A Comparative Study for Arabic Text Classification Based on BOW and Mixed Words Representations

Rouhia M.Sallam
Faculty of Applied Sciences, Taiz University, Yemen
Email: rohiya79@yahoo.com
Hamdy M. Mousa and Mahmoud Hussein
Faculty of Computers and Information, Menoufia University, Egypt
Email: {hamdimmm@hotmail.com and mahmoud.hussein@ci.menofia.edu.eg}

***Abstract*- This paper compares two methods for features representation in Arabic text classification. These methods are bag of words (BOW) that mean the word-level unigram and mixed words representations. The mixed words use a mixture of a bag of words and two adjacent words with different proportions. The main objective of this paper is to measure the accuracy of each method and to determine which method is more accurate for Arabic text classification based on the representation modes. Each method uses normalization and stemming. The results show that the use of mixed words in features representation achieves the highest accuracy by 98.61% when normalization is used.**

***Keywords—Arabic Text Categorization, Frequency Ratio Accumulation Method, Term and Document Frequency, Features Selection, bag of words 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 are a tremendous number of Arabic text documents that are available online which are growing every day. As a consequence, 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. The language has a very complex morphology and high inflection. It consists of 28 letters and is written from right to left. In addition, most of the Arabic words have a tri-letter root [2]. However, 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- 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].

In this paper, we compare two methods that use for represented the features in Arabic text classification. These methods are bag of words (BOW) and the mixed words which is a mixture of a bag of words and two adjacent words with different proportions. Also used Term Frequency (TF) technique 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.

This paper is organized as follows. In Section 2, an overview of the related work is presented. Section 3 introduces the proposed comparative process for the two representations. 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 their roots.

Al-Shargabi [10] has compared three techniques for Arabic text classification based on stop words elimination. These techniques are: Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Naïve

Bayesian (NB), and J48 [6]. He have used vector space techniques in features representation, and Arabic data set contains 2363 documents divided into six categories: Sport, Economic, Medicine, Politic, Religion, and Science, where 60% of them is used for training and the rest “40%” are used as testing. The results of accuracy using these techniques achieved 94.8%, 89.42% and 85.07% respectively.

Wahbeh A. et al. [11] introduced an approach for automatic Arabic text classification with and without the stemming application. Using Support vector machine (SVM), Decision Trees (C4.5), Naive Bayes (NB) Classifiers, unique words were extracted in representing the features. They collected a corpus from different trusted websites. The corpus consists of 1100 documents that classified into nine categories: agriculture, art, economics, health and medicine, law, politics, religion, science, and sports. They used accuracy, recall, precision and F-measure to evaluate experiments. The results achieved are 87.79%, 88.54% for the accuracy with SVM and Naïve Bayes respectively when stemming is used. On the other hand, the results achieved when stemming is not applied have lower accuracy (i.e. 84.49% and 86.35%).

Nezreg et al. [12] introduced multiple methodologies for the categorization of English text automatically, these methodologies combine Bag-of-Words, Bag-of-Concepts, and Bag-of-Words with Bag-of-Concepts in representation patterns. Three classifiers were used: SVM, decision trees and KNN and two corpus were used that have 11 categories of reuters-21578 articles and 7 categories of 20 newsgroup. They used precision for evaluating the classification. The results with decision trees give better results than SVM and KNN for their semantic aspect in the classification. The improvements that happened is 22.85% for 20 newsgroups corpus and 4.65% for the Reuters-21578 corpus.

Mesleh et al. [13] proposed an approach for Arabic text classification using SVMs compared to other classification methods (SVMs, Naïve Bayes and k-NN). They collected a corpus from Al-Jazeera, AlNahar, Al-hayat, Al-Ahram, and Al-Dostor. The corpus consists of 1445 documents that classified into nine categories (Computer, Economic, Education, Engineering, Law, Medicine, Politics, Religion and Sport). They used the vector space representation to represent the Arabic documents. CHI square method was used as a feature selection. They used normalization but not stemming because it is not always beneficial for text categorization since many terms may be conflated to the same root [14]. The experimental results show that classification effectiveness is 88.11% using SVMs.

Hadni et al. [15] introduced a new method for Arabic multi word terms (AMWTs) extraction based on a hybrid approach. They used linguistic AMWTs approach to extract the candidate MWTs based on Part Of Speech (POS). A statistical approach is also used to incorporate the 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 [16] for evaluation by two steps (i.e. reference list and validation).

Diab [17] has used multi-word features in Arabic document classification and two similarity functions: 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 used different light stemmers on multi-word features. The dataset was collected from well-known Arabic websites that contain 300 documents. The results show that unordered pairs produce 2% improvement compared to ordered pairs while ordered triples produce bad results.

Zhang et al. [18] used multi-word features representation with support vector machine as a classifier to improve document classification. They proposed a method based on the adaptation of mutual information (MI) and context dependency for compound words extraction from very large Chinese Corpus, and they report that their method is efficient and robust for Chinese compounds extraction. Two strategies were developed based on the different semantic level of the multi-words. The first is the decomposition strategy using general concepts for representation and the second is combination strategy using subtopics of the general concepts.

Suzuki and Hirasawa [19] proposed a new classification technique called the Frequency Ratio Accumulation Method (FRAM). N-gram character and the word N-gram are used as feature terms. The performance for FRAM outperforms the Naive Bayes method (baseline method).The technique is evaluated through a number of experiments using newspaper articles from Japanese CD-Mainichi 2002, and English Reuters-21578. The classification accuracy is the highest when word N-grams is used as feature terms, the results of accuracy are 87.3% for Japanese CD-Mainichi 2002 and 86.1% for English Reuters-21578.

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].

III. EXISTING APPROACH

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. The main steps/stages of our approach for each method are BOW and mixed words for features representation in Arabic text classification, Fig.1 shows the stages that 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.

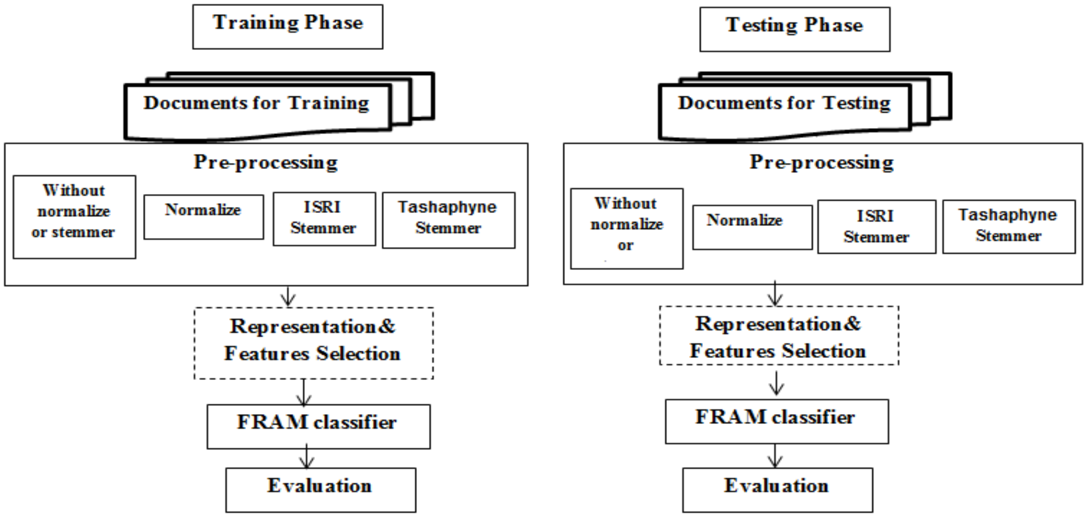

Fig. 1: Arabic Text Categorization

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 [20, 21]. Finally, words with length less than three letters are eliminated; often these words are not important and are not useful in TC. The preprocessing stage includes also normalization and stemming.

A. Normalization

For normalization, a very efficient normalization technique is applied (i.e. Tashaphyne normalization) [22]. 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 "حاســــــــب" to "حاسب" are performed.

B. Stemming

We have 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 [23, 24].

a) ISRI Stemmer

The Information Science Research Institute's (ISRI) Arabic stemmer shares many features with the Khoja stemmer [25]. 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 [26].

b) 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 [27]. Then, segmentation and stemming of the input is performed using a default Arabic affix lookup list for various levels of stemming and rooting [27].

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

B. 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 [28]. In addition, a lot of work has been done to extract MWT in many languages. Many of researchers use AMWTs features to improve Arabic document classification [15, 16, 17, 18].

We are dealing with a huge feature spaces. Therefore, a feature selection mechanism is needed. The most popular feature selection method is term frequency [29].

a) BOW Representation

In the first method, BOW is used in the features representation step. First, the frequencies for every term in all categories are calculated and sorted according to the largest frequency. Second, we take the top 25% of the features when normalization and stemming are not used and take 50% of the features otherwise. These two percentages have been defined experimentally.

b) Mixed words Representation

In the second method, 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, for both the two methods, the frequency ratio (FR) is calculated by the FRAM classifier in each category as follows [9]:

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

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

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

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:

$$\mathrm{E_{di}}(\mathrm{C_I}) = \sum_{\mathrm{tn} \in \mathrm{di}} \mathrm{FR}(\mathrm{t_n}, \mathrm{t_k}) \qquad (3)$$

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

$$\mathrm{C}_{^\wedge\mathrm{k}} = \ \mathrm{argmax}\ \mathrm{c}_{\mathrm{k} \in \mathrm{c}} \mathrm{E_{di}}(\mathrm{c_k}) \qquad (4)$$

IV. EXPERIMENTAL RESULTS

The proposed methodology is implemented using Python 3.4.2 [30, 31]. In addition, our experiments are conducted on a laptop with the following specifications: 2.5 GHz Intel core i5 processor with 4 GB of RAM, and windows 8 enterprise.

A. *Evaluation Metrics*

Four standard evaluations are used: accuracy, recall, precision, and F-measure. The categorization accuracy of the approaches is computed by the equation [32]:

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

Precision, Recall and F-measure are defined as follows [33]:

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

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

$$\text{F} - \text{measure(F1)} = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} \tag{8}$$

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 [34, 35]. 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 are used for training while 30% of the documents are used for testing. These percentages are defined experimentally.

B. *The Results of Experiments using BOW*

In Table 1, the results show that the highest precision achieved for Dataset1 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 [36].

Table 1: 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	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

The results in Table 2 (for Dataset2) 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.

For Dataset3, in Table 3, 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.

Table 2: 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	0.989	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

Table 3: 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	0.900	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

From overall experiments for Arabic text classification by using BOW, the results investigate that normalization can enhance categorization process of documents and gives better evaluation than without normalization and stemming.

C. The Results of Experiments using Mixed Words

In Table 4, for Dataset1, 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.

Table 4: 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

The results in Table 5 indicate that the highest recall, precision and F-measure for Dataset2 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.

For Dataset3, Table 6 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.

Table 5: 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

Table 6: 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

From the previous results for 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.

D. Discussion

Table7and Fig. 2 show a comparison between the two representations: BOW and mixed words. The results show that the highest accuracy achieved by used mixed words when normalization is 98.61% with Dataset2, while in BOW method by used normalization achieved 97.50% in the same dataset. The mixed words methods showed the highest accuracy in all datasets and all experiments, except for the use of stemming the results are a significant decrease but in some categories have achieved 100% accuracy as shown in Tables 4, 5 and 6.

.

Table 7: 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%

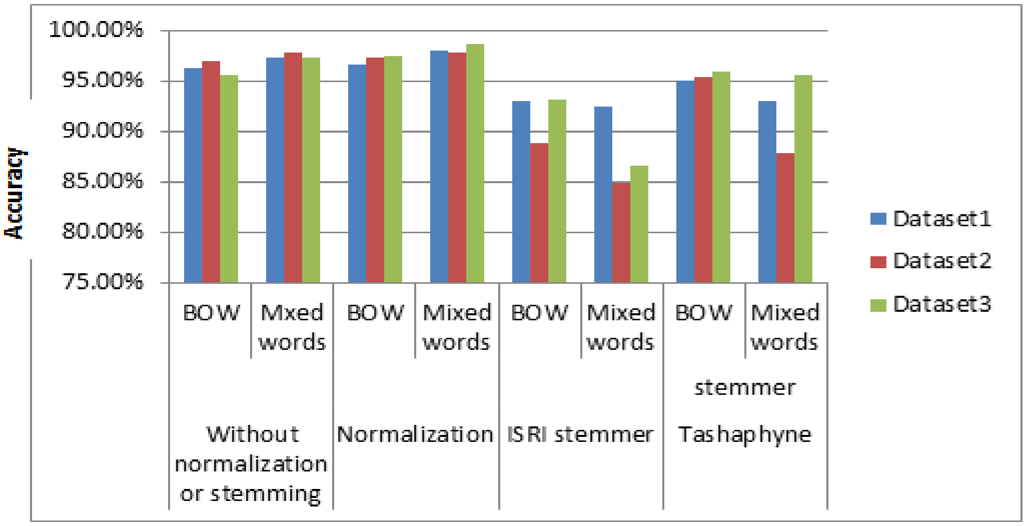

Fig. 2: Comparison between the two representations BOW and mixed words

Table 8 shows the results for execution time for all stages of classification with different datasets and the four experiments for two methods. The first method by using BOW took less execution time in all experiments and all datasets. It was less time for execution with the ISRI stemmer where it took is 46 seconds in Dataset3. While the second method by using mixed words took 78 seconds with Dataset3 (See Fig.3 that shows the comparison between BOW and mixed words in the execution time).

Table 8: Results the execution time of for the two methods

	Without normalization or stemming		Normalization		ISRI stemmer		Tashaphyne stemmer	
	BOW	Mixed words	BOW	Mixed words	BOW	Mixed words	BOW	Mixed words
Dataset1	133s	500s	125s	488s	71s	147s	80s	158s
Dataset2	88s	341s	85s	266s	56s	120s	64s	127s
Dataset3	77s	205s	74s	198s	46s	78s	51s	86s

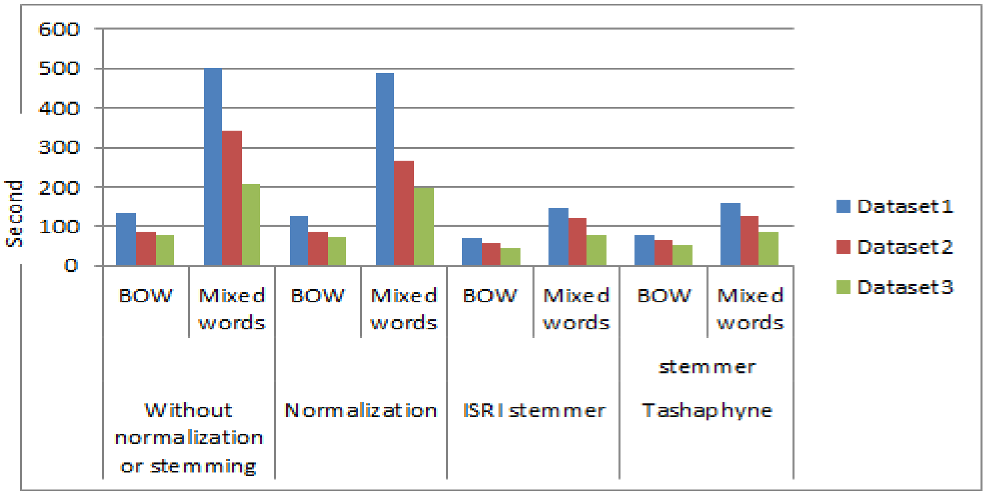

Fig. 3 Comparison between the two methods in the execution time

By analyzing the previous results for the two methods for features representation in Arabic text classification (i.e. BOW and mixed words), we observed the following. On the one hand, the results in mixed words method showed the highest accuracy in all datasets and all experiments. But, it takes more execution time. On the other hand, the use of BOW achieves less accuracy and takes less execution time in all experiments.

V. CONCIUSION

In this paper, we have compared two methods in features representation for categorizing Arabic text. The first method applies BOW while the second method applies technique in the features representation (i.e. mixed words). Each method uses a simple efficient technique for features selection. Also, we have applied Frequency Ratio Accumulation Method classifier with normalization and two stemming mechanisms: ISRI and Tashaphyne stemmers are used.

The results show that the use of mixed words achieves the highest classification accuracy of 98.61% with normalization, while the use of BOW achieves 97.22% with normalization. In addition, the use of BOW method has less execution time in all experiments.

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 |, ,pp.78-82, 2006.

[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, pp.77-83,2008.

[5] F.Harrag, E.ElQawasmeh "Neural Network for Arabic text classification", Applications of Digital Information and Web Technologies, 2009. ICADIWT '09. Second International Conference, pp. 778 – 783, 2009.

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

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

[11] 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, pp.54-70, 2011.

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

[13] A. Mesleh. "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System", Journal of Computer Science, pp. 430-435. 2007.

[14] T. Hofmann, "Introduction to Machine Learning", Draft Version 1.1.5, November10, 2003.

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

[16] H. Meryem, A. Lachkar, S. Ouatik, " Multi-Word Term Extraction Based On New Hybrid Approach for Arabic Language" ,Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014 ,pp. 109-120,2014.

[17] D. Abuaiadah, "Arabic Document Classification Using Multiword Features",International Journal of Computer and Communication Engineering, Vol. 2, No. 6, pp.659-664. November 2013.

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

[19] 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.

[20] 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 .

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

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

[23] Sh. Oraby, Y. El-Sonbaty and 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.

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

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

[26] A. Kreaa, A. Ahmad and K. Kabalan "Arabic Words Stemming Approach Using Arabic WordNet" International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.4, No.6, November 2014.

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

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

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

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

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

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

[33] 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.

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

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

[36] R. Sallam, H. Mousa, M. Hussein "Improving Arabic Text Categorization using Normalization and Stemming Techniques,” International Journal of Computer Applications (0975 – 8887) Volume 135 – No.2, February 2016.