Clustering with Probabilistic Topic Models on Arabic Texts: A Comparative Study of LDA and K-Means

Abdessalem Kelaiaia[1] and Hayet Merouani[2]
[1]Department of Computer Sciences, University of May 08, Algeria
[2]Department of Computer Sciences, University of Badji Mokhtar, Algeria

Abstract: *Recently, probabilistic topic models such as Latent Dirichlet Allocation (LDA) have been widely used for applications in many text mining tasks such as retrieval, summarization and clustering on different languages. In this paper, we present a first comparative study between LDA and K-means, two well-known methods respectively in topics identification and clustering applied on Arabic texts. Our aim is to compare the influence of morpho-syntactic characteristics of Arabic language on performance of first method compared to the second one. In order to, study different aspects of those methods the study is conducted on four benchmark document collections in which the quality of clustering was measured by the use of four well-known evaluation measures, Rand index, Jaccard index, F-measure and Entropy. The results consistently show that LDA perform best results more than K-means in most cases.*

Keywords: *Clustering, topics identification, arabic text, LDA, k-means, preprocessing.*

Received December 7, 2012; accepted December 24, 2013; published online April 1, 2015

1. Introduction

Document clustering is a fundamental and enabling tool for efficient document organization. Recently, probabilistic topic models methods such as Latent Dirichlet Allocation (LDA) have been used in clustering (integrated in Mallet framework, Gensim framework and hierarchical clustering [18, 20, 21]) and take good results. Arabic language is greatly inflectional and derivational language which makes text difficult task. To our knowledge and until this writing, there is no study that highlights the influence of the morpho-syntactic characteristics of this language on performance of such methods in document clustering. For this reason, this paper will compare the influence of morpho-syntactic characteristics of Arabic language on performance of LDA and K-means which are well-known methods in topics identification and clustering.

The rest of paper is organized as follows: The next section document clustering, section 3 describes the clustering evaluation, section 4 presents Arabic language and related works, section 5 presents in details our experimentation procedure and evaluation, section 6 describes and discusses results, finally section 7 concludes.

2. Document Clustering

Clustering is a process of grouping objects represented in the same form in uniform groups (clusters). In document clustering objects become documents (texts). The need for such grouping is explained by the large number of texts that are often contained in a document collection. In the following two sections, we will describe the two methods used in this study.

2.3. K-Means Presentation

The K-means method belongs to the family of partitioning algorithms. This type of algorithm and its variants are best known in the community of data classification. In this type, each cluster is represented by an average (mean) or a weighted average called the "centroid" [25] which is the closest to all other elements in the cluster. This centroid is calculated using Equation 1.

$$C_j = \frac{1}{n_j} \Sigma X_i \in j\, X_i \tag{1}$$

Where C_j is the centroid of cluster j, X_i is an element of this cluster and n_j is the number of these elements. The K-means functioning is described below (here, we consider the basic K-means): Initially, the K random elements are selected arbitrary and assumed to be centroids. Centroids are points in the cluster which are the closest to all other elements in the cluster. All other elements are assigned to the nearest centroids and a new centroid is recomputed. The process is reiterated until no further elements move from one cluster to another.

2.3. LDA Presentation

Since, its first introduction by [5] the LDA continues to attract a considerable interest from the statistical machine learning and natural language processing communities. The idea behind LDA is that each document in the collection is modeled as a mixture over an underlying set of topics and each topic is modeled as a probability distribution over the terms in the vocabulary. According to this, the process of generating a collection is as follows (here, we describe the smoothed LDA with symmetric dirichlet priors [11]):

1. For each topic *z*, a multinomial distribution ϕ_z is sampled from a dirichlet distribution (β).
2. For each document *d*, a multinomial distribution θ_d over topics is sampled from a dirichlet distribution (α).
3. For each word *w* in the document *d*, a single topic is chosen according to distribution θ_d.
4. Each word is sampled from a multinomial distribution ϕ_z over words specific to the sampled topic *z*.

Thus, the likelihood of generating a whole collection is:

$$P(D_{1..N} \big| \alpha, \beta) = \iint \prod_{Z=1}^{K} P(\varphi_Z / \beta) \prod_{d=1}^{N} P(\theta_d / \alpha)$$
$$(\prod_{i=1}^{N_d} \sum_{Z_i=1}^{K} P(z_i / \theta) P(w_i / z, \varphi)) d\theta \, d\varphi \qquad (2)$$

Where *K* is the number of topics and *N* is the total number of documents in collection.

2.3. LDA and Clustering

According to [16] generally, there are two ways of using topic models for document clustering. The first approach uses a topic model to reduce the dimension of representation of documents (from word representation to topic representation) and then applies a standard clustering algorithm like K-means in the new representation whereas, the other approach uses topic models more directly. The idea is that each topic *z* becomes, after estimating the parameters ϕ and θ, a new cluster and the documents assigned to this cluster are the documents with the highest probability Equation 3 of assigning of the topic *z* to these documents.

$$arg\ max_{Z=1..K}\ \theta_d \qquad (3)$$

In the present study, we will use the second approach which allows us to measure the performance of LDA compared to traditional methods of clustering like K-means.

3. Clustering Evaluation

To evaluate the quality or goodness of produced clusters, two types of measures are usually used, internal and external measures [17, 25]. When we do not have an external knowledge about (predefined sets of classes) that allows us to compare different produced clusters, we use the first type and otherwise, we use the second one.

Many external measures are presented in the literature. To provide further evidence for the results in this study, we use four well-known evaluation measures, Rand index, Jaccard Index, F-measure and Entropy. The values of these measures are between 0 and 1 and higher is better, except entropy, for which lower is better.

3.1. Rand Index

The Rand index or Rand statistics [19] is a simple criterion used to compare a produced clustering structure with a predefined structure. The latter is computed by examining all pairs of documents in the collection after clustering. If two documents are in the same emplacement in both the predefined class and the clustering result, this counts as an agreement. If two documents are in different emplacement in both the predefined class and the clustering result, it is also an agreement. Otherwise, there is a disagreement. The Rand index is calculated according to Equation 4.

$$RI = \frac{(A+D)}{(A+B+C+D)} \qquad (4)$$

Where *A* is the number of pairs of documents that are in the same cluster and in the same class, *B* is the number of pairs of documents that are in the same cluster and in different classes, *C* is the number of pairs of documents that are in different clusters and in the same class and *D* is the number of pairs of documents that are in different clusters and in different classes.

3.2. Jaccard Index

Similar to Rand index the Jaccard index [12] is computed Equation 5 by examining all pairs of documents in the collection except it does not take into account the number of pairs of documents that are in different clusters and in different classes (*D*).

$$JI = \frac{A}{(A+B+C)} \qquad (5)$$

3.3. F-Measure

The F-measure [15] is a harmonic combination of two measures precision and recall [26] which has a long history in Information Retrieval (IR) domain for clustering evaluation. The cluster is viewed as the result of a query for a specific class and class is viewed as a relevant set of documents relevant for a query. Precision is the fraction of correctly retrieved documents (attribution of document from cluster to correct class) Equation 6 and recall is the fraction of correctly retrieved (classed) documents out of all matching documents Equation 7.

$$Precision(i, j) = \frac{n_{ij}}{n_j} \tag{6}$$

$$Recall(i, j) = \frac{n_{ij}}{n_i} \tag{7}$$

Where n_{ij} is the number of documents of class i in cluster j, n_j is the number of documents of cluster j and n_i is the number of documents of class i. The F-measure of cluster j and class i are then given by:

$$F\ (i, j) = \frac{(2 * Precision\ \ (i, j) * Recall\ \ (i, j))}{(Precision\ \ (i, j) + Recall\ \ (i, j))} \tag{8}$$

Note that, [16] present another definition of precision and recall calculated on entire collection:

$$Precision = \frac{A}{A+C} \tag{9}$$

$$Recall = \frac{A}{A+B} \tag{10}$$

3.4. Entropy

The entropy is a function of the distribution of classes in the resulting clusters [27] it indicates the quantity of disorder in these clusters [24]. The lower value of entropy indicates a better clustering (higher homogeneity). The greater entropy means that the clustering is not good.

Given a class i and a cluster j, the entropy of a cluster j is defined as:

$$E_j = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_{ij}}{n_j} \log \frac{n_{ij}}{n_j} \tag{11}$$

Where n_{ij} is the number of documents of the class i that were assigned to the cluster j, q is the total number of classes in the document collection and n_j is of documents in cluster j. Thus, the overall entropy is weight average of all cluster entropies:

$$Entropy = \sum_{j=1}^{k} \frac{n_j}{n} E_j \tag{12}$$

Where n_j is the size of cluster j, k is the number of clusters and n is the total number of documents in whole document collection.

4. Arabic Language

4.1. Particularities

The Arabic language has an alphabet containing 28 consonants that change their layout according to their position. Unlike English, Arabic is an agglutinative language; articles, prepositions and pronouns stick to adjectives, nouns, verbs and particles which they relate, which creates ambiguities during morphological analysis.

An Arabic word can represent a phrase in English; it may be composed of a stem (base), proclitics such as prepositions or conjunctions, prefixes and suffixes, which express grammatical features and indicate the functions of cases, verb mode and modalities (number, gender, etc.,) and enclitics, which are personal pronouns [9]. For example the word أتأكلونها, which mean: "do you eat it?" is decomposed as follows:

Table 1. Arabic word decomposition.

Enclitic	Suffix	Stem	prefix	Proclitic
ها	ون	أكل	ت	أ

The collage of flexional elements (proclitics, prefixes, suffixes, enclitics) creates patterns [8]. The flexion of a root may generate up to 150 different patterns حمل→ محمول، محمل، حامل، .... This property makes the application of preprocessing techniques such as stemming very useful especially in IR systems.

4.2. Arabic Language and Clustering

The nature of the Arabic language, the writing system, writing orientation, omission of vowels and morphological structure has slowed research into this language, especially in automatic classification (categorization or clustering). In the literature, most of the research is focused on the morphological aspect of this language [1, 14] via developing preprocessing tools such as stemming and their influence on IR or on supervised classification (categorization), but only a small number of research projects focus on document clustering, we identified two major works, on a morphological analysis based on the language and using the n-gram. Sawaf *et al.* [23] used a statistical approach (based on the technique of entropy maximization) for the clustering of an Arab-based articles covering several areas such as politics, economics, etc., [13] developed an algorithm (integrated into the standard software clusters TEMIS insight discoverer) that, from descriptors in Arabic, contains similar documents in classes according to their semantic similarity and proximity topic.

4.3. Arabic Language and Probabilistic Topic Models

Regarding the use of probabilistic topics models in Arabic language, we mentioned a single major work [6] which investigates the influence of root-based stemming approach on LDA supervised classification.

5. Experimentation Procedure

5.1. Document Collections

To evaluate the performance of the two methods, LDA and K-means, we used four document collections shown in Table 2. The details of each document collections are described as follows:

- Corpus of Contemporary Arabic (CCA): Compiled by El Sulaiti [10] from radio Qatar and includes 432 text documents.
- Alwatan on-line Newspaper: Colleted from Alwatan on-line newspaper during 2004 by mourad [2] includes 20291 text documents.
- BBC Arabic Corpus: Member of Open Source Arabic Corpora (OSAC), collected from bbcarabic.com by Saad [22] includes 4763 text documents.
- OSAc: Member of OSAC, collected from multiple sites by Saad [22] includes 22429 text documents.

Table 2. Document collections.

Documents Collections	Number of Classes	Number of Documents	Number of Words (Million)	Size (Mbytes)
CCA	15	432	0,82	10
Al Watan Online Newspaper	6	20291	10	118
BBC Arabic Corpus	7	4763	1,86	29
Osac	11	22429	18,18	182

5.2. Text Preprocessing

Preprocessing aims to standardize the representation of texts to be classified. There are commonly four steps:

1. First, we convert text files to UTF-8 encoding and remove non letters, punctuation marks and diacritics.
2. Second, we need to give a transliterated form to each word in each document in document collection to be useful for clustering process.
3. Third, we need, after tokenization, remove stop words such as 'أين', 'كان', 'حيث', etc., since they are frequent and carry no information. During our study we collected about 875 stop words.
4. Fourth, we need to stem the word to its origin, which means we only consider the root form of words. Stemming aims to obtain the lexical root or stem for words in natural language, by removing affixes attached to them, i.e., it's regrouping under a single identification words whose root is common. For example, the words حملة ،محمل ،يحمل are flexions of stem حمل. For this, stemmers are developed; they are generally designed for a specific language on which a certain expertise should be developed. Reference [14] considers that the use of a dictionary for stems and morphological analysis are other forms of stemming. Several stemming algorithms have been studied for different languages. For Arabic, there are several stemmers, the most famous are Al-Stem and StemmerLight10 [7, 14]. In our case we choice employ Al-Stem due to its performance [8] and we re-implemented it to work with entire text.

Note that, before steps 3 and 4 we save each text document in cleaned form and stemmed form in the goal to test the influence of removing of the stop words and stemming on clustering process. According to this, we will have three forms for each document collection, raw form, cleaned form and stemmed from.

5.3. Clustering Process

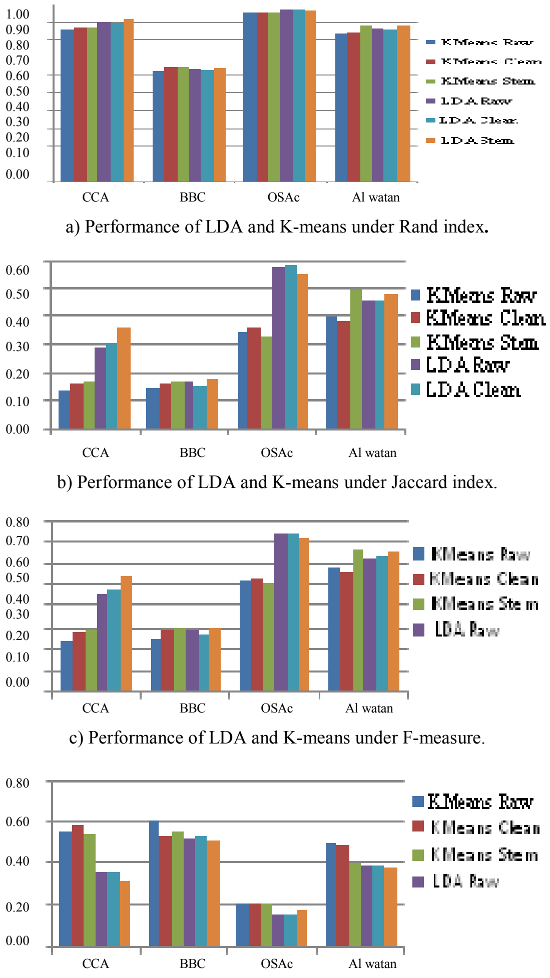
In this phase we submitted the preprocessing results of the four document collections (raw form, cleaned form and stemmed form) on the clustering process with LDA and K-means methods.

5.4. Evaluation

After clustering process we computed the evaluation measures, Rand index, Jaccard index, F-measure and Entropy for all clustering results compared to predefined structure showing in Table 2.

6. Results and Discussion

First, we present the results of conducted experiments in Figure 1 summarized in Table 3 which clearly demonstrate at first sight that for all document collections, LDA perform best results more than K-means. As mentioned above the results showing in Figure 1 are computed according to the predefined structure Table 2.



a) Performance of LDA and K-means under Rand index.

b) Performance of LDA and K-means under Jaccard index.

c) Performance of LDA and K-means under F-measure.

d) Performance of LDA and K-means under Entropy.

Figure 1. Performance of LDA and K-means.

6.1. Performance of LDA over K-Means

From the Table 3 we observe that LDA provides a substantial performance improvement over K-means with four metrics on the three collection forms. This performance is in contrast to research of [16] which states that the use of probabilistic topic models in clustering is not as accurate as traditional clustering methods such as K-means in respect to the functioning of this topic models. [16] Research has been conducted on Reuters 21578 and TDT2 document collections. Both collections are in English, this may be the cause of decreasing of performance of probabilistic topic models with respect to K-means. This leads us to say that the morpho-syntactic characteristics of language (inflectional and derivational characteristics in our case) have a great influence on the performance of probabilistic topic models in clustering and not only their functioning.

Table 3. Average performances of LDA over K-means on the four document collections in raw, cleaned and stemmed forms.

	Rand	Jaccard	F-mesure	Entropy
Raw Form	2,35%	11,43%	12,77%	10,74%
Cleaned Form	1,33%	10,76%	11,34%	9,12%
Stemmed form	1,49%	10,29%	11,39%	7,97%

Note that, all metrics indicate that the results performed with LDA over K-means on raw form are more than those obtained on the two other forms. We will go back on in the following section.

6.2. Influence of Elimination of the Stop Words on Clustering Quality

Table 4 shows the comparison of the results of clustering performed on raw and cleaned forms with both methods where K-means appears to perform better results than those obtained with LDA (decreasing in Rand and entropy metrics). This leads us to say that with LDA, removing stop words in Arabic text decrease the performance of the quality of obtained clusters. These results are in line with those obtained in [16] and confirm their belies to the usually unstated, but widespread, assumption in papers [5, 11] on LDA that the removal of stop words is a necessary preprocessing step.

Table 4. Average performances of LDA and K-means with removing stop words over the four document collections (comparison between raw and cleaned forms).

Methods	Rand	Jaccard	F-Measure	Entropy
K-Means	0,80%	1,19%	1,84%	1,42%
LDA	-0,22%	0,53%	0,41%	-0,19%

6.3. Influence of Stemming on Clustering Quality

Table 5 shows the comparison of the results of clustering performed on raw and stemmed forms with both methods. In this table, we notice that the stemming has improved quality of the obtained clusters. This is perhaps mostly due to the effect of stemming, which helped remove the flexions of words that have the same root, so the documents relating to the same topic will have a greater chance of being in the same cluster. Also we noticed that, similar to elimination of stop words, K-means appears to perform better results than those obtained with LDA.

Table 5. Average performances of LDA and K-means with stemming over the four document collections (comparison between raw and stemmed forms).

Methods	Rand	Jaccard	F-Measure	Entropy
K-Means	1,66%	3,49%	4,12%	4,33%
LDA	0,80%	2,35%	2,74%	1,56%

7. Conclusions and Future Plan

The present work compared between LDA and K-means in order to examine the reaction of LDA in clustering of Arabic texts which is a very flexional language. The experiment was conducted on four benchmark Arabic document collections, CCA, Alwatan on-line newspaper, BBC Arabic corpus and OSAc. Four metrics were used, rand index, Jaccard index, F-measure and Entropy. We started by doing the comparison between the results obtained by the two methods on the four document collections and the predefined structure. The results consistently show a clear improvement of LDA over K-means on raw, cleaned, and stemmed forms of document collections. Afterwards, we investigated the influence of the preprocessing tasks, elimination of stop words and stemming on performance of the studied methods. We found that LDA reacts less than K-means in both cases, especially in first one.

In the second case, both methods show a good improvement. Such improvement is due to the fact that stemming attenuates the flexional characteristics of the Arabic language despite the ambiguities that may result in some cases. These cases are, in our opinion, more than compensated for by the high rate of correct stems extracted.

Bearing in mind the obtained results, our future work will extend the present study to other important parameters such as lemmatization, a very important preprocessing operation in Arabic language, and investigate other variant of LDA such as Dynamic Topic Model (DTM) [4] and Correlated Topic Model (CTM) [3].

References

[1] Ababneh M., Al-Shalabi R., Kanaan G., and Al-Nobani A., “Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness,” *the International Arab Journal of Information Technology*, vol. 9, no. 4, pp. 368-372, 2012.

[2] Abbas M., Smaili K., and Berkani D., “Multi-Category Support Vector Machines for Identifying Arabic Topics,” *Advances in Computational Linguistics*, *Special Issue of Journal of Research in Computing Science*, vol. 41, pp. 217-226, 2009.

[3] Blei D. and Lafferty J., "A Correlated Topic Model of Science," *the Annals of Applied Statistics*, vol. 1, no. 1, pp. 17-35, 2007.

[4] Blei D. and Lafferty J., "Dynamic Topic Models," *in Proceedings of the 23rd International Conference on Machine Learning*, New York, USA, pp. 113-120, 2006.

[5] Blei D., NG Y., and Jordan I., "Latent Dirichlet Allocation," *the Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[6] Brahmi A., Ech-Cherif A., and Benyettou A., "Arabic Texts Analysis for Topic Modeling Evaluation," *Information Retrieval*, vol. 15, no. 1, pp. 33-53, 2012.

[7] Darwish K. and Oard W., "Evidence Combination for Arabic-English Retrieval," available at: https://terpconnect.umd.edu/~oard/pdf/trec02.pdf, last visited 2002.

[8] Darwish K., Hassan H., and Emam O., "Examining the Effect of Improved Context Sensitive Morphology on Arabic Information Retrieval," *in Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, USA, pp. 25-30, 2005.

[9] Diab M., Hacioglu K., and Jurafsky D., "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," *in Proceedings of the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/ Human Language Technologies Conference*, USA, pp. 149-152, 2004.

[10] El Sulaiti L., "L'arabe Contemporain," Radio Qatar, Qatar, 2003.

[11] Griffiths L. and Steyvers M., "Finding Scientific Topics," *in Proceedings of the National Academy of Science*, USA, pp. 5228-5235, 2004.

[12] Hamers L., Hemeryck Y., Herweyers G., Janssen M., Keters H., Rousseau R., and Vanhoutte A., "Similarity Measures in Scientometric Research: The Jaccard Index versus Salton's Cosine Formula," *Information Processing and Management*, vol. 25, no. 3, pp. 315-318, 1989.

[13] Huot H. and Coupet P., "Le Text Mining sur la langue Arabe : Application au Traitement des Sources Ouvertes," TEMIS SA, Paris, France, 2005.

[14] Larkey S., Ballesteros L., and Connell E., *Light Stemming for Arabic Information Retrieval, Arabic Computational Morphology*, Springer, 2007.

[15] Larsen B. and Aone C., "Fast and Effective Text Mining using Linear-Time Document Clustering," *in Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, CA, USA, pp. 16-22, 1999.

[16] Lu Y., Mei Q., and Zhai C., "Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA," *Information Retrieval*, vol. 14, no. 2, pp. 178-203, 2011.

[17] Manning D., Raghavan P., and Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, UK, 2008.

[18] Mccallum K., "MALLET: A Machine Learning for Language Toolkit," available at: http://mallet.cs.umass.edu, last visited 2002.

[19] Rand M., "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846-850, 1971.

[20] Řehůřek R. and Sojka P., "Gensim-Python Framework for Vector Space Modelling," Masaryk University, Brno, Czech Republic, 2011.

[21] Rosen-zvi M., Griffiths T., Steyvers M., and Smyth P., "The Author-topic Model for Authors and Documents," *in Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Alberta, Canada, pp. 487-494, 2004.

[22] Saad K. and Achour W., "OSAC: Open Source Arabic Corpora," *in Proceedings of the 6th International Symposium on Electrical and Electronics Engineering and Computer Science*, European University of Lefke, pp. 118-123, 2010.

[23] Sawaf H., Zaplo J., and Ney H., "Statistical Classification Methods for Arabic News Articles," available at: http://www.abdelali.net/ref/Sawaf_ArabicClassification.pdf, last visited 2001.

[24] Shannon E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.

[25] Steinbach M., Karypis G., and Kumar V., "A Comparison of Document Clustering Techniques," available at: http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf, last visited 2000.

[26] Van Rijsbergen J., *Information Retrieval*, London, UK, 1979.

[27] Zhao Y. and Karypis G., "Criterion Functions for Document Clustering: Experiments and Analysis," available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.4633&rep=rep1&type=pdf, last visited 2001.

Abdessalem Kelaiaia received his Engineer degree from Annaba University, Algeria in 1996, and his MS degree in Computer Science from the Guelma University, Algeria in 2008. Currently, he is working as an Assistant Professor at the University of May 08, Algeria and he is preparing the PhD degree at Annaba University. His current research field is text mining.

Hayet Merouani received her Engineer degree from Annaba University, Algeria in 1984, PhD degree from Robert Gordon University, UK. Actually, she is full Associate Professor at Badji Mokhtar University, Annaba. She also, leads Research group of Pattern recognition as a national program research of breast cancer. Her current works focus on the computer vision, medical imaging and Biometry.