ASTES

*Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 900-905 (2017)*

www.astesj.com

*Special Issue on Recent Advances in Engineering Systems*

**ASTESJ**

**ISSN: 2415-6698**

# Features based approach for indexation and representation of unstructured Arabic documents

Mohamed Salim El Bazzi[*1], Driss Mammass[1], Abdelatif Ennaji[2], Taher Zaki[1]

[1]*IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco.*

[2]*LITIS Laboratory, University of Rouen, France.*

ARTICLE INFO

*Article history:*
*Received: 05 April, 2017*
*Accepted: 03 June, 2017*
*Online: 28 June, 2017*

*Keywords:*
*Unstructured documents*
*Arabic text mining*
*Keyphrases*
*Indexation*
*Classification*

ABSTRACT

*The increase of textual information published in Arabic language on the internet, public libraries and administrations requires implementing effective techniques for the extraction of relevant information contained in large corpus of texts. The purpose of indexing is to create a document representation that easily find and identify the relevant information in a set of documents. However, mining textual data is becoming a complicated task, especially when taking semantic into consideration. In this paper, we will present an indexation system based on contextual representation that will take the advantage of semantic links given in a document. Our approach is based on the extraction of keyphrases. Then, each document is represented by its relevant keyphrases instead of its simple keywords. The experimental results confirms the effectiveness of our approach.*

## 1. Introduction

Document indexing is an essential step in the text mining process because it determines how the knowledge contained in the documents is represented. It takes place each time a document is added to the corpus. Indexation must then deal with two main problems, the choice of the most representative terms of each document and the evaluation of their power of representation.

Several approaches are proposed in the literature, particularly in English, but they are unusable for documents in Arabic, given its specific characteristics and the complexity of its morphology, grammar and vocabulary. For instance, association rules and semantic approaches based on external references are more complicated to use on Arabic documents.

The indexation, under all its forms, aims to extract the most relevant descriptors from a given document. The more sophisticated the selection is, the more accurate are the subsequent tasks of text mining using indexation systems, like classification and information retrieval.

Document indexing involves extracting keywords that best represent a document. In spite of the essential role of this phase in the next step of the natural language processing process, few are the works identified at this level [1][2][3].

While preparing this paper, we stopped particularly on descriptor's type extracted from a document in a text mining process. We noticed that stemming is the most used method to represent a document.

This paper is an extension of work originally presented in the 11th International Conference on Intelligent Systems: Theories and Applications (SITA 2016) [4]. It presents a comparison of the two text representation methods: stem-based representation and keyphrase-based representation, as well as their applications and compatibility with the Arabic language. However, our proposed approach can be applied to other languages.

The rest of this paper is organized as following. The second part introduces the related works. The third part is dedicated to the presentation of a conventional text mining process. The fourth part describes the proposed indexing systems and the used methods. We will present and discuss the experimental results in

* Mohamed Salim El Bazzi, IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco, Email: elbazzi@yahoo.fr

the fifth section and we conclude by highlighting our contribution and any possible improvements.

## 2. Related Works

Before starting the keywords extraction, a very important task must be accomplished, which is cleaning and normalizing the text. Tokenization, stop words removal and normalization are frequently used as a preprocessing step. While stemming is the most complicated and discussed issue.

The stemming method consists of extracting the root of a word. In other words, associating the words linked morphologically to the same root. The number of terms is then reduced, which makes the system more effective. This technique reveals a major drawback which is ambiguity.

A root can be defined as a word that cannot be created from other word. For example, the root of the Arabic word ("التصفيات", *the qualifications*) is ("صف", *row*). Although stemming methods cause a big loss of meaning from the original words, these techniques are competitive and provide good classification results. Hence, several works on Arabic text indexing opt for roots as document's descriptors.

Mansour et al. in their work [5], they perform a morphological analysis of the words in a document to extract indexes. First, the authors propose a process of extracting stems. Second, they set up a recognition system of names and verbs, based on rhymes (grammatical patterns) and grammatical rules. A weight is then assigned to each stem taking into account its occurrence and introducing a function indicating how the word is spread in the document

El-Khoribi [6] applied the stemming as a representation of features. Features are represented as vectors that consist of a number of elements equal to the number of classes which the probability of belonging to a class where stems occur. Then a stems lookup table is built from roots and labels of the classes to which they belong. Then, a Hidden Markov Model is used to evaluate the membership of a new document to a class.

Al Molijy et al. [7] used a linguistic method for the extraction of indexes from Arabic documents. Their method is based on the parsing of words document. The proposed algorithm consists of breaking words into N-grams (where N equals 3, 4 or 5), calculate their frequencies, and then retain the first 100 most frequent words, constituting thereby an N-gram document profile. The advantage of this method is the reduction of the number of words representing a document.

Bawaneh et al. [8] compared between the two classifiers K-Nearest Neighbor (KNN) and Naïve Bayesian (NB). The light stemmer was used as feature and TFIDF (Term Frequency – Inverse Document Frequency) measurement as a method of weighting features. The K-NN classifier was considered more efficient.

Gharib et al. [2] applied four classifiers on Arabic corpus which are: SVM (Support Vector Machine), NB, K-NN and Rocchio method, using stemming as features representation technique and TFIDF as weighting method. The Rocchio classifier works better when the feature space is small but the SVM performs better when space becomes larger.

Raheel et al. [3] have shown in a comparative study the influence of the choice of entities representing a document, on manipulating the performance of classifiers. They selected as descriptors, words in their original form, lemmas, roots, and the n-grams. Two classifiers were used, the SVM and Naive Bayesian Networks. SVM based on the 3-grams gave better classification results.

Harrag et al. [9] conducted a comparative study of three pretreatment techniques: light stemming, root-based stemming and dictionary lookup stemming in order to reduce the feature space. Then two classifiers were tested, Artificial Neural Networks and SVM.

Since the stemming is the method of representation which is the most common in text mining preprocessing, several studies have contributed to its improvement [10][11][12][13][14] [15]. However, the extraction of keyphrases has not experienced the same progress [16].

## 3. Arabic Text Mining Process

المغرب يتجاوز تداعيات أزمة المال هذه السنة
أبدى بنك المغرب ارتياحاً لأداء الاقتصاد في المغرب، خلال 2009، وتوقع أن يتراوح معدل النمو بين 5 إلى 6 في المائة للعام الثاني على التوالي، رغم تداعيات أزمة الاقتصاد العالمية وتأثيرها على إيرادات المغرب من العملات الصعبة والاستثمارات الخارجية.
وتوقع المركزي أن يحقق النمو الاقتصادي 3 إلى 4 في المائة في 2010، ارتباطاً بالإنتاج الزراعي وتحسن أسعار القطاعات الصناعية.

Figure 1: Original text from the used corpus.

### 3.1. Preprocessing

Before starting the document indexing phase, a very important task is to clean up and standardize the text. The following are the most commonly used steps:

- *Tokenization*

Tokenization is the production of a sequence of segments separated by spaces or punctuation marks. The output is a list of words without neither punctuation, nor special characters.

المغرب يتجاوز تداعيات أزمة المال هذه السنة أبدى بنك المغرب ارتياحاً لأداء الاقتصاد في المغرب خلال 2009 وتوقع أن يتراوح معدل النمو بين 5 إلى 6 في المائة للعام الثاني على التوالي رغم تداعيات أزمة الاقتصاد العالمية وتأثيرها على إيرادات المغرب من العملات الصعبة والاستثمارات الخارجية وتوقع المركزي أن يحقق النمو الاقتصادي 3 إلى 4 في المائة في 2010 ارتباطاً بالإنتاج الزراعي وتحسن أسعار القطاعات الصناعية

Figure 2: text after tokenization.

- *Stop words removal*

Stop words are non-significant terms in a text. It is a about eliminating words whose occurrence is very frequent and do not bring any added value to the process of indexation. These words are usually pronouns, articles and conjunctions.

المغرب يتجاوز تداعيات أزمة المال السنة أبدى بنك المغرب ارتياحاً لأداء الاقتصاد المغرب خلال توقع يتراوح معدل النمو المائة للعام الثاني التوالي تداعيات أزمة الاقتصاد العالمية تأثيرها إيرادات المغرب العملات الصعبة الاستثمارات الخارجية توقع المركزي يحقق النمو الاقتصادي المائة ارتباطاً بالإنتاج الزراعي تحسن أسعار القطاعات الصناعية

Figure 3: text after stop words removal.

- *Stemming*

Stemming can be defined as the process of removing prefixes, infixes and suffixes from words to reduce these words to their stems (roots). To remedy the problem of big loss of meaningful information, the notion of light stemming has been evoked in several works and consists in eliminating just the prefixes and suffixes of a given word without having to go back to its root.

- *Conversions*

The first conversion that can be applied to a document is the elimination of diacritical marks. Diacritical marks are signs added to the top or bottom of the Arabic letters to specify the pronunciation of the word. This phonological role also influences the meaning of this word. Indeed, two words can be written in the same way but differentiated by the addition of different diacritical signs. For example, if the word "عالم" is pronounced (عالم, âalim) it means "scientist", and if pronounced (عالم, âalam) it means "world". This procedure aims to standardize the documents because it is rare to find a fully accented corpus.

مغرب جوز دعا ازم مول سنة ابدى بني مغرب ارتياحا لاداء قصد مغرب خلل وقع روح عدل نمي عوم ثنى ولى رغم دعا ازم قصد علم اثر ردا مغرب عمل صعب ثمر خرج وقع مركز حقق نمي قصد ارتباطا توج زرع حسن سعر قطع صنع

Figure 4: text after stemming and conversions.

The second conversion is that of characters which has for purpose to normalize the letters which can be written in several forms. Thus the characters "إ" ,"أ" and "آ" are replaced by "ا", similarly "ة" is converted into "ه" and "ئ" ,"ي" into "ى". In our system, we did not consider this conversion.

### *3.2. Term weighting*

The weighting of an indexing term is the association of numerical values called weight to that term to represent its power of discrimination for each document in the collection. This characterization is linked to the informativeness of the term for the given document. This notion refers to the amount of meaning that a word carries.

There are several ways to determine the weight of a term. In our case, we have used the TF-IDF (Term-Frequency Inverse Document Frequency) method.

$$Weight_i(j) = TF_{ij} \times IDF_j \quad (1)$$

Where $TF_{ij}$ is the occurrence frequency of term the **j** in a document **i** and $IDF_j$ is the inverse absolute frequency of term **j** in the collection. Thus the weight of a term increases if it is frequent in the document and decreases if it is frequent in the collection.

### *3.3. Document representation*

Document representation is one of the techniques used to reduce the complexity of documents and makes them easier to manipulate, and the document is then transformed from its textual version to a matrix [Document × Term] (Figure 5). The representation of the most used document is the Vector Space Model. The documents are represented by vectors of words. This representation has its own drawbaks as the large dimension of representation and loss of correlation between adjacent words, the thing that leads to the loss of the semantic relationship that exists between the terms in the original document. To overcome these problems, weighting methods are used to assign appropriate weights to the term as shown in Figure 5:

| | $T_1$ | $T_2$ | … | $T_m$ | |
|---|---|---|---|---|---|
| $D_1$ | $w_{11}$ | $w_{12}$ | … | $w_{1m}$ | $C_a$ |
| $D_2$ | $w_{21}$ | $w_{22}$ | … | $w_{2m}$ | $C_b$ |
| … | … | … | … | … | … |
| $D_n$ | $w_{n1}$ | $w_{n2}$ | … | $w_{nm}$ | $C_k$ |

Figure 5 : Document × Terme Matrix.

Each input represents a vector of terms, where $w_{nm}$ is the weight of the term $T_m$ in the document $D_n$, and $C_i$ is the class assigned to the document $D_i$.

### *3.4. Classification*

The text classification is an important part of the text mining. It consists in providing a set of learning data (labeled documents) to the classification system. The task is then to determine a classification model that is capable of affecting the correct class for a new document.

Lately, the task of automatic text classification has been widely studied and the progress seems to be efficient in this field [2]. Several classification methods have been compared and proven to be effective, including the Bayesian classifier, decision trees, K-nearest neighbor, Support Vector Machines, neural networks.

The classification of textual documents presents many challenges and difficulties. First, it is difficult to grasp high level semantics and abstract concepts of natural language only a few keywords. This confirms that the efficiency of the indexing step is paramount and entirely decisive.

### *3.5. Evaluation*

Experimental evaluation of classifiers is the last step in the indexing process, and generally attempts to evaluate the effectiveness of a classifier, namely, its ability to make categorization decisions. To this end, there are many measures, each emphasizing one or other property of the system.

We used the most commonly used measures: recall, precision and the f-measure which synthesizes the first two formulas. Let's consider the following nominations:

TP (True positive), FN (False Negative), FP (False positive), TN (True Negative).

$$\textbf{Recall} = \frac{TP}{TP + FN} \quad (2) \qquad \textbf{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\textbf{F-measure} = \frac{2 \times recall \times Precision}{(recall + Precision)} \quad (4)$$

## 4. Proposed Approach: Keyphrases based indexing

### *4.1. Text features*

Descriptors are the units of text that represent its content. There are several categories of descriptors, which attempt, on the one hand, to reduce the size of the document, and on the other hand to preserve its semantic aspect. We quote:

*Word*: the text is simply segmented into single words after removing stop words.

*Lemma*: the words of the text are reduced to their canonical forms. The process of lemmatization consists on using grammatical rules to replace verbs by their infinitives and names by their singular forms.

*Root (or stem):* the stemming process extracts the roots of each word from the text. The words are then associated with their common root.

*Concepts*: they take the form of a word or group of words, coming from a list of controlled lexicon, generally dictionaries or ontologies.

*Keyphrase*: it is a sequence of words that describes more perfectly a document than a single word. They maintain better context and semantics. For example, the term "middle east" is more descriptive than "middle" and "east" taken distinctly.

### *4.2. Keyphrases Extraction Methotds*

In this section, we try to give an overview of different used methods for unsupervised keyphrases extraction (table 1). Consequently, our proposed system is completely automatic. The unsupervised methods have the particularity to abstract the specific nature of the processed data.

Table 1: Overview of unsupervised methods for keyphrases extraction

| Reference | Approach |
|---|---|
| Nabil et al. [26] | Combining linguistic methods with statistical methods |
| Ali et al. [27] | enrich statistical and linguistic information |
| Mihalcea et al. [28] | A Graph based ranking algorithm |
| El-Beltagy et al. [29] | Consider the position of the first occurrence of any given phrase |
| Elshishtawy et al. [30] | Combining the linguistic knowledge and the machine learning techniques |
| Najadat et al. [31] | Phrase frequency (PF), summation of phrase terms frequencies (Tf), PFIDF (Phrase Frequency - Inverse Document Frequecy), Phrase Position, Title Threshold and phrase distribution |
| Liu et al. [32] | The best keyphrase must be: understandable, semantically relevant with the document and have high coverage of the whole document |
| Sarkar et al. [33] | A candidate keyphrase is considered as a sequence of words that does not contain neither punctuations nor stop words, and then this sequence is broken into smaller phrases |

### *4.3. The Used Keyphrases Extraction Method*

Our implemented method for keyphrases extraction is inspired from Najadat's [31] approach which uses Phrases frequency. Hence, to extract keyphrases we follow those steps:

1. Segmentation of the text into simple words;
2. Calculating the number of co-occurrences of two (or more) consecutive words. To get a candidate keyphrase, the consecutive words should co-occur at least two times in the given document.
3. Calculate the TFIDF of obtained keyphrases to get the final list of indexes.

## 5. Results and Discussion

In this work, we have conducted a study about the influence of the choice of features' type on the efficiency of documents indexing. Consequently, experimental results should clarify our hypothesis which considers that keyphrases describe better the content of a document than simple keyword.

For validation, we tested our system on a corpus of 1000 documents of electronic press, extracted from Aljazeera [1] and Alarabiya [2]. This corpus is classified into 3 categories. These categories are: Economics, Politics and Sport.

In this section, a series of experiments was conducted. For each experiment, we used TFIDF as weighting method and the optimal and same Thresholds' values of K-Nearest Neighbor and features selection. Tables 2 and 3 show different classification results obtained for each feature type. These results are expressed through recall, precision and F-measure criteria.

Table 2: classification results using stems

| | Recall | Precision | F-measure |
|---|---|---|---|
| Economics | 0.164 | 0.392 | 0.231 |
| Politics | 0.477 | 0.326 | 0.387 |
| Sport | 0.393 | 0.351 | 0.371 |
| **Average** | **0.345** | **0.356** | **0.350** |

Table 3: classification results using keyphrases.

| | Recall | Precision | F-measure |
|---|---|---|---|
| Economics | **0.447** | **0.461** | **0.454** |
| Politics | **0.537** | **0.327** | **0.406** |
| Sport | 0.136 | **0.36** | 0.197 |
| **Average** | **0.373** | **0.382** | **0.378** |

[1]http://www.aljazeera.net

[2]http://www.alarabiya.net

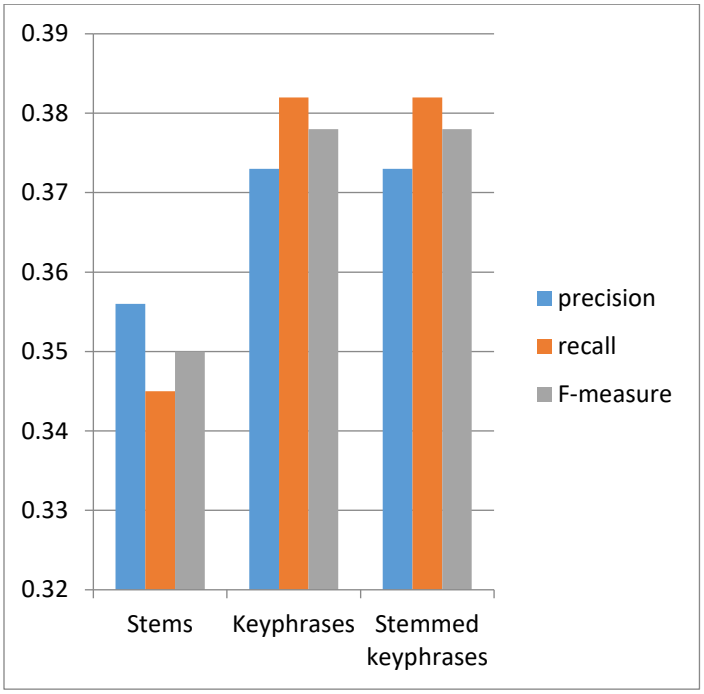
We conducted another test by performing the stemming firstly and then proceeding to extract keyphrases. After this combination, we obtained a list of stemmed-keyphrases. Classification results (table 4) do not show a significant impact of stemming, which explains that the system can stand alone.

Table 4: classification results using stemmed keyphrases.

| | Recall | Precision | F-measure |
|---|---|---|---|
| Economics | 0.149 | 0.370 | 0.212 |
| Politics | **0.731** | 0.365 | **0.487** |
| Sport | 0.242 | **0.410** | 0.304 |
| **Average** | **0.373** | **0.382** | **0.378** |

The obtained results show that our approach has an impact on the behavior of the indexes in the classification phase.

Figure 6: Graphic representing the obtained results.




Despite the fact that we were not particularly selective on the keyphrases extraction method, the results are encouraging.

Figure 7: Recall curves depending on each class.

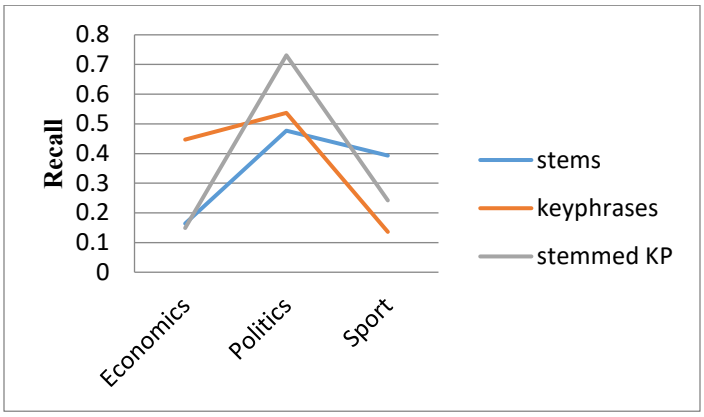
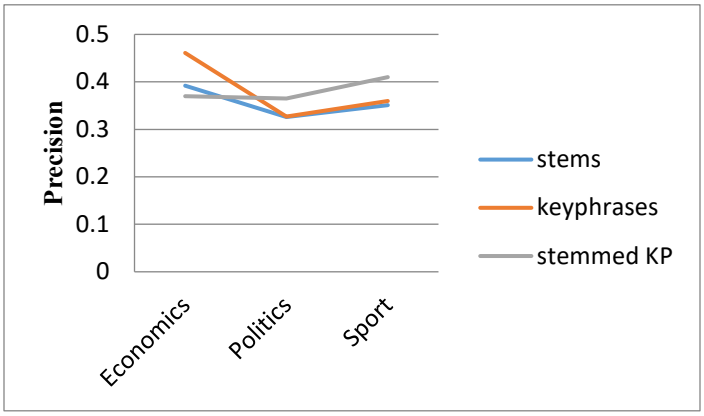



Figure 8: Precision curves depending on each class.




Despite the fact that the results are close for the two approaches (keyphrases and stemmed keyphrases) on average recall and precision (table 3and 4), we clearly distinguish the changes in values according to each domain (figure 7 and 8). This is probably due to the homogeneity of the classes in the corpus. However, our approach reaches up to 73% of recall and 46% of precision in some cases. This leads us to test our approach on other bigger corpus with various classes, and considering adding other improvements.

The experimental results confirms our assumption that keyphrases express better documents than simple keywords. However, this comparison does not, in any way, thwart the approaches presented in related works. Nevertheless, our work could be taken as an advance on what is proposed in the state of the art of Arabic documents indexing.

## 6. Conclusion

. In this paper, we introduced keyphrases based indexing for Arabic documents. After segmenting texts into simple words, we select candidate terms using the co-occurrence method, then we used the TFIDF formula to obtain the final list of indexes. Our objective is to assess the ability of keyphrases to represent a given document.

On one hand, stemming techniques are used in Arabic text preprocessing to reduce multiple forms of the word to one root. However, the results that we have obtained show that the stem's power of discrimination is relatively low. This may be due to the loss of semantic linking between words during the stemming process. On the other hand, we have proposed a method that extract keyphrases in order to represent document. Although the keyphrases extraction method we used is basic, and based on calculation of frequencies, the results of indexation outperform those obtained by the stem based indexing.

In our future works, we will continue to reveal the semantic information by developing new indexing methods. Nonetheless, we consider developing an efficient method in order to extract Arabic keyphrases. Many other improvements are planned for Arabic document indexation.

## References


[1] H. Al-Mahmoud, M. Al-Razgan. "Arabic Text Mining: A Systematic Review of the Published Literature 2002-2014", International Conference on Cloud Computing (ICCC), 2015.

[2] T. F. Gharib, Habib M. B,.and Z. T. Fayed. "Arabic Text Classification Using Support Vector Machines". International Journal of Computers and Their Applications ISCA, vol. 16, no. 4, pages 192–199, 2009.

[3] S. Raheel and J. Dichy. "An empirical study on the feature qs type effect on the automatic classification of arabic documents". In Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing, CICLing'10, pages 673–686, Berlin, Heidelberg, 2010.

[4] M.S. EL Bazzi, T. Zaki, D. Mammass, A. Ennaji. "Stemming Versus Multi-words Indexing For Arabic Documents Classification". 11th International Conference on Intelligent Systems: Theories and Applications, SITA, 2016.

[5] N. Mansour, R.A. Haraty, W. Daher, M. Houri. "An auto-indexing method for Arabic text". Information Processing and Management, volume: 44 issue: 4, pages: 1538-154, 2008.

[6] R. El-Khoribi, A. and M. A. Ismael. "An Intelligent System Based on Statistical Learning For Searching in Arabic Tex"t. ICGST International Journal on Artificial Intelligence and Machine Learning, AIML, vol. 6, pages 41–47, 2006.

[7] A. Al Molijy, I. Hmeidi et I. Alsmadi. "Indexing of Arabic documents automatically based on lexical analysis". International Journal on Natural Language Computing (IJNLC) Vol. 1, No.1, April 2012.

[8] M.J. Bawaneh, M. S. Alkoffash and A. I. Al Rabea. "Arabic Text Classification using K-NN and Naive Bayes". Journal of Computer Science, vol. 4, pages 600–605, 2008.

[9] F Harrag,, E. El-Qawasmah and A. Al-Salman. "Stemming as a Feature Reduction Technique for Arabic Text Categorization". In Proceedings of The 10th International Symposium on Programming and Systems, ISPS 2011, pages 128–133, April 2011.

[10] S.Khoja and R.Garside (1999). "Stemming Arabic Text". Computing Department, Lancaster University, Lancaster, http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps.

[11] M. Y. Al-Nashashibi, D. Neagu, and A. A. Yaghi, "Stemming techniques for Arabic words: A comparative study", 2nd International Conference on Computer Technology and Development, 2010, pp. 270–276.

[12] M. Y. Al-Nashashibi, D. Neagu, and A. A. Yaghi, "An improved root extraction technique for Arabic words", 2nd International Conference on Computer Technology and Development, 2010, pp. 264–269.

[13] M. N. Al-Kabi, "Towards improving Khoja rule-based Arabic stemmer", IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2013, pp. 1–6.

[14] A. Zitouni, A. Damankesh, F. Barakati, M. Atari, M. Watfa, and F. Oroumchian, "Corpus-Based Arabic Stemming Using N-Grams," in Information Retrieval Technology, vol. 6458, P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 280–289.

[15] S. Ghwanmeh, G. Kanaan, R. Al-Shalabi, and S. Rabab'ah, "Enhanced Algorithm for Extracting the Root of Arabic Words," 2009, pp. 388–391.

[16] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents," Inf. Syst., vol. 34, no. 1, pp. 132–144, Mar. 2009.

[17] G. Wei, X. Gao and S. Wu. "Study of Text Classification Methods for Data Sets With Huge Features". In Proceedings of the 2nd International Conference on Industrial and Information Systems, volume 1, pages 433–436, 2010.

[18] I. Mallak. "De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information". Human-Computer Interaction. Université Paul Sabatier – Toulouse III, 2011.

[19] A. Alajmi, E.M. Saad, R.R. Darwish. "Toward an ARABIC Stop-Words List Generation". International Journal of Computer Applications (0975-8887) Volume 46- No.8, May 2012.

[20] I. Hmeidi, B.Hawashin, E.El-Qawasmeh. "Performance of KNN and SVM classifiers on full word Arabic articles". Advanced Engineering Informatics 22 (2008) 106–111.

[21] B. Trstenjak, S. Mikac, D. Donko. "KNN with TF-IDF Based Framework for Text Categorization" . 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

[22] T. Zaki, D. Mammass, A. Ennaji. "A semantic proximity based system of Arabic text indexation". International Conference on Image and Signal Processing (ICISP) 2010.

[23] A. Mountassir. "Sentiment Analysis: Classification supervisée de documents arabes". Proceedings of 7th International Conference on Intelligent Systems : Theories and Applications. May 16-17, 2012, Mohammedia, Morocco.

[24] S. Alsaleem. "Automated Arabic Text Categorization Using SVM and NB". International Arab Journal of e-Technology, vol. 2, no. 2, 2011.

[25] B. Al-Salemi. and M. J. A. Aziz. "Statistical Bayesian Learning For Automatic Arabic Text Categorization". Journal of Computer Science, vol. 7, no. 1, pages 39–45, 2011

[26] M. Nabil, M.Aly, Amir F. Atiya. "New Approaches for Extracting Arabic Keyphrases". 2015 First International Conference on Arabic Computational Linguistics.

[27] N. G. Ali, N. Omar. "ARABIC KEYPHRASES EXTRACTION Using a Hybrid of STATISTICAL AND MACHINE" Learning Methods. International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia

[28] Mihalcea R. et P. Tarau. Textrank: Bring- ing order into texts. In Proceedings of EMNLP, pages 404–41, 2004.

[29] S. R El-Beltagy and Ahmed Rafea. Kp-miner: A keyphrase extraction system for english and arabic documents. Information Systems, 34(1):132–144, 2009.

[30] El-shishtawy T.A. & Al-sammak A.K. "Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques". Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2009

[31] Hassan M. Najadat, Mohammed N. Al-Kabi, Ismail I. Hmeidi, Maysa Mahmoud Bany Issa. "Automatic Keyphrase Extractor from Arabic Documents". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016

[32] Liu, Z. , Peng, L. , Yabin, Z. et S. Maosong. 2009. "Clustering to find exemplar terms for keyphrase extraction". In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 257–266.

[33] Kamal Sarkar. "A Hybrid Approach to Extract Keyphrases from Medical Documents". International Journal of Computer Applications (0975 – 8887) Volume 63– No.18, February 2013