al.* [11] used hierarchical clustering algorithms to cluster 345 documents into 12 categories. They used lemma-based similarity measure that is based on shared key-phrases among documents. They reported a high purity of around 0.95; however, the data set is very small (each category has an average of 28 documents) and the key phrases extraction process is not clear. Froud *et al.* [12] applied k-means clustering algorithm on Corpus of Contemporary Arabic (CCA)) which is composed of 12 categories. They used different similarity measures and report the highest purity of 0.77 using Euclidean distance measure. However, the dataset they used is different from what is found in the literature. Also, the dataset has few number of documents (432) and large number of categories (16).

B. TOPIC MODELLING

Topic modeling techniques choose a set of topics each with a group of words using statistical methods; they try to find a set of topics in a group of text documents; where each topic is defined as a distribution over a set of words. This is achieved using statistical modeling. There are different flavors of topic modeling [13]. In this study, we use LDA for topic modeling with algorithms such as: Gibbs Sampling [14], variational expectation-maximization (VEM) [15], VEM fixed and Correlated Topic Modeling (CTM) [16].

Topic Modeling aims at extracting main topics from a set of text documents. It has been shown that LDA outperforms other models such as Latent Semantic Analysis (LSA) [17]–[19]. LDA has been applied to many fields of study such as NLP [20]–[22].

The idea behind topic modeling is that a set of words are represented by a probabilistic distribution. First, words in the document are assigned with random probabilities, and during the running of the algorithm, these probabilities are updated to infer the latent structure of topics in that document. In LDA, Dirichlet distributions are used to infer such structures. More details about topic modeling can be found in [14], [15], [23], and [24].

C. EXTRACTING TOPICS FROM ARABIC DOCUMENTS

There exist studies that exploit topic modeling techniques to extract topics from the Arabic documents. For example, Ayadi *et al.* [25] used topic modeling techniques (LDA) to extract the main topics of an in-house Arabic corpus. They show that using the reduced word space after applying LDA, produces more accurate results when classifying documents. Siddiqui *et al.* [26] applied LDA to a sample of the holy Quran to extract thematic structure and also main topics. In one setup, results show classification of chapters into two categories: Makki and Madni (time/place of revelation). In another setup, topics with 5,10 and 15 terms are extracted. Although there are some stop words not removed and no definite topics are noticed, still results are promising. Also, Alhawarat [27] applied LDA techniques to a sample of the holy Quran to extract main topics. Although results are promising, but they show few number of coherent topics. Brahmi *et al.* [28] studied the effect of stemming algorithms on topic modeling of Arabic Text. They show that stemming induces improvement in the results of extracting accurate topics.

In a different context, Kelaiaia and Merouani [29] compared between LDA and k-means on Arabic text documents. The results show that LDA outperforms k-means using external measures such as F-measure.

At last, very few studies considered combining topic modeling with clustering algorithms. For example, Xie and Xing [3] proposed a new methodology that integrates topic modeling with clustering algorithms in two manners. First, they used topic modeling to improve the quality of clustering. Second, they used clustering to extract local and global topics. They have applied their methodology on both Reuters-21578 and 20-Newsgroups datasets. Results of their experiments showed better quality of clustering compared to different other techniques using coherence measure. They showed that topic modeling and clustering are two related and mutual techniques.

III. CLUSTERING ALGORITHMS AND VALIDATION TECHNIQUES

This section discusses clustering algorithms and validation methods in general including those used in this study.

A. CLUSTERING ALGORITHMS

Clustering algorithms might be divided into two types: partitioning and hierarchal. In partitioning methods, the number of clusters must be specified before clustering takes place. Once this is specified, then random initial centers are chosen, and then objects are assigned to the nearest center according to the distance between objects and centers. This process is repeated until no further improvement. Examples of clustering algorithms of this type are: k-means and k-medoids [30], [31].

On the other hand, hierarchal methods have no pre-specified number of clusters, because this type either considers each object as a cluster (agglomerative), or considers the whole data as one cluster (divisive). Then it starts to either increase or decrease the number of clusters until a criterion is met [32].

In this study, k-means algorithm is used for several reasons: simplicity, performance and wide usage. Although better flavors of k-means exist such as Bisect K-Means [30]–[32], still the aim of this study is not to compare between clustering algorithms; but instead to improve the quality of clusters.

B. CLUSTERING VALIDATION TECHNIQUES

Clustering is an unsupervised learning technique, where labels are not provided or even do not exist in some cases. Although there are automated and semi-automated techniques for labeling data, still they may not be accurately used to validate clustering.

Validation of clustering is very important to decide on configuration of parameters and methods to be used for a specific data. Validation methods are usually divided into three categories [31]:

- External: are based on external information about clusters in order to evaluate accuracy of clustering.
- Internal: are based on calculating indices without having labels to decide the quality of one clustering.
- Relative: are used to compare results of two clusterings for the same data, using different parameter settings or different clustering algorithms.

External measures can be used if class labels exist for the data. Evaluation is then used to benchmark resulted clusters to validate quality of clusters. In contrast, internal measures are used when class labels are not available. Relative measures have the same purpose of internal measures, and is used to compare the quality of two clusterings for the same data with either different clustering algorithms or different parameter settings.

Since the labels exist for data in this study, then external measures are used. External validation methods can be classified to different categories [33], [34]. The following are the categories with examples:

- Matching based measures: purity, recall, precision and F-measure.
- Entropy and information based measures: entropy, normalized mutual information (NMI) and normalized variation of information (NVI).
- Pairwise and counting measures: accuracy (rand-index) and jaccard index.

These are some of the most validation measures used in the literature, and are used in this study to validate the quality of the resulted clusterings. The following sub-subsections will give a very brief description to the aforementioned validation methods. For more information on these please refer to [35]–[38].

1) PURITY

Purity [39] is an evaluation measure of how pure is a cluster with regard to the dominant class in that cluster. Purity is then computed based on the percentage of all objects of dominant classes for each cluster with regard to the number of all objects:

$$purity = \frac{1}{N}\sum_{k} max_j|\omega_k \cap c_j| \quad (1)$$

where N is the number of all objects, k is the number of clusters, ω_k is the dominant class, and c_j is the real class (ground truth). The largest the value of purity the better clustering with maximum value of one if the dominant class of a cluster represents all objects in that cluster.

2) F-MEASURE

This measure [40] is the harmonic mean of both recall and precision. Recall represents the fraction of documents of one category in one cluster out of all documents of that category. Whereas precision is the fraction of documents of one category in one cluster out of all documents in that cluster. Note from such definitions that values of precision and recall in isolation will not give a correct indication of the quality of clustering for several reasons found in the literature, therefore a combination of the two makes sense when appear in one measure, viz., the F-measure. To compute recall, precision and F-measure, then confusion matrix is usually used which is composed of four values as table 1 shows.

TABLE 1. Confusion matrix for Clustering.

	Same cluster	Different cluster
Similar documents	True Positive (TP)	False Negative (FN)
Different documents	False Positive (FP)	True Negative (TN)

The confusion matrix for clustering is based on all possible combination-pairs of all documents chosen from all clusters, where:

- TP: indicates that the two documents are similar and belong to the same cluster.
- FN: indicates that the two documents are similar and belong to different clusters.
- FP: indicates that the two documents are different and belong to the same cluster.
- TN: indicates that the two documents are different and belong to different clusters.

Based on these values, then we can calculate recall, precision and F-measure according to the following equations:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F - measure = \frac{2 \times P \times R}{P + R} \quad (4)$$

where, P represents precision, and R represents recall. Greater values for F-measures means better and precise clustering.

3) ENTROPY

Entropy [41] represents the class distribution of objects within each cluster. If a cluster contains objects with the same class then entropy is 0. Otherwise the value increases with more mixed classes in the same cluster with a value that might exceed one. To calculate entropy for a cluster, then class distribution of objects in each cluster is calculated as:

$$E_j = \sum_{i} p_{ij} log(p_{ij}) \quad (5)$$

Then the sum is computed for all classes. After that Entropy is calculated as follows:

$$E = \sum_{j=1}^{m} \frac{n_j}{n} E_j \quad (6)$$

where m is the number of clusters, n_j is the size of cluster j, and n is the number of all objects.

4) NORMALIZED MUTUAL INFORMATION

Mutual information is a popular statistical measure that compares the shared information between two clusterings, usually the resulted clustering and the ground truth of the data. Although this is a good measure, but it cannot be used to compare different data clusterings. Instead, if normalized then it can be used to compare the quality of different clustering results. The NMI is defined as [42]:

$$NMI(X, Y) = \frac{I(X, Y)}{\sqrt{H(X) \times H(I)}} \quad (7)$$

where X and Y represents class and cluster labels respectively, I(X,Y) represents the mutual information between X and Y, and H(X) and H(Y) represent the entropy of X and Y respectively. Greater value of NMI means more mutual information and hence more similarity between clusterings.

5) NORMALIZED VARIATION OF INFORMATION

This is another measure that is based on entropy and information. It depends on the lost and gained information when comparing two clusterings. NVI is defined as [43]:

$$NVI(X, Y) = \begin{cases} \frac{H(X|Y) \times H(Y|X)}{H(X)} & H(X) \neq 0 \\ H(Y) & H(X) = 0 \end{cases} \quad (8)$$

where H() is the entropy function and $H(X|Y)$ and $H(X|Y)$ are conditional entropy. When NVI approaches 0 then this means total agreement between cluster labels of X and Y, hence homogeneous clusterings. When the value gets larger, this means decrease in agreement, and when it reaches 1, this means total heterogeneous clusterings.

6) ACCURACY OR RAND-INDEX

Quantify the similarity between two clusterings based on the confusion matrix. It represents the percentage of correct matches of documents in clusters, this is also known as accuracy. Rand-index is calculated according to the following formula [44]:

$$Rand - index = \frac{TP + TN}{TP + FP + FP + FN} \quad (9)$$

In complete similarity between clusterings data, the Rand-index has a value of 1, whereas in complete dissimilarity it has a value of 0.

7) JACCARD INDEX

This measures the similarity between two clusterings based on confusion matrix. This measure quantifies the similarity between data clustering and ground truth labels of data. It is defined as [45]:

$$Jaccard - index = \frac{TP}{TP + FN + FP} \quad (10)$$

Greater value means higher similarity between two clusterings.

IV. DATA PREPARATION AND METHODOLOGY

A. MAIN DATASET

The main dataset used in this study represents Modern Standard Arabic (MSA) news documents taken from [4]. It is available online at http://diab.edublogs.org/dataset-for-arabic-document-classification/. The dataset is composed of 2700 documents of 9 categories. Each category contains 300 documents. The categories are: Religion, Economy, Health, Politics, Law, Literature, Sports, Art and Technology.

The dataset has five versions as following:

1) V1: Documents with no preprocessing.
2) V2: Documents with stop words removed.
3) V3: Documents after stop words removed and stemmed with Light10 algorithm [46].
4) V4: Documents after stop words removed and stemmed with Chen's algorithm [47].
5) V5: Documents after stop words removed and stemmed with Khoja's algorithm [48].

TABLE 2. Details of the main Dataset.

Version	Nº Classes	Nº Docs.	Nº Terms	Nº Unique Terms	Doc. Avg. Length
V1	9	2,700	878,726	96,859	325
V2	9	2,700	600,627	89,757	222
V3	9	2,700	600,552	42,571	222
V4	9	2,700	600,477	30,488	222
V5	9	2,700	600,602	13,803	222

The dataset is preprocessed by removing diacritic marks, English words and numbers. Table 2 shows basic information about the dataset, for more information about the dataset please refer to [4] and [5].

This data will be used as the input for the implantation of LDA to reveal the main topics, and k-means algorithm is then used. Figures 1-2 illustrate the methodology of the study.

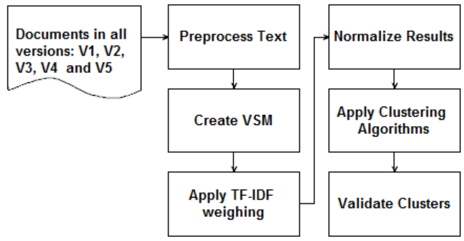

FIGURE 1. Algorithm for clustering documents.

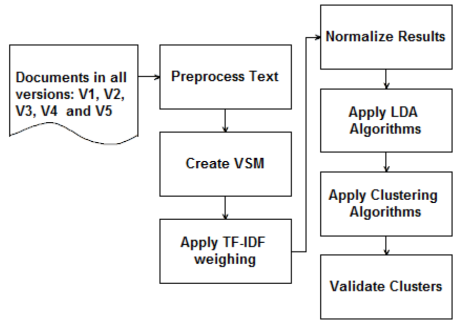

FIGURE 2. Algorithm for clustering topics.

B. METHODOLOGY

Initially, we preprocess the text by removing punctuation marks. On one hand, the main process for documents clustering is performed as shown in figure 1. First, we create Vector Space Model for the documents based on bag-of-words model. Then, Term Frequency - Inverse Document Frequency (TF-IDF) weighting is applied to the VSM. This removes unnecessary frequent terms that appears in most documents. Before clustering documents, data is mean-normalized so that Euclidean distance computes comparable results, this is calculated as follows:

$$m = \sqrt{\sum_{i=1}^{n} x_i^2} \quad (11)$$

Where n is the number of terms in each row in the DTM. Then each value in the row is divided by m. This is applied for each row.

Lastly, we apply k-means clustering algorithm on the normalized data in all versions, to have best results each clustering run starts from different 25 initial centers and the best result is then taken. On another setup, the normalization step is not performed.

On the other hand, the main process for topics clustering is performed as explained in figure 2. The same VSM model for documents clustering is used here after applying the TF-IDF weighting. Then data is mean-normalized, this is presented as an input to the LDA algorithms. Then, we use different algorithms (VEM, VEM fixed, Gibbs and CTM) to generate topic models for the data for all versions. The configuration for LDA algorithms are taken from [27]. After that, we mean-normalize the probabilities of topics in documents for all models. Again, the normalization step is needed so that Euclidean distance make sense while calculating distances between vectors and centers, this is achieved in the same way done previously. Finally, we apply k-means clustering algorithm on the normalized data for all versions, and again for each run of clustering, the best of the 25 runs starting from different initial centers is used.

The final step is then to evaluate clusters for both techniques: clustering documents and clustering topics. We compute purity, precision, recall, F-measure, entropy, NMI, NVI, accuracy and jaccard-index for the resulting clusters for all five versions. These results are used to compare quality of clusters for both techniques.

C. MORE DATASETS

To verify the correctness of the combined method, more datasets are used in this study. Table 3 shows the main information about these datasets.

TABLE 3. Details of the extra Datasets used in the experiments.

Dataset	Nº Classes	Nº Docs.	Nº Terms	Nº Unique Terms	Doc. Avg. Length
Aljazeera [49]	5	1,500	388,653	50,099	259
Alkhaleej [50]	4	5,690	2,472,763	122,162	435
Alwatan [50]	6	20,291	9,876,786	261,909	487
BBC [50]	7	4,763	1,794,123	88,953	377
CNN [50]	6	5070	2,166,109	105,047	427

These datasets are freely available on the internet and all represent news articles in MSA form. The same aforementioned methodology will be applied on these datasets, except for both Alkhaleej and Alwatan. Due to memory limitations, in these two datasets, the top-ranked terms are used based on TF-IDF weighting. All terms with TF-IDF weight above the third-quartile statistics are selected and used in the experiments.

V. EXPERIMENTS AND RESULTS

A. CALCULATING CLUSTERS WITHOUT NORMALIZATION

The first set of experiments calculate clusters using k-means algorithm. The configuration of the experiments follow the methodology specified in figure 1 except that **normalization phase is not performed**. The number of clusters is 9, for each version of the dataset the experiment is repeated 20 times and then the average as well as standard deviation are computed. These results -in terms of purity and entropy- are shown in table 4 for all versions of the text documents. Note that only purity and entropy calculated here for the purpose of comparison with recent similar study. The other validation measures are computed later.

B. CALCULATING CLUSTERS WITH NORMALIZATION

The second set of experiments calculate clusters using k-means algorithm but with normalization applied to the data. The configuration of the experiments follow exactly the methodology that is specified in figure 1. The number of clusters is 9, for each version of the dataset the experiment is repeated 20 times and then the average as well as standard deviation are computed. These results -in terms of purity and entropy- are shown in table 5 for all versions of the text documents. Again, the other validation measures are computed later due to the aforementioned reason in previous subsection.

TABLE 4. Average values and standard deviation for purity and entropy using k-means algorithm without normalization.

Version	Avg. Purity	Std. Dev.	Avg. Entropy	Std. Dev.
V1	0.5357	0.0422	0.4908	0.0462
V2	0.5267	0.0499	0.5028	0.0560
V3	0.5849	0.0634	0.4249	0.0735
V4	0.5890	0.0444	0.4344	0.0446
V5	0.6289	0.0552	0.3885	0.0447

TABLE 5. Average values and standard deviation for purity and entropy using k-means algorithm with Normalization.

Version	Avg. Purity	Std. Dev.	Avg. Entropy	Std. Dev.
V1	0.7450	0.0106	0.2970	0.0127
V2	0.7457	0.0172	0.2931	0.0231
V3	0.8058	0.0131	0.2294	0.0142
V4	0.8057	0.0103	0.2314	0.0125
V5	0.8232	0.0100	0.2158	0.0223

TABLE 6. Matching-based Evaluation measures for main dataset.

Dataset	Purity	Precision	Recall	F-measure
V1 K-means	0.7478	0.4617	0.77	0.5773
V1 Combined	0.8807	0.7804	0.786	0.7832
V2 K-means	0.7415	0.4536	0.7692	0.5707
V2 Combined	**0.9252**	0.8546	0.8635	**0.8590**
V3 K-means	0.8030	0.5772	0.8136	0.6753
V3 Combined	0.8926	0.8042	0.8103	0.8072
V4 K-means	0.8026	0.5786	0.8129	0.6760
V4 Combined	**0.9233**	0.8558	0.8607	**0.8582**
V5 K-means	0.8215	0.6065	0.7917	0.6869
V5 Combined	**0.9330**	0.8713	0.8752	**0.8732**

C. CALCULATING CLUSTERS BASED ON TOPICS

The third set of experiments compute the topic models with 9 topics, this number represents the number of clusters for the dataset. After that, the resulted probabilities for words on the topics are used as an input to the k-means algorithm. This is applied for different LDA models: VEM, fixed VEM, Gibbs and CTM. The methodology used in these experiments follow what is shown in figure 2. Again, the k-means algorithm is repeated 20 times for all versions of the dataset and then the average as well as standard deviation are computed.

TABLE 7. Entropy-based evaluation measures for main dataset.

Dataset	Entropy	Normalized Mutual Information	Normalized variation of information
V1 K-means	0.1774	0.7513	0.3983
V1 Combined	0.2356	0.7634	0.3827
V2 K-means	0.1789	0.7468	0.4041
V2 combined	0.1473	**0.8515**	0.2587
V3 K-means	0.1556	0.7985	0.3354
V3 Combined	0.1924	0.8066	0.3241
V4 K-means	0.1587	0.7955	0.3395
V4 Combined	0.149	**0.8503**	0.2605
V5 K-means	0.1762	0.7874	0.3507
V5 Combined	0.1384	**0.8611**	0.2440

TABLE 8. Pairwise evaluation measures for main dataset.

Dataset	Accuracy (Rand-index)	Jaccard index
V1 K-means	0.8751	0.4058
V1 Combined	0.9518	0.6436
V2 K-means	0.8718	0.3992
V2 Combined	**0.9686**	0.7529
V3 K-means	0.9133	0.5098
V3 Combined	0.9571	0.6768
V4 K-means	0.9137	0.5106
V4 Combined	**0.9685**	0.7517
V5 K-means	0.9200	0.5231
V5 C1ombined	**0.9718**	0.775

TABLE 9. Matching-based Evaluation measures for other datasets.

Dataset	Purity	Precision	Recall	F-measure
Aljazeera K-means	0.6927	0.444	0.7259	0.551
Aljazeera combined	**0.8993**	0.8154	0.8171	**0.8163**
Alkhaleej K-means	0.4580	0.3039	0.8171	0.4431
Alkhaleej combined	**0.8534**	0.7908	0.7124	**0.7495**
Alwatan K-means	0.2565	0.1754	0.8884	0.2929
Alwatan combined	0.6403	0.5113	0.4982	0.5047
BBC K-means	0.5581	0.3725	0.1828	0.2452
BBC combined	0.5860	0.4895	0.2121	0.296
CNN K-means	0.4493	0.2063	0.5201	0.2954
CNN combined	0.5943	0.4366	0.3818	0.4074

Evaluation of the results for all experimentation setups is applied according to section IV. Tables 6-8, show the results for all validation measures for the combined results compared with those for k-means algorithm.

D. VERIFICATION EXPERIMENTS

In this subsection more experiments are conducted on other datasets. This is to make sure that previous results are consistent and our methodology extends to different datasets.

TABLE 10. Entropy-based evaluation measures for other datasets.

Dataset	Entropy	Normalized Mutual Information	Normalized variation of information
Aljazeera K-means	0.2607	0.6221	0.5485
Aljazeera combined	0.2416	**0.7580**	0.3897
Alkhaleej K-means	0.2806	0.0985	0.9482
Alkhaleej combined	0.3185	**0.6880**	0.4756
Alwatan K-means	0.1351	0.0593	0.9694
Alwatan combined	0.5431	0.4544	0.7060
BBC K-means	0.8467	0.1213	0.9354
BBC combined	0.7908	0.2318	0.8689
CNN K-means	0.4209	0.2682	0.8452
CNN combined	0.6840	0.3206	0.8091

TABLE 11. Pairwise evaluation measures for other datasets.

Dataset	Accuracy (Rand-index)	Jaccard index
Aljazeera K-means	0.764	0.3803
Aljazeera combined	**0.9266**	0.6896
Alkhaleej K-means	0.3956	0.2846
Alkhaleej combined	**0.8599**	0.5994
Alwatan K-means	0.2456	0.1716
Alwatan combined	0.828	0.3375
BBC K-means	0.6046	0.1398
BBC combined	0.6454	0.1737
CNN K-means	0.5237	0.1733
CNN combined	0.7868	0.2558

TABLE 12. Comparing purity values for K-means in this study and K-means in [5].

Version	Purity (This Study)	Purity (as in [5])
V1	0.75	0.11
V2	0.75	0.11
V3	0.81	0.25
V4	0.81	0.30
V5	0.82	0.43

The resulted clusterings are then validated using different external measures as discussed previously, the results are shown in tables 9-11.

VI. DISCUSSION

In this study the k-means algorithm achieved better results than those reported in [5] on the same dataset. This is due to two reasons: first, the TF-IDF values in DTM are mean normalized and second, each run represents the best of runs which start from different 25 initial centers. These together increased the purity dramatically. Table 12 shows a comparison between results of applying K-means algorithm on all versions using our methodology and those resulted from applying K-means algorithm in [5].

The results shown in previous section indicate that the quality of clusters using clustering algorithms alone is inferior. This is due to several reasons including curse of dimensionality. In this study, the dimensions of the VSM are in thousands. These are very sparse and high dimensional matrices. In such cases, there are available different solutions including Singular Value Decomposition (SVD), Latent Semantic Analysis (LSA) [51] and subspace clustering [52], [53]. In all these methods, the main point is to reduce dimensionality but preserve, hopefully, representative dimensions. Another technique that could be used to achieve dimensionality reduction is LDA. This technique achieves two roles: reduces the number of dimensions and uncovers latent semantic variables in text documents.

Tables 6-8 show the quality of clusters according to several measures for all versions of the text. These results suggest two things; the combined algorithm is superior to k-means algorithm, and that text in V5 gives the best results with purity of **0.9330** and F-measure of **0.8732**. These are much better than results of k-means algorithm, where purity is **0.8215** and F-measure is **0.6869**. Results are clearly confirmed by the other measures.

The best results for external measures are achieved when V5 and V4 are used. Text in V5 represents text processed by removing stopwords and then stemmed with Khoja's algorithm, which is a root-based stemmer. Also, V4 is the same but is stemmed with Chen's algorithm. Notwithstanding, the analysis of the effect of stemming algorithms on clustering is out of the scope of this study. However, this is discussed in different research papers that study the effect of stemming on the performance of clustering or classification on Arabic documents [54]–[56].

One can notice that V2 -which represents text preprocessed by removing stopwords only and no stemming applied- has near best results with purity of **0.9252** and F-measure of **0.8590**. This suggests that using Gibbs Sampling gives very good results for text in original format with stopwords only being removed.

The combined method is applied to other Arabic datasets, and the same previous results on the main dataset are confirmed as shown in tables 9-11. It is clear from these results that the combined algorithm attains much better results even when applied to different datasets.

It is vital here to stress that the combined algorithm may not be applied on short text documents. This is because short text lacks enough content and hence has its special techniques and methods in processing [57]. In such cases more NLP techniques [58], [59] are used, also text expanding [60], [61] is used to overcome the shortage in shared features which help much in the clustering algorithms.

The combined algorithm has achieved excellent results based on the simple K-means algorithm combined with LDA and using simple Euclidean distance compared with similar studies that use different distance and similarity measures and sophisticated clustering algorithms [5], [7], [12].

Although the methodology used in this study is simple, however it achieves a much better clustering results compared to k-means algorithm. Especially, mean normalization of the TF-IDF weights in VSM enhance the results dramatically. Also, Applying Topic modeling first on the datasets served as both feature-selection and feature-reduction. These are very important in data mining applications and algorithms including clustering.

VII. CONCLUSION

Regular clustering algorithms might not give good results due to at least the high dimensionality nature of text. Therefore, this study utilizes a combined solution for Arabic text using clustering algorithms and topic modeling techniques.

Clustering Arabic text documents is a challenging task due to several reasons, as mentioned in the introduction. In this study, we show that the quality of clusters for Arabic text documents is dramatically improved by exploiting topic modeling techniques in the clustering process based on external clustering measures.

This study uses news text dataset composed of five versions. This is used in evaluating both k-means clustering algorithm and topic modeling/k-means combined method.

The results of this study emphasize that plugging in normalization in the VSM enhances the results of the simple K-means algorithm with the simple Euclidean measure compared with similar study.

Also, the results of this study show that the combined method gives much better results compared with simple K-means algorithm. This is confirmed by the results of experiments conducted on other five datasets.

Working with Arabic text, although challenging, but still there is a large space for improvement and development. Future work might include building a word embedding model for Arabic Text. This task needs a large size of text in order to give acceptable results. Existing models such as word2vec or GloVec have been applied successfully to some languages including English and gave reasonable results.

REFERENCES

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, pp. 14:1–14:22, Dec. 2009.

[2] M. Saad and W. Ashour, "Arabic morphological tools for text mining," in *Proc. 6th Int. Symp. Elect. Electron. Eng. Comput. Sci.*, Lefke, Cyprus, 2010, pp. 112–117.

[3] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling," *CoRR*, Sep. 2013.

[4] D. Abuaiadah, J. E. Sana, and W. Abusalah, "Article: On the impact of dataset characteristics on arabic document classification," *Int. J. Comput. Appl.*, vol. 101, no. 7, pp. 31–38, Sep. 2014.

[5] D. Abuaiadah, "Using bisect k-means clustering technique in the analysis of arabic documents," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 15, no. 3, pp. 17:1–17:13, Jan. 2016.

[6] H. S. Al-Sarrayrih and R. Al-Shalabi, "Clustering arabic documents using frequent itemset-based hierarchical clustering with an N-grams," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, Jun. 2009.

[7] H. Froud and A. Lachkar, "Agglomerative hierarchical clustering techniques for arabic documents," in *Proc. 3rd Int. Conf. Comput. Sci., Eng. Inf. Technol. (CCSEIT)*, Konya, Turkey, vol. 1, D. Nagamalai, A. Kumar, and A. Annamalai, Eds. Berlin, Germany: Springer, 2013, pp. 255–267.

[8] S. H. Ghwanmeh, "Applying clustering of hierarchical k-means-like algorithm on arabic language," *Int. J. Inf. Technol.*, vol. 3, no. 3, pp. 168–172, 2007.
[9] H. Sawaf, J. Zaplo, and H. Ney, "Statistical classification methods for arabic news articles," in *Proc. Arabic Natural Lang. Process. (ACL)*, 2001.
[10] M. El-Haj, U. Kruschwitz, and C. Fox, "Exploring clustering for multi-document arabic summarisation," in *Information Retrieval Technology*, M. V. M. Salem, K. Shaalan, F. Oroumchian, A. Shakery, and H. Khelalfa, Eds. Berlin, Germany: Springer, 2011, pp. 550–561.
[11] M. Hussein, A. Alsammak, and T. Elshishtawy, "Keyphrase-based hierarchical clustering for arabic documents," in *Proc. 10th Int. Conf. Inform. Syst. (INFOS)*, 2016, pp. 61–67.
[12] H. Froud, R. Benslimane, A. Lachkar, and S. A. Ouatik, "Stemming and similarity measures for arabic documents clustering," in *Proc. 5th Int. Symp. I/V Commun. Mobile Netw.*, Sep. 2010, pp. 1–4.
[13] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.
[14] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning*, T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Eds. Mahwah, NJ, USA: Laurence Erlbaum, 2006.
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[16] D. Blei and J. Lafferty, "Correlated topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18, 2006, p. 147.
[17] I. Biro, "Document classification with latent Dirichlet allocation," Ph.D. dissertation, Eötvös Loránd Univ., Budapest, Hungary, 2009.
[18] P. Crossno, A. Wilson, T. Shead, and D. Dunlavy, "TopicView: Visually comparing topic models of text collections," in *Proc. 23rd IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2011, pp. 936–943.
[19] A. Kelaiaia and H. Merouani, "Clustering with probabilistic topic models on arabic texts," in *Modeling Approaches and Algorithms for Advanced Computer Applications* (Studies in Computational Intelligence), vol. 488, A. Amine, A. M. Otmane, and L. Bellatreche, Eds. Springer, 2013, pp. 65–74.
[20] G. K. Gerber, R. D. Dowell, T. Jaakkola, and D. K. Gifford, "Automated discovery of functional generality of human gene expression programs," *PLoS Comput. Biol.*, vol. 3, no. 8, p. e148, 2007.
[21] J. Boyd-Graber, D. M. Blei, and X. Zhu, "A topic model for word sense disambiguation," in *Empirical Methods in Natural Language Processing*. 2007.
[22] S. Gerrish and D. M. Blei, "Predicting legislative roll calls from text," in *Proc. ICML*, L. Getoor and T. Scheffer, Eds. Madison, WI, USA: Omnipress, 2011, pp. 489–496.
[23] W. M. Darling, "A theoretical and practical implementation tutorial on topic modeling and Gibbs sampling," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, Dec. 2011, pp. 642–647.
[24] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.
[25] R. Ayadi, M. Maraoui, and M. Zrigui, "Latent topic model for indexing arabic documents," *Int. J. Inf. Retr. Res.*, vol. 4, no. 1, pp. 29–45, 2014.
[26] M. A. Siddiqui, S. M. Faraz, and S. A. Sattar, "Discovering the thematic structure of the quran using probabilistic topic model," in *Proc. Taibah Univ. Int. Conf. Adv. Inf. Technol. Holy Quran Sci.*, Dec. 2013, pp. 234–239.
[27] M. Alhawarat, "Extracting topics from the holy Quran using generative models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, pp. 288–294, Dec. 2015.
[28] A. Brahmi, A. Ech-Cherif, and A. Benyettou, "Arabic texts analysis for topic modeling evaluation," *Inf. Retr.*, vol. 15, no. 1, pp. 33–53, Feb. 2012.
[29] A. Kelaiaia and H. F. Merouani, "Clustering with probabilistic topic models on arabic texts: A comparative study of lda and k-means," *Int. Arab J. Inf. Technol.*, vol. 13, no. 2, pp. 332–338, 2016.
[30] T. Tarczynski, "Document clustering-concepts, metrics and algorithms," *Int. J. Electron. Telecommun.*, vol. 57, no. 3, pp. 271–277, 2011.
[31] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. San Diego, CA, USA: Academic, 2008.
[32] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.
[33] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Inf. Retr.*, vol. 12, no. 4, pp. 461–486, 2009.
[34] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Jan. 2010.
[35] S. Wagner and D. Wagner, "Comparing clusterings: An overview," Ph.D. dissertation, Univ. Karlsruhe, Fakultät Inf. Karlsruhe, Germany, 2007.
[36] A. N. Albatineh and M. Niewiadomska-Bugaj, "Correcting Jaccard and other similarity indices for chance agreement in cluster analysis," *Adv. Data Anal. Classification*, vol. 5, no. 3, pp. 179–200, 2011.
[37] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge Univ. Press, 2008.
[38] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.
[39] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*, vol. 39. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[40] N. Chinchor and B. Sundheim, "MUC-5 evaluation metrics," in *Proc. 5th Conf. Message Understanding*, 1993, pp. 69–78.
[41] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 623–656, Jul./Oct. 1948.
[42] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.
[43] R. Reichart and A. Rappoport, "The NVI clustering evaluation measure," in *Proc. 13th Conf. Comput. Natural Lang. Learn.*, 2009, pp. 165–173.
[44] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.
[45] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
[46] L. S. Larkey, L. Ballesteros, and M. E. Connell, *Light Stemming for Arabic Information Retrieval*. Dordrecht, The Netherlands: Springer, 2007, pp. 221–243.
[47] A. Chen and F. C. Gey, "Building an arabic stemmer for information retrieval," in *Proc. TREC*, 2002, pp. 631–639.
[48] S. Khoja and R. Garside, "Stemming arabic text," M.S. thesis, Dept. Comput. Lancaster Univ., Lancaster, U.K., 1999.
[49] D. Said, N. M. Wanas, N. M. Darwish, and N. Hegazy, "A study of text preprocessing tools for arabic text categorization," in *Proc. 2nd Int. Conf. Arabic Lang.*, 2009, pp. 230–236.
[50] M. K. Saad and W. Ashour, "OSAC: Open source arabic corpora," in *Proc. 6th ArchEng Int. Symp., (EEECS)*, vol. 10, 2010.
[51] S. T. Dumais, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2004.
[52] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *ACM SIGMOD Rec.*, vol. 27, no. 2, pp. 94–105, Jun. 1998,
[53] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 90–105, Jun. 2004.
[54] Q. W. Bsoul and M. Mohd, "Effect of isri stemming on similarity measure for arabic document clustering," in *Proc. Asia Inf. Retr. Symp.* Springer, 2011, pp. 584–593.
[55] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Stemming as a feature reduction technique for arabic text categorization," in *Proc. 10th Int. Symp. Program. Syst. (ISPS)*, 2011, pp. 128–133.
[56] R. Duwairi, M. Al-Refai, and N. Khasawneh, "Stemming versus light stemming as feature selection techniques for arabic text categorization," in *Proc. 4th Int. Conf. Innov. Inf. Technol. (IIT)*, 2007, pp. 446–450.
[57] D. Pinto, J.-M. Benedí, and P. Rosso, "Clustering narrow-domain short texts by using the Kullback-Leibler distance," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Berlin, Germany: Springer, 2007, pp. 611–622.
[58] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2009, pp. 919–928.
[59] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 775–784.
[60] P. Makagonov, M. Alexandrov, and A. Gelbukh, "Clustering abstracts instead of full texts," in *Proc. Int. Conf. Text, Speech Dialogue*. Springer, 2004, pp. 129–135.
[61] D. Pinto, P. Rosso, and H. Jiménez-Salazar, "A self-enriching methodology for clustering narrow domain short texts," *Comput. J.*, vol. 54, no. 7, pp. 1148–1165, 2011.

M. ALHAWARAT was born in Jordan in 1975. He received the B.Sc. degree (Hons.) in computer science from the University of Mu'tah, Jordan, in 1997, and the Ph.D. degree in chaotic neural networks from the School of Technology, Oxford Brookes University, U.K. Then, he was a Programmer and Web Developer with the Computer Center, University of Mu'tah, from 1997 to 1999. After that, he was an Oracle Developer with the IT Department of a private hospital in Saudi Arabia from 1999 to 2000. Then, he was a Team Leader and System Analyst in the Computer Center, Petra University, Jordan, from 2000 to 2003, where he was an Assistant Professor for one year. During the same period, he has been awarded a scholarship from Petra University to pursue his higher studies in U.K. Then, in 2008, he joined the College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Saudi Arabia, where he become an Associate Professor in 2015. His research interests include chaotic neural networks, Arabic natural language processing, text mining, and machine learning. He published several research papers mainly in the field of Arabic NLP. He has been a Professional Member of the ACM since 2013.

M. HEGAZI received the Ph.D. degree from the Sudan University of Science and Technology, Khartoum, Sudan, in 2004. His Ph.D. thesis was entitled—An Approach for Heterogeneous Distributed Database Systems Integration. He is currently a Professor of computer science with the Department of Computer Science, Prince Sattam Bin Abdulaziz University. His research interests are database, data mining, data science, and computer applications.