I.J. Information Technology and Computer Science, **2016, 11, 74-81**
Published Online November 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.11.09

Arabic Text Categorization Using Mixed Words

Mahmoud Hussein, Hamdy M. Mousa
Faculty of Computers and Information, Menoufia University, Egypt
E-mail: {mahmoud.hussein@ci.menofia.edu.eg, hamdimmm@hotmail.com}

Rouhia M.Sallam
Faculty of Applied Sciences, Taiz University, Yemen
E-mail: rohiya79@yahoo.com

Abstract—There is a tremendous number of Arabic text documents available online that is growing every day. Thus, categorizing these documents becomes very important. In this paper, an approach is proposed to enhance the accuracy of the Arabic text categorization. It is based on a new features representation technique that uses a mixture of a bag of words (BOW) and two adjacent words with different proportions. It also introduces a new features selection technique depends on Term Frequency (TF) and uses Frequency Ratio Accumulation Method (FRAM) as a classifier. Experiments are performed without both of normalization and stemming, with one of them, and with both of them. In addition, three data sets of different categories have been collected from online Arabic documents for evaluating the proposed approach. The highest accuracy obtained is 98.61% by the use of normalization.

Index Terms—Arabic Text Categorization, Frequency Ratio Accumulation Method, Term and Document Frequency, Features Selection, and Mixed Words.

I. INTRODUCTION

Text Categorization (TC) is an automatic process for grouping documents based their contents into pre-defined categories that are known in advance [1]. There is a tremendous numbers of Arabic text documents available online which is growing every day. Consequently, text categorization becomes very important and a fast growing research field. The developments of such text classification systems for Arabic documents are a challenging task because of the complexity of the Arabic language and its very complex morphology and high inflection. The language consists of 28 letters and is written from right to left. Most of the Arabic words have a tri-letter root [2]. There is still a limited research for the Arabic text categorization due to the complex and rich nature of the Arabic language compared to other languages [3, 4].

There are several different techniques for automatic text classification including Support Vector Machines (SVM), K- Nearest Neighbor (KNN), Neural Networks (NN), Decision Trees (DT), Maximum Entropy (ME), Naïve Bayes (NB), and Association Rules [5, 6, 7, 8]. Most of these techniques have complex mathematical and statistical models and power consuming and do not usually lead to accurate results for the categorization [9].

The proposed approach aims to improve Arabic text categorization by using a new technique that is the mixed bag of words and two adjacent words collected with different proportions to represent the features. Term Frequency (TF) technique is used in features selection. In addition, a simple mathematical model is used which called Frequency Ratio Accumulation Method (FRAM). Normalization and stemming approaches are also used.

The two adjacent words are composed of two terms that together refer to a concept where each word is taken separately refers to another unrelated concept. To illustrate this, consider the concept "فقرالدم" (Anemia), two words together refer to illness but the word "فقر" in Arabic means (Poverty) and the other word "الدم" means (Blood). Therefore, separation of the words may refer to another unrelated concept that reduces the classification accuracy.

This paper is organized as follows. In Section 2, an overview of the related work is presented. Section 3 introduces the proposed approach. Section 4 presents the experimental results. Finally, conclusions and future work are put forward in Section 5.

II. RELATED WORK

Arabic text documents available online are growing every day. Arabic language has complex internal word structures and the complicated construction of Arabic words from roots. Addition to the mentioned above, the problem of diacritics and text normalization. Thus, there are few work in Arabic text categorization. Recently, a number of approaches have been proposed for automatic Arabic text categorization. In the following, we describe and analyze these approaches.

Laila Khreisat has used N-gram frequency statistics in Arabic text classification [2]. The text corpus is collected from online Arabic newspapers where 40% of the text is used as training and 60% for testing. The experiments compared the performance of tri-gram technique using the Manhattan measure and Dice measure [10]. The results for the tri-gram method using the Dice measure exceed Manhattan measure reaching its highest recall value of 1 for the weather category.

Sharef et al. introduced a new Frequency Ratio

Accumulation Method (FRAM) approach for classifying Arabic text documents [9]. Using the BOW as feature representation and CHI to select these features, the result achieved is 94.1%.

Suzuki and Hirasawa proposed a new classification technique called Frequency Ratio Accumulation Method (FRAM) [11]. N-gram character and N-gram word are used as feature terms. The technique is evaluated through a number of the experiment using newspaper articles from Japanese CD-Mainichi 2002 and English Reuters-21578. The results accuracy is 87.3% for the Japanese CD-Mainichi 2002 and 86.1% for the English Reuters-21578.

Hadni et al. introduced a new method for Arabic multi word terms (AMWTs) extraction based on a hybrid approach [12]. They used linguistic AMWTs approach to extract the candidate MWTs based on Part Of Speech (POS). A statistical approach is also used to incorporate contextual information by using a proposed association measure based on Term-hood and Unit-hood for AMWTs extraction. They used three statistical measures C-Value, NC-Value and NTC-Value [13]. For evaluation they have used two steps: reference list and validation.

Al-Shargabi compares three techniques for Arabic text classification based on stop words elimination [14]. These techniques are Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Na ïve Bayesian (NB), and J48 [6].The results of accuracy using these techniques achieved 94.8%, 89.42% and 85.07% respectively.

Diab has used multi-word features in the Arabic document classification and two similarity functions [15]: the cosine and the dice similarity functions. He also applied inverse document frequency (IDF) to prevent frequent terms from dominating the value of the function and he used different light stemmers on multi-word features.

An approach for automatic Arabic text classification is introduced by applying stemming and without using stemming [16]. Using Support vector machine (SVM), Decision Trees (C4.5), Naive Bayes (NB) Classifiers. The results achieved 87.79%, 88.54% accuracy with SVM and Na ïve Bayes respectively when stemming is used. On the other hand, the results without application of stemming achieved lower accuracy are 84.49% and 86.35%.

Ezreg et al. proposed a conceptual representation based on terms disambiguation for text classification using WordNet [17]. They have used three representations (terms, concepts, terms + concepts) with three classifiers: SVM, Decision trees and KNN. Decision trees give better results than SVM and k-NN.

Abdelwadood Mesleh proposed an approach for Arabic text classification using SVMs [18]. The proposed approach used CHI square method as a feature selection. He 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 experiments results showed classification effectiveness of 88.11%.

Zhang et al. have used a multi-word technique for features representation with support vector machine as classifier to improve document classification [19]. Two strategies were developed for feature representation based on the different semantic level of the multi-words. The first is the decomposition strategy using general concepts and the second is combination strategy using subtopics of the general concepts.

Oraby et al. have used three different rooting libraries to derive the roots of each of the input words for Arabic sentiment analysis [20]: Khoja Arabic Stemmer, Information Science Research Institute (ISRI) Arabic stemmer and Tashaphyne Light Arabic Stemmer. The best stemmer is Tashaphyne by utilization Opinion Corpus for Arabic (OCA). The obtained accuracy is 92.2% with Khoja stemmer, 93.2% with Tashaphyne stemmer, and 92.6% with ISRI stemmer.

Ezzeldin et al. have introduced an approach for Arabic answer selection [21]. It analyzes the test documents instead of questions, and utilizes sentence splitting, root expansion and semantic expansion by 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.

A lot of the approaches in text classification treat documents as a bag-of-words with the text represented as a vector of a weighted frequency for each of the distinct words or tokens. This simplified representation of text has been shown to be quite effective for a number of applications [6, 22].

III. Proposed Approach

The Arabic language differs syntactically, morphologically and semantically from Latin languages. Arabic is written from right to left and has twenty eight letters that compose of twenty five consonants and three vowels. The Arabic letters take different shapes dependent on their position in the word. Arabic language has vowel diacritics that are written above or under letter that give the desired sound and meaning of word. Due to the increase of availability of digital Arabic documents and the important need of automated text categorization, many approaches are proposed. But, they did not achieve researchers' satisfaction and have high computation cost. For enhancing the automatic Arabic text classification, we have proposed an approach. The main steps/stages of our approach are shown in Fig. 1. The stages include: Arabic text pre-processing, normalization, stemming, and feature representation and selection. These stages are used in both: training and testing phases. In the following, we describe these stages in detail.

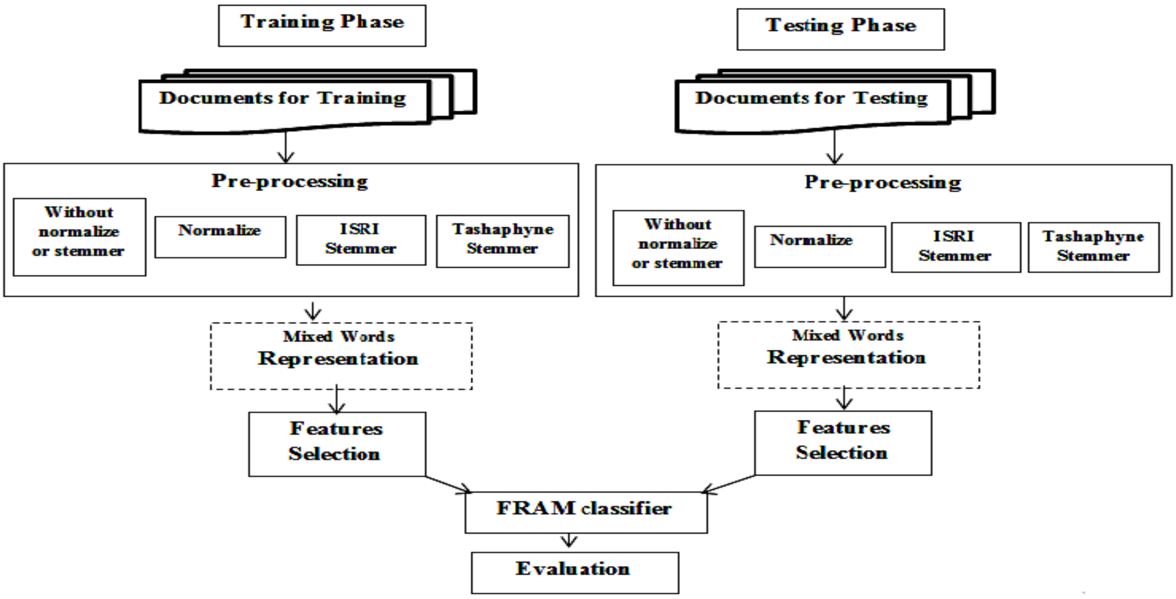

Fig.1. Arabic Text Categorization Using Mixed Words

A. *Text Preprocessing*

This stage necessary due to the variations in the way that text can be represented in Arabic. First, the text documents are converted to UTF-8 encoding. Then, The Arabic stop words are removed. Some Arabic documents may contain foreign words, special characters, numbers [23, 24]. Finally, words with length less than three letters are eliminated, often these words are not important and are not useful in TC.

B. *Normalization*

For normalization, a very efficient normalization technique is applied (i.e. Tashaphyne normalization) [25]. Normalization of some Arabic letters such as “ة” to “ه” and “ي” to “ى” and “آ, أ, إ ” to “ا”. In addition, diacritics such as “تَشْكِيلُ” to “تشكيل”, and elongation“محمـــــــد” to “محمد” is performed.

C. *Stemming*

We applied two efficient stemming algorithms: Information Science Research Institute’s (ISRI) stemmer and Tashaphyne stemmer. Because they have better performance in comparison with other stemmers [20, 21].

D. *ISRI Stemmer*

The Information Science Research Institute’s (ISRI) Arabic stemmer shares many features with the Khoja stemmer [26]. 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 [27].

E. *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 [28]. Then, segmentation and stemming of the input is performed using a default Arabic affix lookup list for various levels of stemming and rooting [28].

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

F. *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 of each other. BOW model is simple and efficient [29].

A lot of work has been done to extract MWT in many languages. Many of researchers use AMWTs features to improve Arabic document classification [12, 13, 30, 31].

We are dealing with a huge feature spaces. Therefore, a feature selection mechanism is needed. The most popular feature selection method is term frequency [32].

In this paper, we have used this method with some modification as the following. First, the frequencies for every term in all categories are calculated and sorted according to the largest frequency. Second, when BOW is applied, we take the top 25% of the features when normalization and stemming are not used and take 50% of the features otherwise. Third, when mixed words are applied, we take the top 50% of the features from BOW, and take the top 3% from two adjacent words in all experiments. These percentages have been defined experimentally. Finally, the frequency ratio (FR) is calculated by the FRAM classifier in each category as follows [9]:

$$\mathrm{FR}(\mathrm{t_n}, \mathrm{c_k}) = \frac{\mathrm{R}(\mathrm{t_n},\mathrm{c_k})}{\sum \mathrm{c_k} \in \mathrm{C}\ \mathrm{R}(\mathrm{t_n},\mathrm{t_k})} \tag{1}$$

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

$$\mathrm{R}(\mathrm{t_n}, \mathrm{c_k}) = \frac{\mathrm{f_{ck}}(\mathrm{t_n})}{\sum \mathrm{t_n} \in \mathrm{T}\ \mathrm{fc_k}(\mathrm{t_n})} \tag{2}$$

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, the category evaluation values or category scores are calculated which indicates the possibility that the candidate document in the testing phase belongs to the category as follows:

$$\mathrm{E_{di}}(\mathrm{C_I}) = \sum\nolimits_{\mathrm{tn}\in \mathrm{di}} \mathrm{FR}(\mathrm{t_n}, \mathrm{t_k}) \tag{3}$$

Finally, the candidate document di is classified into the category $\mathrm{C_{^k}}$ for which the category score is the maximum, as follows:

$$\mathrm{C_{^k}} = \ \mathrm{argmax}\ \mathrm{c_{k\in c}} \mathrm{E_{di}}(\mathrm{c_k}) \tag{4}$$

IV. Experimental Results

The proposed methodology is implemented using Python 3.4.2 [33, 34]. 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.

Four standard evaluations are used: accuracy, recall, precision, and F-measure. The categorization accuracy of the approaches is computed by the equation [35]:

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

Precision, Recall and F-measure are defined as follows [36]:

$$\mathrm{Recall(R)} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{6}$$

$$\mathrm{Precision(P)} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{7}$$

$$\mathrm{F-measure} = \frac{2*\mathrm{P}*\mathrm{R}}{\mathrm{P+R}} \tag{8}$$

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 data sets (i.e. Dataset 1, Dataset 2, and Dataset 3) are collected from the website: www.aljazeera.net [37, 38]. They are used to evaluate the efficiency of our proposed approach.

Dataset1 consists 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.

Dataset3 has 1200 documents which are separated into four categories: international, literature, science and sport.

The datasets are divided into 70% of the documents for training while 30% of the documents are used for testing. These percentages are defined experimentally.

Table1 shows a comparison between the two representations: BOW and Mixed Words.

The results show that the highest accuracy achieved when normalization and stemming are not used is 97.78% with Dataset 2, while it is 98.61% when normalization is used. In the case of using stemmers, the highest accuracy is 92.41% and 95.56% when ISRI and Tashaphyne stemmers are used respectively. While there is a significant decrease in the results when using stemming, but in some categories have achieved 100% accuracy as shown in Tables 2, 3 and 4.

In Table 2, the results show that the highest recall and precision achieved is 100% in art and sport categories when normalization and stemming is used. Also, the results show that the highest recall, precision and F-measure achieved when normalization used is 100% with sport category. The highest recall is 100% in sport category when ISRI stemmer is used. Recall, precision and F-measure achieved is 98.8% in Art and sport when Tashaphyne stemmer is used (see Fig. 2, accuracy results with Dataset 1).

Table 1. Comparison between the two representations BOW and mixed words

	Without normalization or stemming		Normalization		ISRI stemmer		Tashaphyne stemmer	
	BOW	mixed words	BOW	mixed words	BOW	mixed words	BOW	mixed words
Dataset1	96.30%	97.22%	96.67%	97.96%	92.96%	92.41%	95.0%	92.96%
Dataset2	96.89%	97.78%	97.33%	97.78%	88.89%	84.89%	95.33%	87.78%
Dataset3	95.56%	97.22%	97.50%	98.61%	93.06%	86.67%	95.83%	95.56%

Table 2. Results of Recall, Precision and F1 for Dataset1

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	1.000	0.947	0.973	1.000	0.957	0.978	0.878	0.975	0.924	0.878	0.988	0.929
Health	0.989	0.947	0.967	0.989	0.947	0.967	0.833	0.974	0.898	0.922	0.943	0.933
Law	0.933	0.988	0.960	0.967	0.978	0.972	0.967	0.853	0.906	0.978	0.779	0.867
Religion	0.967	0.978	0.972	0.967	1.000	0.983	0.933	0.875	0.903	0.889	0.976	0.930
Sport	1.000	0.989	0.995	1.000	1.000	1.000	1.000	0.947	0.973	0.989	0.978	0.983
Technology	0.944	0.988	0.966	0.956	1.000	0.977	0.933	0.944	0.939	0.922	0.965	0.943
Average	97.22	97.29	97.21	97.97	98.03	97.97	92.40	92.81	92.38	92.96	93.80	93.10

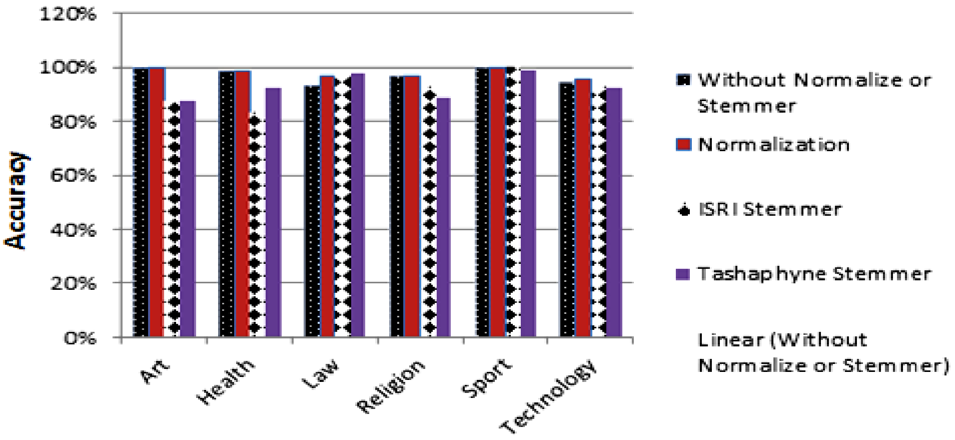

Fig.2. Accuracy for the proposed approach for Dataset1

The results in Table 3 indicate that the highest recall, precision and F-measure achieved when normalization and stemming are not used is 100% in sport category and it is the same percentage when normalization is used with sport category. When stemmers are used, the highest recall is 100% in economic and sport categories when ISRI stemmer is used. Also, when Tashaphyne stemmer is used, the highest recall and precision of 100% in art, economic, politics and sport categories is achieved (see Fig.3, accuracy results with Dataset 2).

Table 4, shows that the highest recall and Precision achieved when normalization and stemming are not used is 100% with Literature and sport categories, and it is the same percentage when normalization is used with literature, science and sport categories. When stemmers are used, the highest Precision is 100% in international and science categories, and the highest recall is 100% in sport category when Tashaphyne stemmer is used (see Fig. 4, accuracy results with Dataset 3).

From the previous results for improving Arabic text classification by using mixed words, best results have been achieved using normalization. Then, in the second place is the results achieved when both normalization and stemming are not used. Finally, the third place is for the results with stemming applied, where Tashaphyne stemmer achieved better results than ISRI stemmer.

Table 3. Results of Recall, Precision and F1 for Dataset2

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Art	0.978	0.978	0.978	0.989	0.978	0.983	0.733	0.970	0.835	0.744	1.000	0.854
Economic	0.967	0.978	0.972	0.967	0.967	0.967	1.000	0.643	0.783	1.000	0.709	0.830
Politics	0.944	0.955	0.950	0.933	0.965	0.949	0.633	0.983	0.770	0.711	1.000	0.831
Science	1.000	0.978	0.989	1.000	0.978	0.989	0.878	0.988	0.929	0.933	0.988	0.960
Sport	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.867	0.92	1.000	0.841	0.914
Average	97.78	97.77	97.77	97.78	97.77	97.76	84.89	88.98	84.91	87.78	90.76	87.76

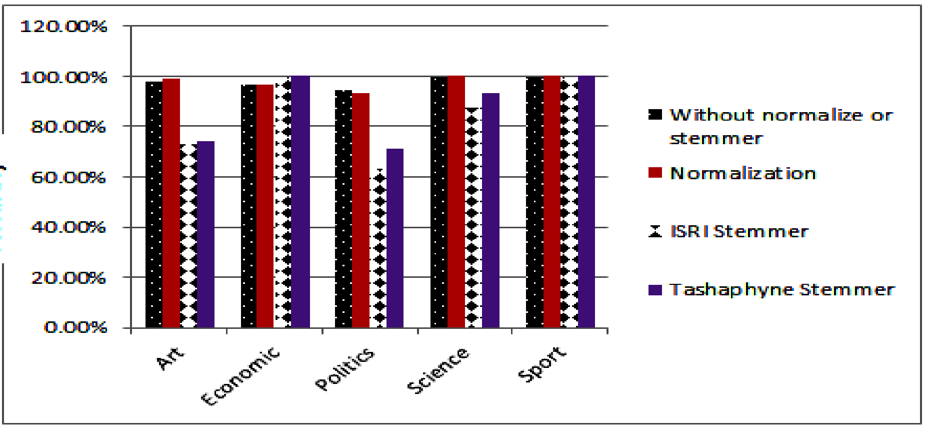

Fig.3. Accuracy for the proposed approach for Dataset2

Table 4. Results of Recall, Precision and F1 for Dataset3

	Without Normalize or Stemmer			Normalization			ISRI Stemmer			Tashaphyne Stemmer		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
International	0.922	0.988	0.954	0.978	0.988	0.983	0.833	1.000	0.910	0.956	0.925	0.940
Literature	1.000	0.968	0.984	1.000	0.957	0.978	0.756	0.944	0.840	0.911	0.965	0.937
Science	0.978	0.946	0.961	0.978	1.000	0.989	0.889	1.000	0.941	0.956	0.956	0.956
Sport	0.989	1.000	0.994	0.989	1.000	0.994	0.989	0.669	0.798	1.000	0.978	0.989
Average	97.22	97.55	97.34	98.61	98.66	98.62	86.67	90.34	87.20	95.56	95.58	95.54

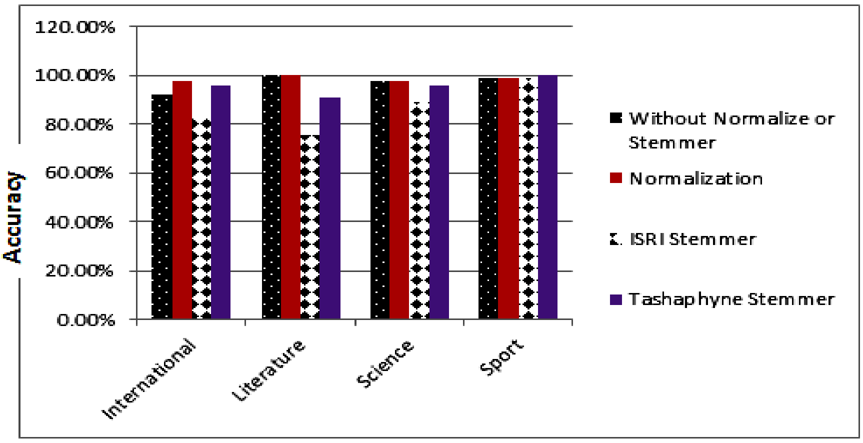

Fig.4. Accuracy for the proposed approach for Dataset3

V. Conclusion

In this paper, the accuracy of categorizing Arabic text has been improved by applying a new technique in the features representation (i.e. mixed words) and a simple efficient technique is used for features selection. Also, the Frequency Ratio Accumulation Method classifier with normalization and two stemming mechanisms: ISRI and Tashaphyne stemmers are used.

The results showed that applying text classification with normalization achieves the highest classification accuracy of 98.61% while it is 97.22% when normalization and stemming are not used. On the other hand, with stemmers less classification accuracy is achieved where the accuracy is 95.56% with Tashaphyne stemmer and it is 92.41%with ISRI stemmer.

In the future work, several approaches that have been applied to English and other languages will be used for improving Arabic text categorization. In addition, new techniques for features representation and selection will be introduced.

References

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

[2] Laila, K.,“ Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study” *Conference*

on Data Mining | DMIN'06 |,2006,pp.78-82

[3] R. Al-Shalabi, G. Kanaan, and M. Gharaibeh "Arabic text categorization using kNN algorithm*",*2006, pp.1-9.

[4] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity "Automatic Arabic text classification*",Journee's internationals d'Analyse statistique des Données Textuelles,* 2008, pp.77-83.

[5] F.Harrag, E.ElQawasmeh "Neural Network for Arabic text classification", pp. 778 – 783,2009.

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

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

[8] Y.Yang and X. Liu" 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,pp. 42-49.*

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

[10] R. Baeza-Yates, and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.

[11] M.Suzuki, S.Hirasawa " 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, 2007,pp. 3535-3540.

[12] H.Meryem, S.Ouatik, A.Lachkar"A Novel Method for Arabic Multi-WordTerm Extraction*"International Journal of Database Management Systems (IJDMS) Vol.6, No.3, June 2014, pp.53-67.*

[13] H.Meryem, A.Lachkar,S.Ouatik"Multi-*Word Term Extraction Based onNew Hybrid Approach For Arabic Language*",2014pp.*109-120.*

[14] B.Al-Shargabi, WAL-Romimah andF.Olayah "A Comparative Study for Arabic Text Classification Algorithms Based on Stop Words Elimination" ACM, Amman, Jordan 978-1-4503-0474-0/04/2011.

[15] W.Zhang, T.Yoshida and X.Tang,"Text classification based on multi-word with support vector machine*"* 2008 Elsevier, pp. 879-886.

[16] A.Wahbeh, M.Al-Kabi, Q.Al-Radaidah, E.AlShawakfa and. I.Alsamdi "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.*

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

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

[19] W.Zhang, T.Yoshida and X.Tang,"Text classification based on multi-word with support vector machine*"* 2008 Elsevier, pp. 879-886.

[20] Sh.Oraby, Y.El-Sonbatyand M.El-Nasr "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.*

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

[22] J.Diederich, J.Kindermann, E.Leopold and G.Paass "Authorship attribution with support vector machines" Applied Intelligence,2003,pp.109-123.

[23] R.Al-Shalabi,G.Kanaan, J.Jaam, A.HasnahandE.Hilat "Stop-word Removal Algorithm for Arabic Language"*Proceedings of 1st International Conference on Information & Communication Technologies: from Theory to Applications,IEEE-France,2004,pp.545-550,CTTA'04* .

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

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

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

[27] A. Kreaa, A. Ahmad and K. Kabalan" ARABIC WORDS STEMMING APPROACH USING ARABIC ORDNET" *International Journal of Data Mining & Knowledge Management Process (IJDKP)* .

[28] https://pypi.python.org/pypi/Tashaphyne/*Vol.4, No.6, November 2014.*

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

[30] H.Meryem, A.Lachkar,S.Ouatik"Multi-*Word Term Extraction Based onNew Hybrid Approach For Arabic Language*",2014pp.*109-120.*

[31] K. El Khatib, A.Badarenh,"Automatic Extraction of Arabic Multi-word Term*"Proceedings of the International Multiconference on Computer Science and Information Technology, 2010,pp.411-418.*

[32] O.Garnes, " Feature Selection for TextCategorization" Master thesis,Norwegian University of Science and Technology, June 2009.

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

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

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

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

[37] http://diab.edublogs.org/dataset-for-arabic-documentclassification/

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

Authors' Profiles

Mahmoud Hussein received his BSc. and MSc. in Computer Science from Menoufia University, Faculty of Computers and Information in 2006 and 2009 respectively and received his PhD in Software Engineering from Swinburne University of Technology, Faculty of Information and Communications Technology in 2013. His research interest includes Software Engineering, Data Mining, Machine Learning, Data Privacy, and Security.

Hamdy M. Mousa received the B.S. and M.Sc. in Electronic Engineering and Automatic control and measurements from Menoufia University, Faculty of Electronic Engineering in 1991 and 2002, respectively and received his PhD in Automatic control and measurements Engineering (Artificial intelligent) from Menoufia University, Faculty of Electronic Engineering in 2007. His research interest includes Intelligent Systems, Natural Language Processing, Privacy, Security, Embedded Systems, GSP applications.

Rouhia M.Sallam is a master student at Menoufia University, Egypt. She works as a teaching assistant in Faculty of Sciences and Computer Sciences, Taiz University, Yemen. Her main research interest is in Natural Language Processing. She received BSc in Computer Science from Taiz University.

How to cite this paper: Mahmoud Hussein, Hamdy M. Mousa, Rouhia M.Sallam,"Arabic Text Categorization Using Mixed Words", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.11, pp.74-81, 2016. DOI: 10.5815/ijitcs.2016.11.09