Improving Arabic Text Categorization using Normalization and Stemming Techniques

Rouhia M. Sallam
Faculty of Applied Sciences
Taiz University
Yemen

Hamdy M. Mousa
Faculty of Computers And Information
Menoufia University
Egypt

Mahmoud Hussein
Faculty of Computers And Information
Menoufia University
Egypt

ABSTRACT
Text Categorization is a technique for assigning documents based on their contents to one or more pre-defined categories. Achieving highest categorization accuracy remains one of the major challenges and it is also time consuming. We proposed approach to tackle these challenges. The proposed approach uses Frequency Ratio Accumulation Method (FRAM) as a classifier. Its features are represented using bag of word technique and an improved Term Frequency (TF) technique is used in features selection. The proposed approach is tested with known datasets. The experiments are done without both of normalization and stemming, with one of them, and with both of them. The obtained results of proposed approach are generally improved compared to existing techniques.The performance attributes of proposed Arabic Text Categorization approach were considered: Accuracy, Recall, Precision and F-measure (F1). The averages of the obtained results are 97.50%, 97.50%, 97.51%, and 97.49% respectively using normalization.

Keywords
Arabic text categorization, Frequency ratio accumulation method (FRAM), Bag-Of-Word (BOW), Features selection, Term and document frequency.

1. INTRODUCTION
Text categorization (TC) is the task of automatically structuring a set of text documents into different categories according to a group structure that is known in advance [1]. Due to There are a tremendous number of text documents available online that is growing each day, text categorization is a very important and a fast growing research field.

The development of text classification systems for Arabic documents is a challenging task due to the complexity and the rich nature of the Arabic language. The language consists of 28 letters and is written from right to left. It has very complex morphology, and the majority of words have a tri-letter root. The other languages have a quad-letter root, a penta-letter root or a hexa-letter root [5].

A number of approaches have been proposed for automatic text categorization such as Support Vector Machines (SVM), K- Nearest Neighbor (KNN), Neural Networks (NN), Naïve Bayes (NB), Decision Trees (DT) Maximum Entropy (ME), and Association Rules [2, 26, 9, 28]. Most of these techniques have complex mathematical models and do not usually lead to accurate results for the text categorization [3].

To improve Arabic text categorization, this paper is proposed approach that used a simple mathematical model which called Frequency Ratio Accumulation Method (FRAM) [3]. In this approach, the Bag-Of-Word (BOW) technique is used to represent the features and an improved Term Frequency (TF) technique is used in features selection. The terms frequencies are sorted according to the largest frequency and then the highest 25% and 50% terms are used as features without and with stemming respectively. Normalization and stemming approaches are also used. For stemming, ISRI (The Information Science Research Institute's) and Tashaphyne Stemmer are used [15, 16, 29, 18]. The proposed approach achieved a high rate of accuracy classification for Arabic text categorization. Its accuracy is 97.50% using normalization and 93.06% using ISRI and 95.83% using Tashaphyne Stemming.

This paper is organized as follows. Section 2 briefly describes related works in the area of automatic text categorization. The proposed approach is described in Section 3. Section 4 outlines the experimental setting and obtained results. Conclusion and future work are presented in Section 5.

2. RELATED WORK
Frequency Ratio Accumulation Method (FRAM) to classify Arabic text documents is introduced and the accuracy achieved is 94.1% with the use of 400 features selected by CHI-FS method [3].

Suzuki and Hirasawa have introduced an approach to classify English and Japanese text documents automatically using FRAM [4]. The technique is evaluated through a number of experiment using newspaper articles from Japanese CD-Mainichi 2002, and English Reuters-21578. The average accuracy reported was 86.1% for EnglishReuters-21578 and 87.3% for Japanese text documents using the character N-gram and the word N-gram as feature terms respectively.

Al-Shargabi has compared three techniques for Arabic text classification to measure the accuracy for each classifier and to determine which classifier is more accurate for Arabic text classification based on stop words elimination [6]. These techniques are: Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Naïve Bayesian (NB), and J48 [29]. Accuracy using percentage split method for SVM classifier;J48 classifier andNB classifier achieve 94.8%,89.42% and 85.07% [6].

Support vector machine (SVM), Decision Trees (C4.5), Naive Bayes (NB) Classifiers and stemming are used to classify Arabic documents [22]. Without applying stemming, the results achieved 87.79%, 88.54% accuracy with SVM and with Naïve Bayes respectively. On the other hand, the results with applied stemming achieved lower accuracy are 84.49% and 86.35%.

Multiple methodologies for categorization of text documents automatically are introduced [7]. These methodologies

combine the Bag-of-Words and the Bag-of-Concepts text representation patterns with the used of Wikipedia as a source of knowledge. Three distinctive instrument learning based classifiers were used.

The author is proposed classification approach using SVMs. This Arabic text documents classifier uses CHI square method as a feature selection method in the pre-processing step. The author also used normalization but not stemming because it is not always beneficial for text categorization since many terms may be conflated to the same root. The accuracy rate for this approach is 88.11% [10].

Comparisons two distinct datasets using roots derived from three different rooting libraries for Arabic Sentiment Analysis is introduced [11]. The obtained accuracy is 92.2% with Khoja Stemmer, 93.2% with Tashaphyne Stemmer, and 92.6% with ISRI Stemmer. The best stemmer is Tashaphyne by utilization Opinion Corpus for Arabic [11].

The proposed system analyzes the test documents instead of questions. It utilizes sentence splitting, root expansion, and semantic expansion using an automatically generated ontology. Three stemmers are used: Khoja, ISRI and Tashaphyne on Arabic language question answer selection in machines (ALQASIM 2.0). The results showed an improvement in performance by using ISRI root stemmer [12].

Vector of a weighted frequency for each of the distinct words or tokens represents documents in classification process. This representation of text is quite effective for a number of applications any simply [8, 9].

Due to the results of the previous work; the technique using a simple mathematical classifier (FRAM) and the best two stemming approaches is proposed.It is alsoused simple and effective technique Bag-Of-Word (BOW) and an improved Term Frequency (TF) technique in the features selection. Furthermore,normalization without stemming is usedto improve performance.

3. PROPOSED APPROACH

The main pedestal of the stages for Arabic text categorization is shown in Figure 1. The process is divided into two phases: training and testing. In each phase, the text documents are pre-processed, features are selected, the FRAM classifier is applied, and the results are evaluated.

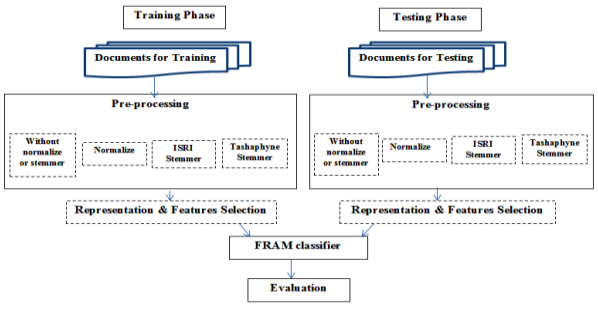

Figure 1: An improved Arabic text categorization using normalization and stemming

3.1 Text Preprocessing

In the preprocessing stage, the text documents are converted to UTF-8 encoding.Then, The Arabic stop words are removed. Such words (e.g. pronouns and prepositions) are not useful in the text categorization. For example,

(.....,بالإشارة,بالنسبة ,الاخري,
ممكن,غير ها,اولئك,بعض,كذلك,كهناك,حاليا)

In addition, some Arabic documents may contain foreign words, special characters, numbers (e.g. ',: ,? , \, $, 1, and 2) [13, 14]. Finally, words with length less than three letters are eliminated. Often these words are not important and are not useful in TC.

3.2 Normalization

For normalization, applied a very efficient normalization technique (i.e. Tashaphyne normalization) [15]. The following are the set of rules used:

Strip Diacritics	ex:	العربية	←	الْعَرَبِيَّةُ
Strip Elongation	ex:	العربية	←	العـــربية
Normalize Hamza	ex:	ء	←	ؤ or ىء
Normalize Alef	ex:	ا	←	آ or أ or إ
Normalize Yeh	ex:	ى	←	ي
Normalize Heh	ex:	ة	←	ه

3.3 Stemming

Two efficient stemming algorithms: Information Science Research Institute's (ISRI) stemmer and Tashaphyne stemmer are applied. Because of they are better Performance in comparison with other stemmer [11, 12].

3.3.1 ISRI Stemmer

In 2005, Kazem et. al. have proposed the Information Science Research Institute's (ISRI) stemmer that shares many features with the Khoja stemmer [16]. It uses a similar algorithm to word rooting of Khoja stemmer. However, it does not employ a root dictionary for lookup. In addition, if a word cannot be rooted, it is normalized by the ISRI stemmer (e.g. removing certain determinants and end patterns) instead of leaving the word unchanged. Furthermore, it defines sets of diacritical marks and affix classes. The ISRI stemmer has been shown to give good improvements to language tasks such as document clustering [17].

Example:

Term	***ISRI***
أفتضربونني	افتضربون
تستلزم	لزم
مكتبة	كتب
محامون	حام

3.3.2 Tashaphyne Light Arabic Stemmer

The Tashaphyne stemmer normalizes words in preparation for the "search and index" tasks required by the stemming algorithm. It removes diacritics and elongation from input words [18]. Then, segmentation and stemming of the input is performed using a default Arabic affix lookup list for various levels of stemming and rooting [18].

Tashaphyne Light Arabic Stemmer provides a configurable stemmer and segmented for Arabic text.

Example:

Term	***Tashaphyne stemmer***
أفتضربوننني	ضرب
تستلزم	لزم
مكتبة	كتب
محامون	حام

3.4 Representation and Features Selection

The representation "Bag-Of-Word" BOW is the most popular document representation scheme in text categorization. In this model, a document is represented as a bag of the terms occurring in it and different terms are assumed to be independent with each other. BOW model is simple and efficient [19].

We are dealing with a huge feature spaces. Therefore, a feature selection mechanism is needed. The most popular feature selection method is document frequency [25].

This paper is used this method with slight modification as the following. First, calculate the frequencies for every term in all categories and sort the frequencies according to the largest frequency. Second, take top 25% of the features when normalization and stemming are not used and take 50% of the features otherwise. These two percentagesdefined experimentally. Third, calculate frequency ratio (FR) by FRAM classifierin each category as follows [3]:

$$FR(t_n, c_k) = \frac{R(t_n, c_k)}{\sum_{c_k \in C} R(t_n, c_k)} \quad (1)$$

Where, the ratio (R) of each feature term for each category is calculated by:

$$R(t_n, c_k) = \frac{f_{c_k}(t_n)}{\sum_{t_n \in T} f_{c_k}(t_n)} \quad (2)$$

Here, $f_{ck}(t_n)$ refers to the total frequency of the feature term t_n in a category *ck*. Thus, in the training phase, the FR of all feature terms are calculated and supported in each category. Then, calculate the category evaluation values or category score, which indicates the possibility that the candidate document in the testing phase belongs to the category as follows:

$$E_{d_i}(c_1) = \sum_{t_n \in d_i} FR(t_n, c_k) \quad (3)$$

Finally, the candidate document d_i is classified into the category $c_{\wedge k}$ for which the category score is the maximum, as follows:

$$c_{\wedge k} = argmax_{c_k \in c} E_{d_i}(c_k) \quad (4)$$

Figure 2 shows a comparison between our proposed technique and the proposed approach by [3]. It also shows a comparison between the classification accuracy of both approaches.

	Proposed Approach in [3]	**Our Proposed Approach**
Normalize	Used Normalize	Used Normalize Tashaphyne
Stemming	Remove only prefixes and suffixes	1- ISRI Stemmer 2- Tashaphyne Stemmer
Representation &Features Selection	Bag-Of-Word and CHI in FS	Bag-Of-Word and (sorted& ratio) in FS
classifier	FRAM	FRAM

Figure 2: Comparison between our approach and proposed approach by [3]

4. EXPERMINTAL AND RESULTS

Implementation the proposed methodology using Python 3.4.2 [24]. In addition, our experiments are conducted on a SONY laptop with the following specifications: 2.5 GHz Intel core i5 processor with 4 GB of RAM, and windows 8 enterprise.

Used four standard evaluations; the accuracy of the text categorization approaches is computed by the equation [23]:

$$Accuracy = \frac{\text{Number of correctly identified documents}}{\text{Total number of documents}} \quad (5)$$

$$\textbf{Recall (R)} = \frac{TP}{TP+FN} \quad (6)$$

$$\textbf{Precision (P)} = \frac{TP}{TP+FP} \quad (7)$$

$$\textbf{F –measure} = \frac{2*P*R}{P+R} \quad (8)$$

Precision, Recall and F-measure are defined in [29] as follows:

Where:

- TP: number of documents which are correctly assigned to the category.
- FN: number of documents which are not falsely assigned to the category.
- FP: number of documents which are falsely assigned to the category.
- TN: number of documents which are not correctly assigned to the category.

Three different known data sets (i.e. Dataset 1, Dataset 2, and Dataset 3) collected from the website www.aljazeera.net [20], [21] are used to evaluate the efficiency of the proposed approach for Arabic text categorization.

Dataset1consists of 1800 documents that are separated into six categories: art, health, religion, law, sport, and technology.

Dataset2 consists of 1500 documents separated into five categories: arts, economic, politics, science and sport. Dataset3has 1200 documents which are separated into four categories: international, literature, science and sport. The datasets are divided into 70% of the documents are used for training while 30% of the documents are used for testing.

Table 1: Datasets used for approach evaluation

Datasets	Number of documents	Training set	Testing set
Dataset1	1800	1260	540
Dataset2	1500	1050	450
Dataset3	1200	840	360

In Table 2 and Figure 3, show the classification accuracy of the different variants of our proposed approach (i.e. without normalization or stemming, with normalization, and with normalization and stemming) and the approach proposed in [3] with the three different datasets .

The results show that the highest accuracy achieved when normalization and stemming are not used is 96.66% with Dataset 2, while it is 97.50% when normalization is used. In case of the use of stemmers, the highest accuracy is 93.06% and 95.83% with Dataset 3 when ISRI and Tashaphyne stemmers are used respectively. Finally, the approach proposed in [3] achieved accuracy of 94.44% with Dataset 3 (see Table 2 and Figure 3).

Table 2: Results of accuracy for the proposed approach and the proposed approachin [3]

Datasets	Without normalization or Stemming	Normalization	ISRI stemmer	Tashaphyne stemmer	Proposed [3]
Dataset1	96.30%	96.67%	92.96%	95.0%	91.94%
Dataset2	96.89%	97.33%	88.89%	95.33%	93.77%
Dataset3	95.56%	97.50%	93.06%	95.83%	94.44%

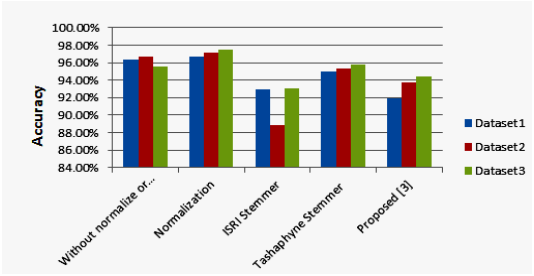

Figure 3: Results of accuracy for the proposed approach and the proposed approach in [3]

In Table 3, the results show that the highest precision achieved is 100% in sport category when Tashaphyne stemmer is used. Also the results shows that the highest recall, precision and F-measure achieved when normalization and stemming are not used is 98.9% with sport category, and it is the same percentage when normalization is used with art and sport categories. In case of the use of stemmers, the highest recall is 98.9% in sport category when ISRI and Tashaphyne stemmers are used (see Figure 4, accuracy results with Dataset 1).

Table 3: Results of Recall, Precision and F1 for the proposed approach for Dataset1

	Without normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	0.978	0.979	0.978	0.989	0.978	0.983	0.900	0.988	0.942	0.978	0.989	0.983
Health	0.978	0.946	0.962	0.967	0.956	0.961	0.888	0.954	0.919	0.922	0.933	0.927
Law	0.933	0.966	0.949	0.967	0.936	0.951	0.944	0.833	0.885	0.989	0.839	0.908
Religion	0.956	0.935	0.945	0.956	0.977	0.966	0.922	0.902	0.912	0.922	0.988	0.954
Sport	0.989	0.989	0.989	0.989	0.989	0.989	0.989	0.967	0.978	0.989	1.000	0.994
Technology	0.944	0.966	0.955	0.933	0.965	0.949	0.933	0.956	0.944	0.900	0.976	0.936
Average	96.30	96.32	96.29	96.67	96.69	96.66	92.96	93.29	93.01	95.00	95.42	95.06

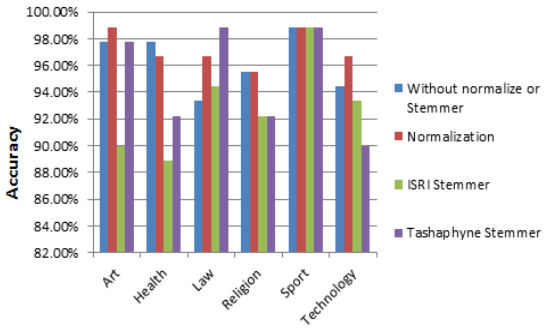

Figure 4: Results of accuracy for the proposed approach for Dataset1

The results in table 4 gives that the highest recall and precision achieved when normalization and stemming are not used is 100 % with economic, science , sport and politics categories, and it is the same percentage when normalization is used with economic, science and sport categories. When stemmers are used, the highest recall and precision is 100% in economic, politics and Science categories when ISRI stemmer is used. Also when Tashaphyne stemmer is used achieved highest recall is100% in sport category (see Figure 5, accuracy results with Dataset 2).

Table 4: Results of Recall, Precision and F1 for the proposed approach for Dataset2

	Without normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	0.989	0.979	0.984	0.979	0.979	0.979	0.867	0.934	0.902	0.978	0.967	0.972
Economic	1.000	0.918	0.956	1.000	0.938	0.968	1.000	0.677	0.807	0.989	0.881	0.932
Politics	0.856	1.000	0.922	0.889	0.988	0.936	0.678	1.000	0.808	0.822	0.961	0.887
Science	1.000	0.978	0.989	1.000	0.989	0.994	0.922	1.000	0.960	0.978	0.989	0.983
Sport	1.000	0.978	0.989	1.000	0.978	0.989	0.978	0.978	0.978	1.000	0.978	0.989
Average	96.89	97.06	96.82	97.33	97.40	97.29	88.89	91.87	89.09	95.33	95.53	95.26

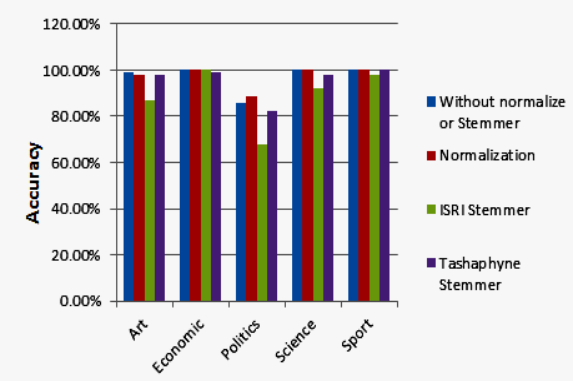

Figure 5: Results of accuracy for the proposed approach for Dataset2

In Table 5, the results shows that the highest recall achieved when normalization and stemming are not used is 98.9% with sport category, and it is the same percentage the highest Precision when normalization is used with science and sport categories. When stemmers are used, the highest Precision is 98.9% in science category when ISRI stemmer and the highest recall is 100% in sport category when Tashaphyne stemmer is used (see Figure 6, accuracy results with Dataset 3).

Table 5: Results of Recall, Precision and F1 for the proposed approach for Dataset3

	Without normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
International	0.900	0.976	0.937	0.956	0.978	0.967	0.933	0.955	0.945	0.956	0.935	0.945
Literature	0.956	0.945	0.950	0.967	0.957	0.961	0.911	0.891	0.901	0.933	0.966	0.949
Science	0.978	0.934	0.957	0.989	0.989	0.989	0.966	0.989	0.976	0.944	0.966	0.955
Sport	0.989	0.978	0.983	0.989	0.978	0.983	0.911	0.901	0.906	1.000	0.979	0.989
Average	95.56	95.88	95.67	97.50	97.51	97.49	93.05	93.39	93.21	95.83	96.11	95.95

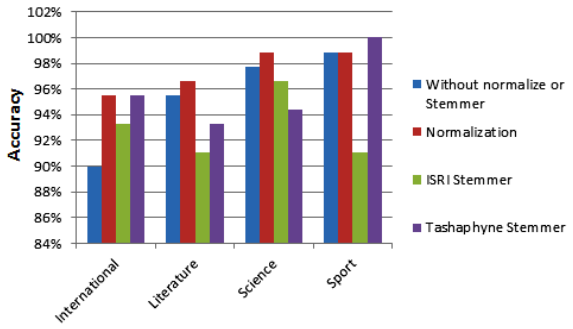

Figure 6: Results of accuracy for the proposed approach for Dataset3

Table 6 shows the results for execution time for all stages of classification with different datasets and the four experiments for our approach and the approach in [3]. Proposed approach took less execution time in all experiments and all datasets. Thus, it is reduced memory usage and power consumption. It was less time for execution with the ISRI stemmer where it took is 46 seconds in Dataset3. While, the proposed approach in [3] is 2820 seconds in with Dataset3.

Table 6: Results the time of executionfor the proposed approach and the approach in [3]

Datasets	Without normalization or Stemming	Normalization	ISRI stemmer	Tashaphyne stemmer	Proposed [3]
Dataset1	133s	125s	71s	80s	8280s
Dataset2	88s	85s	56s	64s	5040s
Dataset3	77s	74s	46s	51s	2820s

5. CONCLUSION

In this paper, the accuracy of Arabic text categorization has been improved by applying an improved features selection technique and using the Frequency Ratio Accumulation Method classifier with normalization and two stemming mechanisms: ISRI and Tashaphyne stemmers.

The results are shown that applying text classification with normalization achieved the highest classification accuracy of 97.50% while it is 96.66% when normalization and stemming are not used. On the other hand, with stemmers less classification accuracy is achieved where the accuracy is 95.83% with Tashaphyne stemmer and 93.06% with ISRI stemmer.

As a future work, applied several approaches that have been applied to English text categorization but have not yet been applied to Arabic text categorization.

6. REFERENCES

[1] Tripathi N., 2012.Level Text Classification Using Hybrid Machine Learning Techniques. PhD thesis, University of Sunderland.

[2] Harrag F., El"Qawasmeh E., 2009.Neural Network for Arabic text classification. 778 – 783.

[3] Sharef B., Omar N., and Sharef Z., 2014. An Automated Arabic Text Categorization Based on the Frequency Ratio Accumulation.*The International Arab Journal of Information Technology*, Vol. 11, No. 2, March 2014, 213-221.

[4] Suzuki M., Hirasawa S.,2007. Text Categorization Based on the Ratio of Word Frequency in Each Categories.*In Proceedings of IEEE International Conference on Systems Man and Cybernetics*, Montreal, Canada, 3535-3540.

[5] Laila, K., 2006. Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study.*Conference on Data Mining | DMIN'06* |,78-82

[6] Al-Shargabi B., AL-RomimahW.andOlayah F.,2011. A Comparative Study for Arabic Text Classification Algorithms Based on Stop Words Elimination. ACM, Amman, Jordan 978-1-4503-0474-0/04/2011.

[7] Nezreg H.,Lehbab H., and BelbachirH. ,2014.ConceptualRepresentation Using WordNet for Text Categorization.*International Journal of Computer and Communication Engineering, Vol. 3, No. 1, January 2014.*

[8] Diederich J., KindermannJ., Leopold E. and Paass G.,2003. Authorship attribution with supportvector machines. Applied Intelligence, 109-123. (2003).

[9] Sebastiani F, 2002. Machine learning in automated text categorization. *ACM Computing Surveys,Vol. 34 number 1. 1-47.*

[10] Mesleh A. A., 2007.Chi Square Feature Extraction Based Svms Arabic Language TextCategorization System. *Journal of Computer Science 3(6): 430-435.*

[11] Oraby Sh., El-SonbatyY.and El-Nasr M., 2013. Exploring the Effects of Word Roots for Arabic Sentiment Analysis.,*International Joint Conference on Natural Language Processing, 471–479,Nagoya, Japan, 14-18 October 2013.*

[12] Ezzeldin A., El-Sonbaty Y. and Kholief.M, 2013.Exploring the Effects of Root Expansion.College of Computing and Information Technology, AASTMT Alexandria, Egypt.

[13] Al-ShalabiR.,Kanaan G., Jaam J.M, HasnahA.andHilat E. 2004.Stop-word Removal Algorithm for Arabic Language. *Proceedings of 1st International Conference on Information & Communication Technologies: from Theory to Applications, IEEE-France, 545-550,CTTA'04*

[14] El-Kourdi M., Bensaid A. and Rachidi T.,2004.Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. *20th International Conference on Computational Linguistics. August, Geneva.*

[15] https://pythonhosted.org/Tashaphyne/Tashaphyne.normalize-module.html

[16] Kazem T., Rania E., and Je.rey C., 2005.Arabic Stemming Without A Root Dictionary. Information Science Research Institute, USA.

[17] Kreaa A., Ahmad A. and KabalanK.,2014 . ARABIC WORDS STEMMING APPROACH USING ARABIC ORDNET.*International Journal of Data Mining &*

Knowledge Management Process (IJDKP) Vol.4, No.6, November 2014.

[18] https://pypi.python.org/pypi/Tashaphyne/

[19] Pu W,.Liu.N,2007.Local Word Bag Model for Text Categorization.*Seventh IEEE International Conference on Data Mining,625-630.*

[20] Abuaiadah D.,El-Sana J.,Abusalah W. 2014.On the Impact of Dataset Characteristics on ArabicDocument Classification. *International Journal of Computer Applications (0975 – 8887)Volume 101– No.7.*

[21] https://sites.google.com/site/mouradabbas9/corpora

[22] Wahbeh A., Al-Kabi M., Al-Radaidah Q., AlShawakfa E. and Alsamdi. I. 2011. The Effect of Stemming on Arabic Text Classification: An Empirical Study. *In International Journal of Information Retrieval Research (IJIRR), vol. 1, no. 3, I. 2011,54-70.*

[23] M. Turk, and A. Pentland.1991.Eigenfaces for recognition. *Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71 -86.*

[24] https://www.python.org/downloads/

[25] GarnesO.,2009. Feature Selection for TextCategorization.Master thesis,Norwegian University of Science and Technology, June 2009.

[26] Sawaf H., Zaplo J., and Ney H.,2001.Statistical Classification Methods for Arabic News Articles.,Workshop on Arabic Natural Language Processing, ACL'01, Toulouse, France, July 2001.

[27] Yang Y. Liu and X.,1999.A Re-examination of Text Categorization Methods.,*Proceedings of 22nd ACM International Conference on Research and Development in Information Retrieval,SIGIR'99, ACM Press, New York, USA, 1999, 42-49.*

[28] *http://www.nltk.org/_modules/nltk/stem/isri.html*

[29] Elhassan R., Ahmed M.2015.Arabic Text Classification on Full Word.*International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 5, May 201 5, 114-120.*