Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 3165 – 3179

doi:10.3906/elk-1808-103

Exploring bigram character features for Arabic text clustering

Dia ABUZEINA*
Department of Computer Science and Information Technology College of Information Technology and Computer Engineering Palestine Polytechnic University, Hebron, Palestine

Received: 12.08.2018 • **Accepted/Published Online:** 06.03.2019 • **Final Version:** 26.07.2019

Abstract: The vector space model (VSM) is an algebraic model that is widely used for data representation in text mining applications. However, the VSM poses a critical challenge, as it requires a high-dimensional feature space. Therefore, many feature selection techniques, such as employing roots or stems (i.e. words without infixes and prefixes, and/or suffixes) instead of using complete word forms, are proposed to tackle this space challenge problem. Recently, the literature shows that one more basic unit feature can be used to handle the textual features, which is the two-neighboring character form that we call microword. To evaluate this feature type, we measure the accuracy of the Arabic text clustering using two feature types: the complete word form and the microword form. Hence, the microword is two consecutive characters which are also known as the Bigram character feature. In the experiment, the principal component analysis (PCA) is used to reduce the feature vector dimensions while the k-means algorithm is used for the clustering purposes. The testing set includes 250 documents of five categories. The entire corpus contains 54,472 words, whereas the vocabulary contains 13,356 unique words. The experimental results show that the complete word form score accuracy is 97.2% while the two-character form score is 96.8%. In conclusion, the accuracies are almost the same; however, the two-character form uses a smaller vocabulary as well as less PCA subspaces. The study experiments might be a significant indication of the necessity to consider the Bigram character feature in the future text processing and natural language processing applications.

Key words: Arabic, text, clustering, features, dimensionality reduction, k-means, principal component analysis, vector space model

1. Introduction

The big data environment with the massive growth of unlabeled data promotes ongoing research to organize or filter data through unsupervised machine learning techniques. Text clustering is an automatic process to partition data into several groups such that each text is assigned to a group based on the similarities of the linguistic features. Intuitively, the texts of the same group are similar while the texts of different groups are dissimilar. With the current huge digital data, text clustering becomes a basic component in many data mining and natural language processing (NLP) tools. For instance, it is used in the monitoring system to detect duplicate content to identify plagiarism. Other applications include information retrieval, recommendation systems, document organization, opinion mining, and feedback analysis.

Text clustering-based applications generally employ the vector space model (VSM) that was proposed in [1]. The VSM represents text as a "bag of words" in which each word corresponds to one independent

*Correspondence: abuzeina@ppu.edu

dimension in the feature vectors. The VSM model is then used as an input to the clustering machine learning algorithms. The standard feature unit of the VSM is the complete word form (also known as the full-form word); however, many other features are used, such as roots, stems, or n-grams, character trigrams. In fact, the type of textual feature that is used is a wide research area which aims at finding the best one that best fits the needs. Recently, it was indicated in [2] that two-neighboring characters is a promising feature type for text classification problems. In this work, we call this "small" word feature "microwords." Accordingly, the goal of this work is to evaluate the proposed microword features by comparing it with the complete word forms for Arabic text clustering. For instance, the microwords of the word "mercy" are "me", "er", "rc", "cy". The interesting use of microwords is that it has a fixed size vocabulary as the Arabic alphabet has 28 letters. Hence, the maximum size of the vocabulary is 28×28 microwords. The standard data mining applications generally use complete word form vocabularies that might contain thousands of words. Hence, employing a space-independent method is extremely important to reduce space and time complexity in data processing. This work examines a new word representation that transforms text into a meaningful little representation of numbers.

To address the intended objective, we employ k-means clustering algorithm along with the principal component analysis (PCA) for dimensionality reduction. PCA is widely used in many scientific areas such as information retrieval and image processing. In the experiments we used, there is a data collection that contains 250 documents of five categories. The experimental results show that the accuracy is almost the same; however, the microwords type used less vocabulary size as well as less PCA subspaces. These outputs were evaluated using a wide range of parameters of PCA values and document frequency (DF) as a feature selection method. In particular, the experimental results show that the accuracy using the complete word form is 97.2% while it is 96.8% using microwords. In particular, the vocabulary size of the complete word form method is 1337 words while it is 301 in the microword approach. In addition, the microword form used less principal components as it used 46 while the complete word form used 67. This is an indication that the microword approach is a promising approach as it reduces the modelling complexity with almost the same performance.

This paper is organized as follows: Section 2 presents the k-means algorithm followed by the background of PCA in Section 3. In Section 4, we present document-term matrix and two worked PCA examples in Section 5. The literature review is demonstrated in Section 6 and the proposed method is in Section 7. The experimental results are presented in Section 8 and the discussion in Section 9. Finally, we conclude in Section 10.

2. K-means clustering algorithm

K-Means is one of the best-known flat (i.e. nonhierarchical) clustering methods which are widely used in data science. It is an unsupervised machine learning technique that is widely used for clustering purposes. It is a simple, fast, and relatively efficient algorithm. K-means is generally available as a package in many programming languages such as Python as well as in many toolkits, such as WEKA, Orange, and RapidMiner. Unlike a supervised machine learning scheme, unsupervised learning such as k-means eliminates the teacher as it does not require correct answers for the observations; hence, the data samples are unlabeled. The task of the k-means is to discover the structure of the input data in order to cluster the observations into a predefined k group (k is a user-defined value of the selected number of clusters). That is, k-means requires defining the number of groups in advance. Other clustering methods include expectation–maximization (EM), self-organizing map (SOM), and agglomerative (i.e. for hierarchical clustering).

To perform clustering, k-means implements two steps given in initial random guesses of the k centroids. The first step is the assignment process to assign the observations to the clusters while the second step is

the update process to find the new centroids based on the new members. For each text feature vector, the distance is measured to find the nearest centroid and to assign the text to that centroid. The centroids are then updated based on the new members. This process is repeated for a predefined number of iterations until there are no further updates for the centroids. This is called the convergence which means that there is no more change in the clusters centroid. Finally, the algorithm's output consists of the final values of the centroids. The classical k-means algorithm uses the Euclidean distance to construct dissimilarities between quantitative variables. However, other distance measures can be used such as the Cosine similarity measure. In this work, we employ k-means since it is suitable to fulfill the goal of this work to explore the performance of the two investigated features: the complete word form and the microword form.

3. Principal component analysis

Text mining applications are characterized by huge vocabularies that are required to use dimensionality reduction techniques to transform the high-dimensional data into much lower dimensional data. PCA (also called the Karhunen Loeve or K-L method) is a well-known statistical unsupervised feature reduction method. The goal of the PCA is to extract the most important "compressed" information while hopefully preserving the most characteristic features of the original data in the constructed PCA subspace. PCA is beneficial in many applications especially that contain high-dimensional multivariate data, such as facial recognition and geographical data applications. In simple words, PCA aims at finding a small number of dimensions of underlying variables that describe the data. Figure 1 shows the concept of dimension reduction as the feature space is reduced into a lower feature space. As shown in the figure, the dimension of the reduced features is less than or equal to the original dimension (i.e. 2, which is less than or equal to 7).

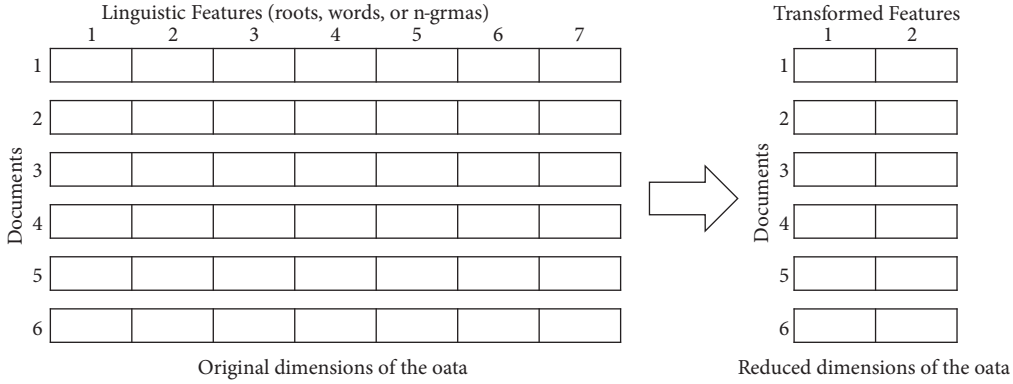

Figure 1. The input space transformation into a lower dimension.

The basis of PCA is to project data to a lower dimensional subspace to facilitate the next processing steps such as text clustering. This process starts by centering the data by subtracting off the mean (the data mean becomes zero) and then finding the direction of the largest variation. The essence is that the large variation demonstrates much variability, which is the core requirement for a successful classification. That is, exposing the variability as well as eliminating some noises (e.g., outliers) is the major goal of the PCA algorithm. The first and largest variation corresponds to the first principal component, which is used to place the direction of the first axis. Similarly, the second principal component is the second largest variation which is also used to place the direction of the second axis, and so on. In fact, the useful information is concentrated in the first components, which means that some of the less important dimensions can be ignored. For an illustration, the

new axes (x', y') in Figure 2 contain the first and the second components based on the largest variations in the original data.

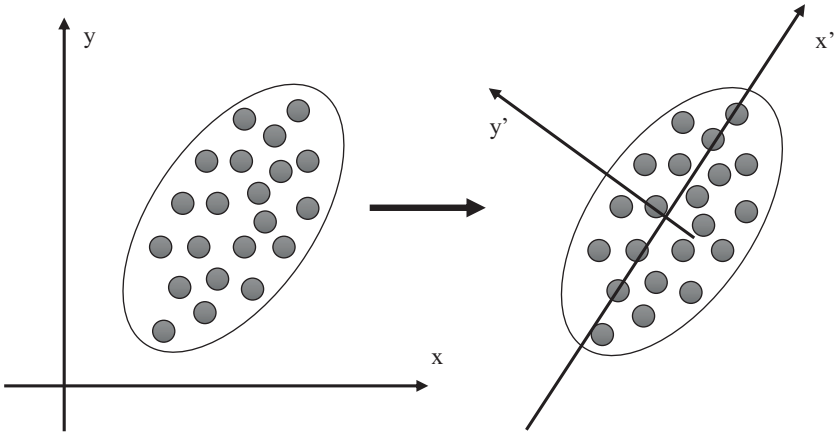

Figure 2. The directions of the largest variations.

To find the largest variations, the PCA finds the Eigenvalues and the corresponding Eigenvectors of the data covariance matrix. The largest Eigenvalues are used to find the Eigenvectors for PCA subspaces. Finally, the user chooses the number of dimensions in the dimensionally reduced subspace. Figure 3 shows the subspace of 500 real values that are transformed into one-dimension features. In the figure, the small point markers are the transformed data that is used for further machine learning tasks. The transformed data is obtained using the dot product of the Eigenvector by the centered data. The next section demonstrates how to compute and plot the transformed data.

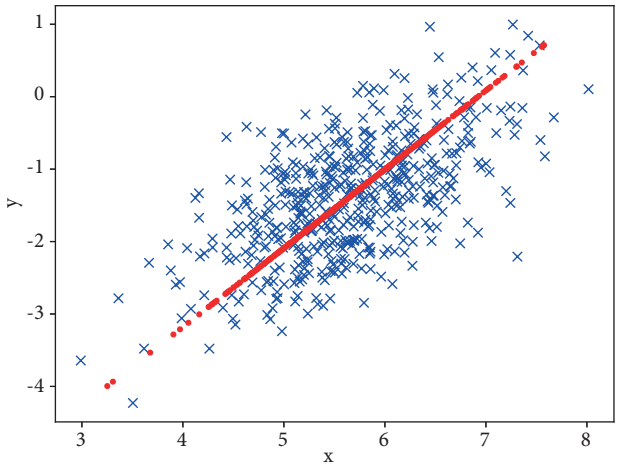

Figure 3. Reconstructed data after PCA for two-dimensions.

4. Document-term matrix

A text clustering system must pass through a set of steps. An initial step is to represent the text as feature vectors based on the data vocabulary (words or terms in the standard VSM). Hence, the word dictionary is employed to form the textual features of the data. The textual feature vectors are generally arranged in a matrix

that is called document-term or term-document matrix. However, in document-term matrix the rows correspond to documents and columns correspond to words. The process of creating a document matrix considers different aspects, such as a stop words list, ignore characters list, weighting schemas, etc. Creating the document-term matrix is a simple function that aims at converting the data collection into a matrix that contains the frequencies or the weights of all words. This study considers two different word forms of textual features: the complete form and the microword form. To illustrate the shape of the document-matrix of the two types, we formulated two matrices of a simple data collection that contained the following five popular English proverbs:

1) A window of opportunity will not open itself
2) in the middle of difficulty lies opportunity
3) luck is a matter of preparation meeting opportunity
4) imagination rules the world
5) the man who has no imagination has no wings

To create the document-matrix of the five sentences above, it was necessary to create the vocabulary. It contains 25 words as the following sorted list (NOTE: logistically, we chose NOT to use the word "a" or "A" – there are actually 26 words in the above sentences/proverbs): ['difficulty', 'has', 'imagination', 'in', 'is', 'itself', 'lies', 'luck', 'man', 'matter', 'meeting', 'middle', 'no', 'not', 'of', 'open', 'opportunity', 'preparation', 'rules', 'the', 'who', 'will', 'window', 'wings', 'world']. Now creating the document-term matrix includes creating a matrix of 5 $rows \times 25$ $words$ as shown in Figure 4.

		Terms																								
Documents	#	difficulty																								world
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	1	0	0
	2	1	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0
	3	0	0	0	0	1	0	0	1	0	1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0
	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
	5	0	2	1	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	1	1	0	0	1	0

Figure 4. Document-term matrix of complete words form.

Creating the document-term matrix for the microword form is almost the same but the basic feature unit is different. In this approach, the vocabulary contains microwords as shown in Figure 5. Figure 5 also contains the document-term matrix.

```
[ag,an,ar,as,at,ck,cu,dd,di,dl,do,ee,el,en,ep,er,es,et,ff,fi,
gi,gs,ha,he,ho,ic,id,ie,if,il,im,in,io,is,it,ld,le,lf,li,ll,
lt,lu,ma,me,mi,na,nd,ng,ni,no,of,on,op,or,ot,ow,pa,pe,po,pp,
pr,ra,re,rl,rt,ru,se,te,th,ti,ts,tt,tu,ty,uc,ul,un,wh,wi,wo]
```

		Microwords									
Documents	#	ag	an	ar	as	at	…	un	wh	wi	wo
	1	0	0	0	0	0		1	0	2	0
	2	0	0	0	0	0		1	0	0	0
	3	0	0	1	0	2		1	0	0	0
	4	1	0	0	0	1		0	0	0	1
	5	1	1	0	2	1		0	1	1	0

Figure 5. Document-term matrix of the microword form.

5. PCA worked examples

In this section, we present two worked examples that exhibit how to extract PCA subspaces. The first example refers to a two-dimension case of 100 observations while the second example considers a three-dimension case of 15 observations. The worked examples follow the PCA algorithm steps below:

1) Create the document-term matrix.

2) Center the document-term matrix.

3) Estimate the covariance matrix (C) of the centered document-matrix matrix.

4) Perform eigenvalues and eigenvectors decomposition of C.

5) Sort the eigenvalues in descending order to choose the largest eigenvalues.

6) Based on the selected largest eigenvalues that represent the required number of dimensions, use the corresponding eigenvectors to create the transformed data in the reduced subspace.

Figure 6 shows how to find the principal components using the PCA algorithm. PCA requires an unlabeled data collection as the input data. Using the input original data, the mean is computed to center the data by subtracting off the mean. Based on the Centered Data Matrix, the covariance matrix is computed to be used for Eigen decomposition. The generated Eigen value and Eigen value are used to find the direction and the values of the transformed data. Figure 6 shows how to find the first principle component for the two-dimensional data.

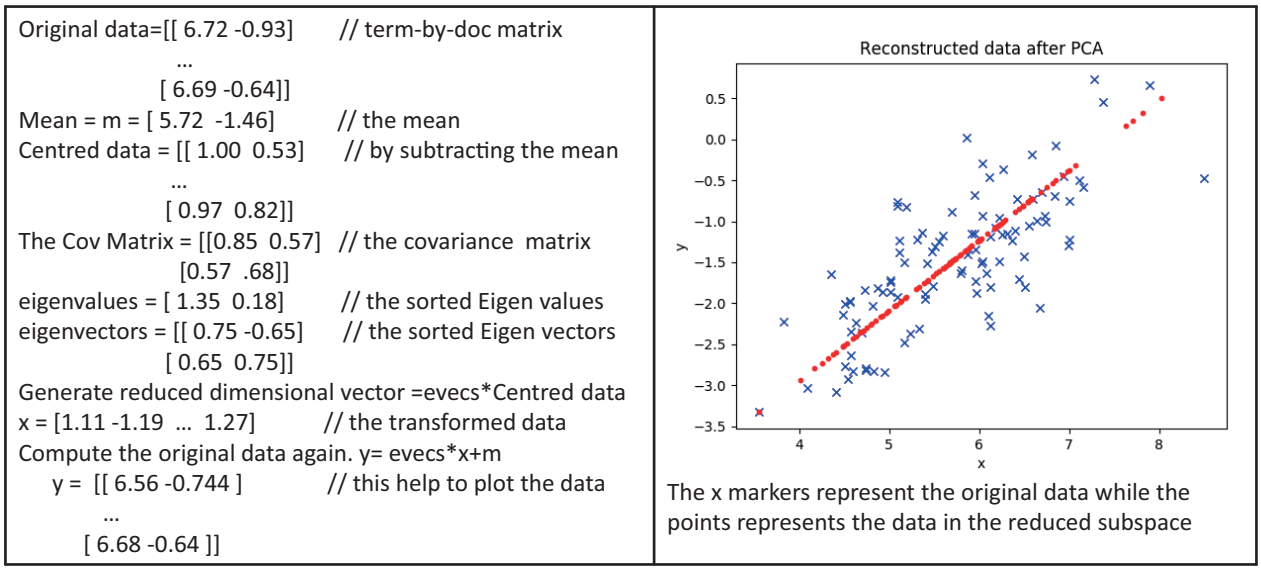

Figure 6. A worked example of PCA for two dimensions.

The second example handles three-dimension feature vectors. The previous example extracts one dimension while this example extracts two dimensions of the original three-dimensions data. The original data consists of 15 observations. The PCA algorithm is implemented as shown in Figure 7.

To visualize the effect of PCA on three attribute data, Figure 8 presents a scatter plot of three-dimensional (3D) space. The figure shows the original data and the transformed data for one, two, and three PCA components. The three components give the original dataset. It is worth to indicate that the original data is just random numbers of three attributes (i.e. three dimensions).

Data=	**Centred data =**			**Covariance Matrix =**			**Transformed data**		**Original Data** = (evecs*x+m)		
[[1 2 1]	[[-7.27	-5.93	-6.53]	[[20.78	20.66	19.20]	=		[[1.54	1.83	0.62]
[2 3 3]	[-6.27	-4.93	-4.53]	[20.66	23.21	19.68]	[[-11.4	-1]	[2.99	2.69	2.31]
[4 1 4]	[-4.27	-6.93	-3.53]	[19.20	19.68	19.70]]	[-9.1	-0.3]	[3.77	1.07	4.16]
[5 4 3]	[-3.27	-3.93	-4.53]	**Eigenvalues** =			[-8.6	2.2]	[4.29	4.22	3.49]
[3 3 1]	[-5.27	-4.93	-6.53]	[60.99	0.94	1.73]	[-6.8	-0.5]	[2.48	3.16	1.36]
[8 7 6]	[-0.27	-0.93	-1.53]	Eigenvectors =			[-9.6	-1.4]	[7.31	7.22	6.48]
[7 6 9]	[-1.27	-1.93	1.47]	[[0.57	0.79	0.20]	[-1.6	-0.3]	[8.06	5.67	8.26]
[8 9 6]	[-0.27	1.07	-1.53]	[0.60	-0.25	-0.76]	[-1.1	2.1]	[7.7	9.09	6.21]
[9 8 8]	[0.73	0.07	0.47]	[0.55	-0.55	0.62]]	[-0.4	-1.8]	[8.76	8.08	8.17]
[10 8 10]	[1.73	0.07	2.47]	**Sorted Eigenvalues** =			[0.7	0.4]	[10.0	8.0	10.0]
[13 14 14]	[4.73	6.07	6.47]	[60.99 1.738 0.94]			[2.4	1.8]	[14.05	13.67	13.27]
[12 15 12]	[3.73	7.07	4.47]	**Sorted Eigenvectors** =			[10	0.3]	[13	14.69	11.3]
[13 11 11]	[4.73	3.07	3.47]	[[0.57	0.20	0.79]	[8.9	-1.9]	[12.14	11.27	11.6]
[14 15 13]	[5.73	7.07	5.47]	[0.60	-0.76	-0.25]	[6.5	0.8]	[14.18	14.94	2.88]
[15 13 12]]	[6.73	5.07	4.47]]	[0.55	0.62	-0.55]]	[10.6	-0.8]	[13.71	13.4	12.9]]
m=[8.2 7.9 7.5]							[9.4	0.3]]			

Figure 7. A worked example of PCA for three dimensions.

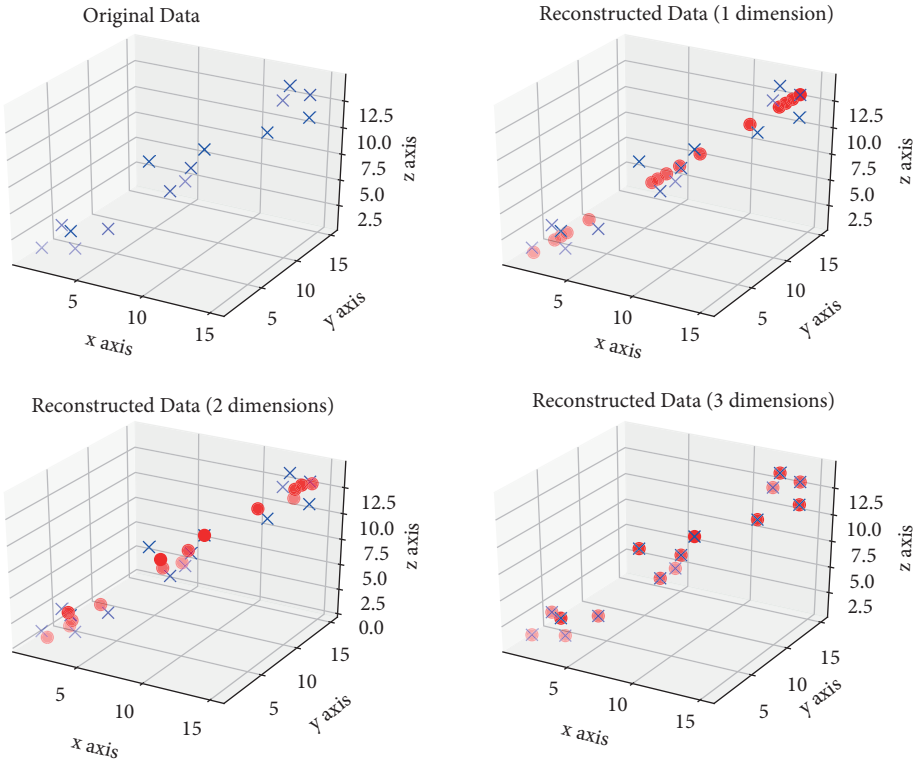

Figure 8. PCA dimensional reduction using three dimensions.

6. Literature review

The literature presents many applications that employ text clustering for different purposes. For instance, some of the text applications employed were Tweets clustering [3], identifying crime patterns on the news [4], document clustering for scientific texts using citation contexts [5], short text clustering using Wikipedia as an additional knowledge source [6], automatic topic clustering of the transcribed speech documents [7], Arabic web pages clustering and annotation using semantic class features [8], clustering for semantically related words [9],

multilingual corpora clustering [10], clustering of DNA texts [11]. Other applications include data compression, image processing, visualization, exploratory data analysis, pattern recognition, and time series prediction.

To highlight the contribution of this work, a comprehensive review is made to explore the reported textual features. The literature shows that the microword approach is unique as most of the previous studies employ roots, stems, complete words, n-grams, character trigrams, or topics. For example, [12] presents a study regarding the impact of stemming on Arabic text classification. Table 1 shows the most widely used textual features that are reported in the literature. It is worth to indicate that the information in Table 1 is not restricted to unsupervised learning as textual features can be used in both types of learning: supervised and unsupervised. In referencing [13], we concluded that they have revealed a thorough review of the data representation methods, especially for Arabic.

Regarding k-means clustering method that was used in this work, the literature shows many applications that employ this clustering technique. For instance, in [14], it was used for underground electrical profile clustering. In addition, k-means was used for object tracking in [15]. The authors is [16] presents a study regarding partitional clustering.

Table 1. A review of textual features.

References	Feature types
[11]	N-grams
[17]	Full-form words
[18]	Roots and Stems
[19]	Full-form words and character trigrams
[20]	Full-form words and character n-grams(3,4,5)
[21]	Microwords and full-form words
[22]	Character trigram
[23]	Topics
[24]	N-grams
[25]	Full-form words and stems
[26]	Full-form words
[27]	N-grams
[28]	Full-form words and N-grams
[29]	full-form words and Stems
[30]	full-form words
[31]	Full-form words and latent semantic indexing (LSI)

7. The proposed method

This study includes a comparison between two forms of the textual features. Hence, the proposed method considers two parts: the word-level form and the microword form. In text mining applications, preparing a data collection is the first step to conduct the experiments. Preparing a data collection mainly includes collecting, cleaning, and using the normalization processes. The procedure of this work includes a standard text clustering step except that it employs a performance evaluation of two different textual features: the complete word form and the microword form. For illustration, Figure 9 shows a part of the Arabic text that is translated using

the Google translator. The figure shows both the complete words and the microwords as the basic feature units. Arabic is read from right to left; hence, the first word in the Arabic text is divided into four parts when generating the microwords. In Figure 9, only the Arabic script is used for clustering, the English is just given as an explanation of the Arabic text.

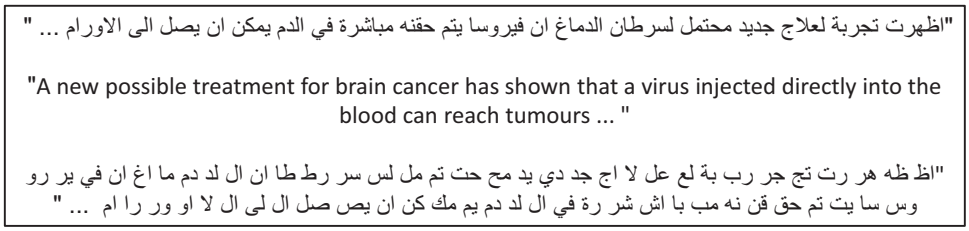

Figure 9. A part of text with the corresponding microwords.

To fulfill the work objective, the following procedure was implemented as follows:

– Prepare and preprocess the data collection: it included cleaning and normalizing the Arabic text such as removing the Hamza that comes over the letter Alif.

– Define the stop words list (to remove some common words such as prepositions).

– Define the ignore characters to remove some unnecessary characters such as numbers and other special characters (e.g., # @ $ + &).

– Define a threshold to discard the small words: Arabic, like English, has small words that can be ignored during classification.

– Define the DF threshold: It is a feature selection method to select the words that appear in more than a certain number of different documents.

– Create the vocabulary: This is a word list that contains the features that will contribute in the clustering process.

– Create the document-term matrix: it is a matrix that contains the data words or terms in the matrix rows and the documents in the matrix columns.

– Define the number of the reduced dimensions based on the chosen eigenvalues.

– Implement the PCA algorithm.

– Implement the k-means clustering algorithm.

– Find the clustering accuracy.

– Repeat the above steps for the microwords.

8. The experimental results

The experiments were performed using the Python programming language. It has many tools to facilitate the NLP processing and data mining applications. For instance, it has a class to create document-term matrices, PCA, k-means, and many others such as performance evaluation techniques. However, it has no class or function to produce microwords. Therefore, we explicitly wrote a program to decompose the words into microwords for the entire data collection. As previously indicated, the employed data collection contains 250 documents divided into five categories. The data was collected from Al-Jazeera TV channel website. It contains 54,472 words whereas the vocabulary contains 13,356 unique words. Table 2 shows the number of documents for each category.

For the complete word form experiments, we discarded the words that are less than four characters in

Table 2. The data collection.

No	Category in English	Number of documents
1	Health	50
2	Sports	50
3	Economy	50
4	Tourism	50
5	Technology	50
	Total	250

length. The smaller words are not important in the clustering problem. In fact, we experimentally verified that discarding small words of one, two, and three letters (i.e. "a", "an", or "the") would not affect the overall performance. In [30], he had a practical study of the effect of discarding the small words for the Arabic language. Table 3 shows the experimental results of the complete word case. The maximum accuracy is 97.2% and it is obtained using the 67 PCA components with the DF of more than 4 different documents. In addition, the table presents the vocabulary size, which is the number of words that were used to contribute to building the document-term matrix (i.e. the document feature vectors). In the table, the explained variances are the percentage of variances explained by the components. Finally, the table contains the accuracy, which is the performance of the clustering task when using the complete words as textual features. Accuracy is simply the ratio of correctly predicted observations to the number of actuals. The accuracies are sorted in descending order. In fact, we investigated all possible values of PCA and DF.

Table 3. The clustering performance using a complete word form.

No	Number of PCA components	DF more than	Vocabulary size	Explained variances	Accuracy (%)
1	67	4	1337	0.603495	97.2
2	113	6	827	0.809425	97.2
3	115	6	827	0.814853	97.2
4	118	6	827	0.823282	97.2
5	119	6	827	0.825613	97.2
6	120	6	827	0.828246	97.2
7	121	7	684	0.846726	97.2
8	129	6	827	0.850974	97.2
9	6	10	445	0.20136	96.8
10	12	1	4758	0.147338	96.8
11	12	12	340	0.328784	96.8
12	14	2	2726	0.19097	96.8
13	14	11	394	0.343187	96.8
14	15	10	445	0.342955	96.8
15	15	11	394	0.356475	96.8

For the microword approach, we measured the accuracy with the distinction to use the microwords instead of the complete words. Here, there was no need to discard small words as what we did in the complete words

were already words of two characters. The maximum accuracy is obtained when the number of components is 46 with DF more than 91. The vocabulary size was 301 words and the explained variance is 0.702172. It is worthy to observe that the DF range is larger in the microword approach. The reason is that when the DF increases the vocabulary words decrease, which leads at some point to a problem to create the textual features for some texts. In the complete words approach, when the DF is set to 35, the system fails as one or more of the Document-Term matrix rows will have zero values. Table 4 demonstrates the collected information regarding the microword approach.

Table 4. The clustering performance using microwords form.

No	Number of PCA components	DF more than	Vocabulary size	Explained variances	Accuracy (%)
1	46	91	301	0.702172	96.8
2	61	93	297	0.785125	96.8
3	36	88	310	0.621367	96.4
4	44	91	301	0.68902	96.4
5	47	89	307	0.702782	96.4
6	59	93	297	0.775311	96.4
7	63	92	298	0.793586	96.4
8	137	83	321	0.963452	96.4
9	12	84	319	0.341497	96
10	17	84	319	0.414478	96
11	19	89	307	0.452359	96
12	23	88	310	0.498572	96
13	27	94	291	0.553153	96
14	29	78	331	0.54015	96
15	32	85	317	0.579963	96

9. Discussion

After analyzing the results, there are three cases of observation. The first case is that both approaches (i.e. the full word form and the microword form) inaccurately cluster the text, the second case is that only the complete words form approach correctly clusters the text, and the third case is that the microword approach correctly clusters the text. Since this study investigates the microword approach, we checked the cases in which the microword approach failed to properly cluster the text. In our experiments, the microword approach had eight errors. Four of the errors also occurred in the complete word form. The reset of the errors that exclusively occurred in the microword approach were investigated to find the error sources. For instance, Figure 10 shows why the health category text was assigned to the Technology cluster. As shown, the text has two words stating "technology," which is the source of the misrecognition case. The microword approach recognized it as a Technology text even though it is a health text, which is a justifiable error and does not degrade the power of the microword approach. In this case, we highlighted the vocabulary which explains why there is an "accepted" reason the microword approach failed to cluster this text correctly.

Another error case is shown in Figure 11. This example demonstrates a part of an Economy category text that is inaccurately assigned to the tourism category. After investigating the text, we found that there

اكتشف فريق من الباحثين بالصدفة ان مستخلصا من بقايا ورق الشاي المهدر يمكن ان يعد فتحا في علاج سرطان الرئة وجاء هذا الكشف
...
لذلك نجمع اوراق الشاي المهدرة لهذه التكنولوجيا المفيدة ومع ذلك نبه الدكتور بيتتشيموتو الى ان هذا العلاج ليس وشيكا لانه يتعين على فريق البحث اولا فهم المزيد عن هذه العملية قبل امكانية بدء التجارب البشرية ويامل الباحثون امكانية نقل كل هذه التكنولوجيا الى التجارب السريرية في الهند بعد عامين وحتى لو نجحت التجارب فان العلاج المتاح على نطاق واسع يمكن ان يكون على الاقل بعد عشر سنوات

The translation using Google Translator

A team of researchers found by chance that extracts from wasted tea leaves could be considered a breakthrough in the treatment of lung cancer. The discovery came when researchers
…
because we get it from the wasted tea leaves, which make up one third of the tea paper crop that is not suitable for drinking and usually go to the landfill so we collect the wasted tea leaves for this useful technology. Dr. Pitchimoto cautioned that this treatment is not imminent because the research team should first understand more about this process before human trials can begin. Researchers hope that all of this technology can be transferred to clinical trials in India after two years. Li wide can be at least ten years later

Figure 10. An example of a health category text.

قالت الحكومة السويسرية الاربعاء انها ستمدد تجميد ارصدة يحتفظ بها رئيسا مصر وتونس السابقان لثلاث سنوات اخرى لمنح البلدين مهلة للتحقيق في مصدر هذه الاموال ويرمي التجميد لمنع اخفاء اية اموال يتم الحصول عليها بطريقة غير قانونية في اسواق مالية بعيدا عن رقابة الاجهزة القضائية وكانت السلطات السويسرية جمدت في نحو مليون فرنك سويسري مليون دولار للرئيس المصري المخلوع حسني مبارك وستين مليون فرنك مليون دولار للرئيس التونسي المخلوع زين العابدين بن علي منذ الاطاحة بهما في مطلع وقالت وزارة الخارجية السويسرية في بيان لها ان تحقيقات جنائية في مصدر هذه الارصدة لم تحقق تقدما كافيا يسمح لسويسرا برفع التجميد الذي يفرض لثلاث سنوات واضافت ان تمديد التجميد يستهدف منح التحقيقات في تونس ومصر مزيدا من الوقت مع الاخذ في الاعتبار التحول السياسي في البلدين النائب العام المصري طلب من سويسرا مد فترة تجميد الارصدة الخاصة بمبارك وعدد من رموز نظامه حتى انتهاء التحقيقات الجنائية بشان مصادر هذه الارصدة وكان النائب العام المصري المستشار هشام بركات قد طلب من المجلس الفدرالي السويسري مد فترة تجميد الارصدة الخاصة بمبارك وعدد من رموز نظامه لمدة ثلاث سنوات اخرى حتى انتهاء التحقيقات الجنائية في مصر بشان مصادر هذه الارصدة فريق خاص وسبق لعلي بن فطيس المري النائب العام في قطر محامي الامم المتحدة لاسترجاع الاموال المنهوبة ان قال للجزيرة في برنامج في العمق الشهر الماضي ان وزير العدل المصري اخبره بان بلاده بصدد تشكيل فريق خاص بملف الاموال المنهوبة وان وزير العدل هو المسؤول عن الملف واضاف المري ان الوضع السياسي في مصر اسهم في تأخير جهود استرجاع الاموال المنهوبة والتي قدرها البنك الدولي بنحو مليار دولار في مصر وبنحو مليار دولار في تونس واعتبر المسؤول الاممي ان تونس تعاملت بذكاء مع ملف استرجاع اموالها المنهوبة حيث طرقت باب الامم المتحدة مبكرا واستطاعت استعادة اموال بحوزة ليلى الطرابلسي في لبنان بقيمة مليون دولار دون ان تدفع دولارا واحدا وكانت السلطات السويسرية اقرت في اكتوبرتشرين الاول الماضي غرامات على ثلاثة بنوك بسبب فشلها في مراقبة اموال تعود لمقربين من بن علي ويتعلق الامر بفرع اتش اس بي سي لإدارة الثروات وبنك يو بي بي وبنك اي اف جي.

The translation using Google Translator

The Swiss government said on Wednesday it would extend a freeze on assets held by the two former presidents of Egypt and Tunisia for another three years to give the two countries time to investigate the source of the money and set the freeze to prevent the concealment of any money illegally obtained in financial markets away from the control of the judiciary. The Swiss Foreign Ministry said in a statement that a criminal investigation into the source of these lands had been carried out by Swiss President Zine El Abidine Ben Ali, Said that the extension of the freeze aims to give investigations in Tunisia and Egypt more time, taking into account the political transition in the two countries Egyptian Attorney General asked Switzerland to extend the freezing of assets of Mubarak and a number of symbols of his regime Egyptian Attorney General Hisham Barakat asked the Swiss Federal Council to extend the freezing of Mubarak's assets and a number of its symbols for another three years until the end of criminal investigations in Egypt on the sources of these funds. Qatar's Attorney-General Ali bin Fattis al-Marri has told the island in a program in depth last month that the Egyptian justice minister had told him that his country was in the process of forming a special team for the looted money file and that the justice minister was responsible for the file. The political situation in Egypt contributed to delaying efforts to recover the looted funds, which the World Bank estimated at about one billion dollars in Egypt and about one billion dollars in Tunisia. The UN official said that Tunisia dealt intelligently with the file of retrieving its looted funds as it knocked the door of the United Nations early and was able to recover money from The Swiss authorities have approved in October last year the fines of three banks for failing to monitor money belonging to Ben Ali's relatives. This concerns the HSBC Wealth Management Branch, UBB Bank, Bank I FJ.

Figure 11. Example of an economy-category text.

are many words that are related to the tourism category. To understand the source of the error, we tracked the decomposing process that generated the microwords of the given text assigning it to the tourism category. In fact, the investigated text has many tourism-related words, of course, based on the employed training data. Figure 11 highlights the tourism-related words such as “Egypt”, “Tunisia”, and “Dollars.” In the used data collection, the tourism category has 61 occurrences of the word “Egypt,” 27 occurrences of the word “Tunisia,” and 52 occurrences of the word “Dollars.” Moreover, some words were decomposed to generate the microwords

given in the tourism-related words such as the underlined word "political", which had one microword "S EE" which can also be related to the tourism microwords in the Arabic script. Again, this analysis supports the microword approach as it is fair to assign this text to the tourism category.

10. Conclusion

The output of this study recognizes the validation of a recently proposed textual feature (i.e. Bigram character features) into what we call the microwords. This method is characterized by having fixed vocabularies regardless of the corpus size, which is extremely important in text mining systems. Therefore, the microword method is considered as a data reduction method since it is able to compact the data while preserving the original characteristics. The intensive experimental evaluation shows that the microwords had scored almost the same accuracy as the complete word forms. In addition to handling the space-challenging problem, the proposed microword approach is beneficial to alleviate the problem of uncommon words as the corpus words are decomposed into microwords. For instance, the users of social networks such as 'Twitter' and 'Instagram' tend to use unusual vocabulary words (i.e. uncommon words), which raises the misrecognition rate. In general, reducing high-dimensional data is extremely important to minimize the execution time in data mining applications. In conclusion, this study reveals that the microword approach is favorable in terms of its effectiveness and efficiency when compared with the complete word form. In future works, the microwords method might be investigated in other text based applications such as text classifications and sentiment analysis.

Acknowledgment

The author would like to thank the Palestine Polytechnic University (PPU) for its support to conduct this research.

References

[1] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management 1988; 24 (5): 503-523.

[2] Al-Anzi FS, AbuZeina D. Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach. Information Processing & Management 2018; 54 (1): 105-115.

[3] Poomagal S, Visalakshi P, Hamsapriya T. A novel method for clustering tweets in twitter. International Journal of Web Based Communities. 2015; 11 (2): 170-87.

[4] Bsoul Q, Salim J, Zakaria LQ. An intelligent document clustering approach to detect crime patterns. Procedia Technology 2013; 11: 1181-1187.

[5] Aljaber B, Stokes N, Bailey J, Pei J. Document clustering of scientific texts using citation contexts. Information Retrieval 2010; 13 (2): 101-131.

[6] Banerjee S, Ramanathan K, Gupta A. Clustering short texts using wikipedia. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; the Netherlands; 2007. pp. 787-788.

[7] Maghawry AM, Omar Y, Badr A. Initial centroid selection optimization for K-Means with genetic algorithm to enhance clustering of transcribed Arabic Broadcast News Documents. In: Proceedings of the Computational Methods in Systems and Software; Cham; 2017. pp. 86-101.

[8] Alghamdi HM, Selamat A, Karim NS. Arabic web pages clustering and annotation using semantic class features. Journal of King Saud University-Computer and Information Sciences 2014; 26 (4): 388-397.

[9] Deepak P, Rao D, Khemani D. Building clusters of related words: an unsupervised approach. In Pacific Rim International Conference on Artificial Intelligence; Berlin, Heidelberg; 2006 pp. 474-483.

[10] Romeo S, Tagarelli A, Ienco D. Semantic-based multilingual document clustering via tensor modeling. In EMNLP, Conference on Empirical Methods in Natural Language Processing; Doha, Qatar; 2014. pp. 10-18.

[11] Volkovich Z, Kirzhner V, Bolshoy A, Nevo E, Korol A. The method of N-grams in large-scale clustering of DNA texts. Pattern Recognition 2005; 38 (11): 1902-1912.

[12] Al-Anzi FS, AbuZeina D. Stemming impact on Arabic text categorization performance: A survey. In: 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA); Marrakesh, Morocco; 2015. pp. 1-7.

[13] Al-Anzi FS, AbuZeina D. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University-Computer and Information Sciences. 2017; 29 (2): 189-195.

[14] Kutbay U, Ural AB, Hardalaç F. Underground electrical profile clustering using K-MEANS algorithm. In2015 23nd Signal Processing and Communications Applications Conference (SIU); Malatya, Turkey; 2015. pp. 561-564.

[15] Hardalaç F, Kutbay U, Şahin İ, Akyel A. A novel method for robust object tracking with K-means clustering using histogram back-projection technique. Multimedia Tools and Applications. 2018; 77 (18): 24059-24072.

[16] Kutbay U. Partitional clustering. In: Recent Applications in Data Clustering 2018 In techOpen.

[17] AbuZeina D, Al-Anzi FS. Employing fisher discriminant analysis for Arabic text classification. Computers & Electrical Engineering 2018; 66: 474-486.

[18] Harrag F, El-Qawasmah E, Al-Salman AM. Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm. In: 2010 First International Conference on Integrated Intelligent Computing; Bangalore, India; 2010. pp. 6-11.

[19] Sawaf H, Zaplo J, Ney H. Statistical classification methods for Arabic News articles. In Proceedings of the Arabic Natural Language Processing Workshop (ACL20001); Toulouse, France; 2001. pp.1-6.

[20] Sharef BT, Omar N, Sharef ZT. An automated arabic text categorization based on the frequency ratio accumulation. The International Arab Journal of Information Technology 2014; 11 (2): 213-221.

[21] Al-Anzi FS, AbuZeina D. A micro-word based approach for arabic sentiment analysis. In2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA. Hammamet, Tunisia; 2017. pp. 910-914).

[22] Khreisat L. A machine learning approach for Arabic text classification using N-gram frequency statistics. Journal of Informetrics 2009; 3 (1): 72-77.

[23] Zrigui M, Ayadi R, Mars M, Maraoui M. Arabic text classification framework based on latent dirichlet allocation. Journal of Computing and Information Technology 2012; 20 (2): 125-140.

[24] Güven A, Bozkurt ÖÖ, Kalıpsız O. Advanced information extraction with n-gram based LSI. In: Proceedings of World Academy of Science, Engineering and Technology 2006; 17: 13-18.

[25] Silva C, Ribeiro B. The importance of stop word removal on recall values in text categorization. In Proceedings of the International Joint Conference on Neural Networks; Portland,USA; 2003. pp. 1661-1666.

[26] Al-Anzi FS, AbuZeina D, Hasan S. Utilizing standard deviation in text classification weighting schemes. The International Journal of Innovative Computing, Information and Control 2017; 13: 4.

[27] Ghiassi M, Skinner J, Zimbra D. Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications 2013; 40 (16): 6266-6282.

[28] Al-Shalabi R, Obeidat R. Improving KNN Arabic text classification with n-grams based document indexing. In: Proceedings of the Sixth International Conference on Informatics and Systems; Cairo, Egypt; 2008. pp. 108-112.

[29] Song F, Liu S, Yang J. A comparative study on text representation schemes in text categorization. Pattern Analysis and Applications 2005; 8 (1-2): 199-209.

[30] Al-Anzi FS, AbuZeina D. A new enhanced variation of TF-IDF scheme for Arabic text classification. Health 2016; 400: 218-4.

[31] Al-Anzi FS, AbuZeina D. Enhanced Search for Arabic Language Using Latent Semantic Indexing (LSI). In: 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC); Bangkok, Thailand; 2018. pp. 1-4.