Text Classification using Bi-Gram Alphabet Document Vector Representation

Fatma Elghannam
Electronics Research Institute, Cairo, Egypt.

Abstract

Text classification TC is the process of assignment of text documents to appropriate categories based on their content. High dimensionality of feature space is a primary challenge in TC. The most common approach for TC is bag of words BOW which is limited due to the continuous increase in the number of features as the volume of vocabulary increases. Many investigators have addressed the issue of management of dimensionality by applying careful preprocessing techniques that include complex morphological phase, in particular for the high inflectional languages including Arabic. In the present study, term frequency of bi-gram alphabet is used to construct document vector. A main contribution of bi-gram alphabet approach is that feature terms are standard and separate from documents contents; this helps to reduce the high dimensionality associated with the increasing the volume of data. In addition, the classification process performs well on both Arabic and English collections without morphological preprocessing requirements. The proposed approach has proved high accuracy results and outperformed other Arabic TC systems.

Keywords: *Text classification, Arabic document, bi-gram alphabet, feature selection, support vector machine.*

I. INTRODUCTION

With the growing amount of electronic documents available, there is a need to classify documents automatically. Text classification TC can be defined as the task of assigning natural language texts into one or more predefined categories based on their content [1]. There are many natural language applications that adopt TC such as email filtering, opinion mining, information retrieval, automatic indexing, summarization, and others. TC process involves common steps. This includes text preprocessing, data division, feature extraction, feature representation, applying a classification algorithm, and performance evaluation [2]. High dimensionality of feature space is a primary challenge in TC. The most common representation for text classification is the bag of words BOW vector. In this method, words that exist in the document collection are used to identify the features. This result in a very large number of features, and therefore the challenge becomes the high dimension of the input feature space, and sparsity of document vectors. So, various researches have presented a set of auxiliary techniques being undertaken in order to overcome the high dimensionality and improve the classification accuracy. This includes, preprocessing, feature selection, filtering and others. In this context, for Semitic languages that have morphological nature like Arabic, some researchers considered root extraction and word stemming as a part of pre-processing [3], [4].

A brief description of the common TC tasks as follows:

A. Pre-processing:

Several tasks can take place in the preprocessing step. The common tasks in preprocessing for TC include the following:

- Removal of functional words: this may include removal of numbers, punctuation, particles, propositions, and others.
- Stemming: reduces all variants of a word into their stem form (e.g., {see, saw, seen} -> "see").
- Normalization: Arabic texts need more consideration due to the writing styles. Some writing forms usually are normalized such as replacing "ة" with "ه".

However, in the proposed approach, no preprocessing phase was applied, except normalization for Arabic characters.

B. Feature Extraction:

It characterized by a definition of the term, and the method to compute term weights. Concerning term definition, vast majority of text classification researches concentrate on the words which occur in the document to identify terms [1], [5]. In order to reduce the number of terms, researches have been considered word stem or root [6], [7]. To cope with the morphological processing, [8], [9] extracted the character tri-gram units that occur in documents to identify terms. For the second issue, determining the weight of term in

document; the most common weighting methods include Term Frequency (TF) [4], [10] which evaluates how many times a term occurs in a document, *Term Frequency* Inverse Document Frequency TF*IDF* [6], [10], [11], and Normalized Frequency [9], [12].

C. Feature Selection:

It also a common technique in order to reduce irrelevant or noisy terms with the aim of improving the learning performance and saving the computation requirements [13]. Various dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, and so on. Details of these selection functions were stated in [1].

D. Classification Algorithm:

There are a number of successful classifiers that have been used in Arabic text classification, such as k-nearest neighbour (KNN) [2], [12]; Naive Bayes [2], [14] and support vector machine (SVM) [2], [6], [12], [14]. SVM has proven that it is very well suiting for text classification for its high accuracy, and an inherent ability to handle large feature spaces [2], [15].

In this work, we introduce and evaluate a novel approach to classify documents using bi-gram alphabet document vector representation and linear support vector classifier. A main contribution of this work to TC area is that we have introduced an approach to classify documents using standard features instead of relying on documents contents to extract features. This guarantee reducing the dramatically increase in features space with increasing volume of data. The current work was applied to classify Arabic and English text collections. The proposed approach is characterized by:

- The approach is language-independent.
- The adopted Features are standard and separate from contents of the documents.
- Term frequency of bi-gram alphabet is used in constructing document vector.
- Documents can be accurately classified without carful preprocessing requirements, especially at the morphological level.
- Number of features does not increase with increasing volume of data.

The rest of the paper is organized as follows: Section II introduces existing related work. Section III presents the proposed approach to classify documents. Experimental results and analysis are reported in Section IV, and the last section concludes this paper.

II. RELATED WORK

Text classification has been extensively studied for more than a decade. The following sections present an overview of a number of studies in the field of Arabic TC.

An important task that has a direct effect on the classification accuracy is the set of features that represent the documents. Most of text classification researches concentrate on the words to express documents, which is called Bag-of-Words (BOW) approach. In this approach, it is possible for a document to include a feature for each word within that document. Therefore, the challenge becomes the high dimension of the input feature space, where tens or hundreds of thousands of unique words in its feature space for some document collections. However, the challenges are greater when dealing with Arabic due to multiple issues, the most important of which is its inflectional and derivational nature. Various researches have considered the root or stem form to reduce the feature space and improve the accuracy. The work of [4], [10] proposed a classification approach for Arabic documents using stems of Arabic words, in which the suffix and prefix were removed from the orthographic form of the word. The work of [16] proposed an Arabic TC and considered the root form in their feature representation. The work of [17] studied the impact of Stemming by applying the Khoja stemmer [18], Information Science Research Institute (ISRI) stemmer [19], and Tashaphyne Light Arabic Stemmer [20] on two datasets of the opinion classification problem. Their results show that the Khoja stemmer is the best for stemming process in TC.

Another approach used a summarizer is presented in [21]. He applied a summarizer to extract informative sentences first, and selected the best words from the summary documents instead of using all words. Words term frequencies (TF), and SVM classifier is used to classify Arabic text documents. His results showed an improvement using the words extracted from summary compared to using all words in feature representation.

The work of [22] presented KNN classifier for Arabic documents. They used N-Gram at the word level (unigrams and bigrams), and compared the classification results by using traditional single words as features. Their results showed that using N-Grams produces better accuracy (f-measure= 0.736) than using single words (f-measure= 0.669) in feature representation for the classification.

The work of [23] proposed an approach for the Arabic TC using Arabic WordNet AWN thesaurus as a lexical and semantic source. In this approach, they

proposed a weighting scheme based on the frequency of the relation in the AWN and the corpus documents. They compared the results of their approach against the BOW as features in the supervised learning of an NB classifier. Their results showed that their suggested approach outperformed the BOW approach.

Another approach proposed by [24] used Polynomial neural Networks (PNs) algorithm for Arabic TC. They applied stemming and used Chi Square for feature selection. Only 1% of each class features is selected to build the classifier. Their results revealed that stemming is a necessary step with PNs classifier, since PNs are usually used with a small number of features due to their high memory requirements. Their results was (F-measure =0.893) with only 135 features using Alj-News Arabic corpus.

BPSO (Binary Particle Swarm Optimization)-KNN as a feature selection method for Arabic text classification is proposed by [14]. They applied preprocessing steps that include removing stop words and rare words (words that occur less than five times in the dataset). No stemming process was applied. They experiment three different classifiers, Support Vector Machines (SVM), Naive Bayes (NB) and Decision Trees (J48). Three separate Arabic datasets have been used to test their method. Their best results on Alj-News Arabic corpus was (F-measure = 0.931) by SVM and 2967 features selected to build the classifier.

While the above-mentioned methods focused on words to represent documents, other approaches used alternative approach where they dealt with characters to cope with the morphological processing. Other work proposed an Arabic text classification using tri-gram frequency [8]. In her work, preprocessing phase was applied first to remove unimportant words. The character tri-gram frequency profile was generated for the training documents. Then for each document to be classified, distance measure between the N-gram (N=3) frequency profile for that document and all the training classes are calculated. She employed the two distance measures, Manhattan and Dice. Her results showed that tri-gram text classification using the Dice measure outperforms classification using the Manhattan measure.

Other work proposed a character based classifier to cope with the sparse data problem instead of feature reduction by the morphological rules [9]. They used continuous sequences of two types of units extracted from the documents. The units can be either full-form words or character tri-grams. They carried out experiments with maximum entropy text classification on a large Arabic corpus and used no preprocessing steps. The best classification accuracy they reported was 62.7% with precision of 50%.

III. TEXT CLASSIFICATION USING BI-GRAM ALPHABET APPROACH

In the current bi-gram alphabet approach, the standard alphabet is used as basics for representing document vector. There is no need any prior knowledge of topics, words, or lexical structure to represent textual documents. The classification process starts by simple normalization in case of Arabic documents, while, no preprocessing was applied for the English documents. A document vector was calculated based on occurrences of bi-gram of the standard alphabet characters in that document. Finally SVM algorithm with k-fold cross-validation technique was used to classify the documents. The detailed steps of the current approach are presented in the following sections.

A. Text Preprocessing

Standard Arabic alphabet contains 28 characters. There is no distinct upper and lower case character forms. Arabic script is written from left to right. There exist three basic short vowels that can be used in generating different pronunciation of a character. In modern standard Arabic publications, short vowels are typically not written and left to the experience of the reader. Some characters have different shapes due to several sounds it represents. And this produces many different forms and shapes to recognize each sound. For example, Alef, the first character in Arabic alphabet can appear in four various shapes to express different sounds. Many modern standard Arabic publications normalize different alef shapes. In the present work, preprocessing step included only normalization of Arabic character shapes. Different aleph shapes "ا, أ, آ, إ" were normalized to "ا" , and Ha shapes " ه ، ة" to "ه".

It should be noted that, in the present work, we avoided applying many preprocessing tasks, such as removal of functional words and morphological process such as stemming. And this was in order to test the efficiency of the current approach in the case of dropping these tasks especially the morphological level, which is difficult task that needs special care for high inflectional languages including Arabic. For the English documents no preprocessing step was applied at all.

B. Feature Vector Calculation

The current document classification approach adopted the standard alphabet to construct document feature vector. The technique starts basically by generating all possible arrangements of two characters in the alphabet. An arrangement of a set of objects is called a permutation. The permutation *P* of a

number of ordered arrangements -with no repetitions- of *r* objects taken from *n* unlike objects is defined by:

$$^{n}P_{r} = \frac{n!}{(n-r)!} .$$

Where:

- *r* is the number of characters chosen to construct a term, which is 2 for bi-gram alphabet term.
- *n* is the number of distinct standard alphabet characters, which is 28 for Arabic.

Thus, total number of bi-gram characters arrangements is 756, which represents the number of adopted terms/features in case of Arabic documents. We generated all arrangements of two Arabic alphabet characters and excluded the repetitive ones, where preliminary experiments showed that their elimination does not affect the classification accuracy. In addition to, repeated characters cannot be detected directly in case of the same character occurs twice in a word, with no vowel between, as they are replaced by one character with a short vowel "shadda" (or gemination).

On the other hand, English alphabet includes 26 characters. In case of English, we did not conduct experiments to find out the importance of existence of repetitive characters. Therefore, repetitive English characters were taken into consideration. In this case, the permutation of a number of ordered arrangements with repetitions of *r* objects taken from *n* unlike objects is: n^r. As demonstrated previously: *r*=2, and *n*=26 for English. Thus, total number of adopted English bi-gram characters arrangements is 676 bi-gram English terms.

After identifying the feature terms, then for each document, document feature vector was calculated based on the normalized term frequency *NTF* of the bi-gram terms. Where *NTF* of a term is the number of times a term occurs in a specific document normalized by the total number of terms exist in that document. *NTF* values ranges from *0* to *1*.

C. Feature Selection

Not all combinations of two characters are common in language; certainly there are some rare compositions. Others have low influence on the classification process. Chi-square measures the maximal strength of dependence between features and different categories. It has been proved to record high accuracy in classifying both Arabic and English texts [24]. In the current work, chi-square measure is used in order to reduce irrelevant or noisy terms with the aim of improving the classifier performance and saving the computation requirements. Top *p* percent features were selected to build the classifier. Results of different *p* values were tested during the experiments.

D. Machine Learning Process

After constructing the document vectors, the phase of choosing the appropriate classifier can be applied. SVM algorithm has been proved to perform successfully in text classification [15], [25], [26]. Therefore, it was chosen in the present work to classify documents. SMO; which is the WEKA [27] version of SVM algorithm was used to build the model. In our experiments, the dataset was tested using k-fold cross-validation technique with k=10. In this technique, the original dataset is randomly partitioned into k subsamples. A single subsample is retained as the validation data for testing the model, and the remaining k−1 subsamples are used as training data. The process is repeated k times, with each of the k subsamples used exactly once as the test data. The k results then can be averaged to produce the final estimation.

Note that, we do not need to do feature extraction or term weighting separately for the training set and the testing set. In the current work, the feature terms are standard and do not extracted from the documents contents. At the same time, term weighting calculation for a document, is not influenced by other documents. Other works [14] that have used BOW and TF.IDF need to do feature preprocessing and weightings separately for the training set and the test set.

IV. EXPERIMENTAL STUDY

We have performed experiments on four different Arabic datasets. To test the applicability of using the present approach in other languages, English dataset was also experimented. In the experiments, precision, recall and F-measure were used as a performance metrics for text classification. In all experiments conducted, the classifier was trained with the 10-fold cross validation technique. Four different experiments were conducted. Both Weka [28], and Rapid Miner [29] tools were used during the experiments. The datasets used, results of experiments conducted, performance evaluation, details of feature reduction, and analysis of results are presented and discussed in the following sections.

A. The Datasets

1) Arabic Datasets

In our experiments, four different Arabic datasets were used; the following sections describe the details of each dataset:

Alkhaleej News Dataset: Alkhaleej-2004 Dataset contains 5690 documents for news articles which correspond to nearly 3 million words. It is available from [30]. Each document is labeled with one of the following four classes {'International News', 'Local News', 'Sport', and 'Economy'}. We choose a random

set of 2000 documents equally distributed among the four classes.

Aljazeera News 9 classes Dataset (Alj-News9): Alj-News9 Arabic Dataset contains 2700 documents for news articles. It is available from [31]. Each document is labeled with one of the following nine classes {'Economy', 'Health', 'Law', 'Literature', 'Politics', 'Religion', ' Sport', 'Technology', and 'Art'} with equal number of documents in each category.

Aljazeera News 5 classes Dataset (Alj-News5): Alj-News5 is another different dataset that contains 1500 documents for news articles. It is available from [31]. Each document is labeled with one of the following five classes {'Art', 'Economic', 'Politics', 'Science', and 'Sport'}. Alj-News5 was used by other two research works [14], and [24]. A comparison of the classification results of the current approach against these two works on Alj-News5 will be presented in the experiments.

BBC-Arabic News Dataset: The dataset contains 4763 documents of BBC-Arabic news collected from the BBC Arabic website. It is available from [32]. Each document is labeled with one of the following seven classes {'Middle East', 'World News', 'business', 'sport', 'newspapers', 'Science', and 'Misc.'}.

*2) **English Dataset***

BBC-English News Dataset is an English dataset that contains 2,225 articles from the BBC news website available from [33]. It is corresponding to stories in five topical areas from years 2004-2005. Each article is labeled with one of the following five classes: {'business', 'entertainment', 'politics', 'sport', and 'tech'}.

B. Bi-Gram Alphabet Distribution in Arabic Vocabularies

The experiment was conducted to study the distribution of bi-gram alphabet in Arabic vocabularies. In this experiment Chi-square was applied to measure the importance of each feature. The experiment was applied to classify documents using the four Arabic datasets separately. In studying the results of Chi-square term weights, it was found that there are typical 125 terms that have zero weights agreed upon all the four Arabic datasets. It was observed that these bi-gram terms almost do not occur in the datasets vocabularies. We therefore considered these terms as rare and difficult to occur in the Arabic vocabularies. Accordingly, the list of rare bi-gram terms can be safely dropped from the feature list as it will be shown in next experiment. Example of zero weight rare bi-grams in Arabic vocabularies that were revealed by this experiment are: (حخ, دذ, سش, صض,شص, ضظ, ذش).

C. Validation of Bi-Gram Alphabet Approach

In this experiment we separately classified different datasets using SVM-SMO classifier and bi-gram alphabet vector representation. The classifier was trained with 10-fold cross validation technique. The experiment was applied on Arabic documents with 631 features (full number of features excluding the zero weight rare ones described in section *B* above). Table I shows the overall summarized classification accuracy in terms of precision, recall, and F-measure on different Arabic datasets: Alkhaleej News, Alj-News9, Alj-News5, and BBC-Arabic news. The English dataset (BBC-English News) was also experimented by the same way. The table shows that, the best recorded F-Measure was (0.949) on Alj-News5, while the lowest F-Measure was (0.874) on BBC-Arabic. The average F-measure over all datasets was (0.92).

The achieved accuracy results reveals that the presented approach has recorded high performance and can be successfully applied to classify Arabic and English documents. More detailed study and comparison against other works will be presented in the following experiments.

Table I: Classification Accuracy of Svm-Smo by Bi-Gram Alphabet on Different Datasets

Dataset	Precision	Recall	F-Measure
Alkhaleej	0.905	0.905	0.905
Alj-News9	0.930	0.930	0.930
Alj-News5	0.949	0.949	0.949
BBC-Arabic	0.874	0.874	0.874
BBC- English	0.926	0.926	0.926

D. Effect of Feature Space Reduction on the Classification Accuracy

The experiments were conducted with different feature space to study the effect of the selected number of features on the achieved accuracy. Initially, features were weighted by Chi-square and ordered by their weights in ascending order. Then the feature dimension was defined by 100%, 90% and so on up to 10% of full features (where full features =756 in Arabic, and 677 in English). The classifier was trained with 10-fold cross validation technique on different datasets separately. The classification accuracy results in terms of F-measure were formulated in the two graphs shown in figures 1, 2. The figures illustrate the effect of selected percentage number of features with Chi-square feature selector on the achieved accuracy over different data sets. Figure 1 shows the accuracy results among different percentages of highest ranking features over different datasets. Figure 2 shows the accuracy deviations from the baseline over different data sets with different percentages of the highest ranking features, where the baseline is the accuracy at

full feature. It was noticed that even the selected features of up to 50% of full features, there was no significant change in the accuracy results over all datasets. This indicates that even if half number of the bi-gram alphabet features are used, the classifier can predict with high accuracy close to the baseline (full features) accuracy. Significantly, from the obtained recorded results it was found that at 50% of full features, the average decrease over all datasets in F-measure was (0.003) with maximum (0.005) and minimum (0.001). More details in this regard and comparison against other methods will be presented in the next experiment.

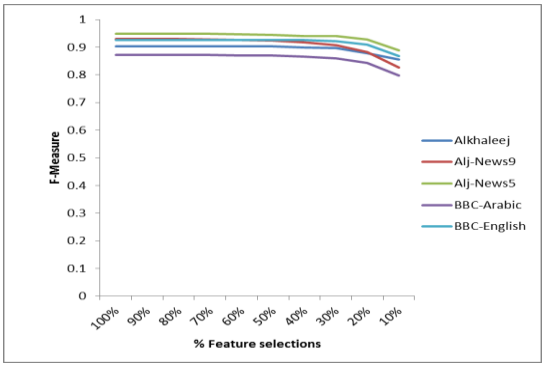

Figure 1 % Features Selection and the Corresponding Classification Accuracy

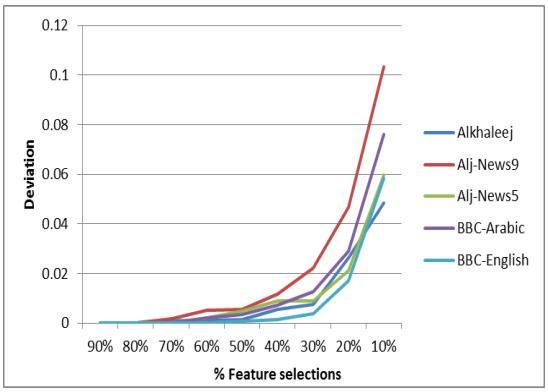

Figure 2 Deviation in Accuracy from the Baseline using Different Percentage Feature Selections

E. Analysis and Comparison against other Methods

The detailed classification results of Alj-News5 were highlighted in order to analyze the results more closely and compare the accuracy against other methods. There are difficulties to make direct comparison with other previous research works in the area of Arabic TC. Among these difficulties are the following: no benchmark Arabic dataset, the dataset used is not available in some cases, researchers selects randomly different amount of documents of the original dataset, different number of instances for each class, different organization of training and testing sets, and lack of clarity on the number of features selected. Alj-News5 was selected for comparison due to the availability of its test results for other works. The detailed results of applying SVM-SMO classifier using the current bi-gram alphabet approach on Alj-News5 is presented in table II, showing precision, recall, and F-measure for each class. The classifier was trained with the 10-fold cross validation technique as described in section *C*. As shown in table II, the top F-measure was on 'Sport' class with 0.995, while the lowest F-measure recorded was 0.901 on 'Politics' class. The overall weighted average F-measure for all classes was 0.949.

Table II - Accuracy By Class for Svm-Smo on Alj-News5

Class	Precision	Recall	F-Measure
Art	0.937	0.937	0.937
Economic	0.946	0.930	0.938
Politics	0.890	0.913	0.901
Science	0.980	0.967	0.973
Sport	0.993	0.997	0.995
Weighted Avg.	0.949	0.949	0.949

The recorder results of experiment *D* for Alj-News5 is presented in table III.

Table III - Different Feature Selections and Corresponding F-Measure On Alj-News5

%Feature Selection	Number of Features	F-Measure
Full	756	0.949
90%	680	0.949
80%	605	0.949
70%	529	0.949
60%	454	0.944
50%	378	0.946
40%	302	0.940
30%	226	0.930
20%	151	0.914
10%	75	0.890

The table shows the detailed view of accuracy results among different percentage of highest ranking features with Chi-square feature selector. The columns of the table indicate the percentage features selection, the corresponding number of features selected, and the recorded F-measure. The table shows that, the best recorded F-Measure at full features was 0.949, and it has been maintained at the same accuracy using up to 70% of full features (i.e. 529 features). While when selecting 50% of full features, the obtained F-measure was 0.946, significantly F-measure was reduced by only 0.003 from the baseline measure. The lowest F-Measure = 0.89 was recorded at 10% of the full features (75 features).

In order to measure the success of the current bi-gram alphabet approach, the results were compared against other two works; System1, proposed by [14] and System2, proposed by [24], for the same dataset Alj-News5. Due to the important impact of feature space in the classification process, it was taken into consideration in the comparison. Table IV presents the best overall F-measure and the number of features selected for the three systems on Alj-News5 dataset.

Despite the difficulty of an accurate comparison, the analysis of the results showed the following:

System1 used BPSO-KNN as a feature selection method and applied preprocessing steps that includes removing stop words and rare words, no stemming was applied. Their best results was (F-measure = 0.931) obtained by SVM classifier. Nevertheless, this accuracy value was at the expense of high memory requirements due to the large number of features (2967) that was used to build the classifier. The results obtained by the current work outperformed system1 in terms of f-measure (0.949), and the number of features used (529). In addition to, in the current approach, to achieve approximately the same accuracy (0.930) that was obtained by System1, only 266 features could be used, as shown in table III.

System2 used PN classifier. They applied several preprocessing steps including stop words removal and stemming. To further reduce the number of features, Chi-square was used for feature selection, and only 1% of each class features is selected to build the classifier. Their results by PN classifier using 135 features was (F-measure =0.893). The current approach has the superiority in f-measure (0.949), with no morphological preprocessing. Although System 2 had used powerful algorithm PNs classifier, it has memory restriction. So, there was a need to reduce the number of features as have been implemented in their work. In this regard, another experiment was applied to test the current bi-gram alphabet approach using the same number of features that was used in system2. The results showed that bi-gram alphabet achieved 0.913 at 135 selected features, compared to 0.893 that was obtained by system2 at the same number of features.

Table IV- Overall Best F-Measure and Number of Features Selected for the Three Systems on Alj-News5 Dataset

Work	F-Measure	Number of Features
System1	0.931	2967
System2	0.893	135
Bi-gram Alphabet	0.949	529

In addition to that the proposed approach had proved high accuracy results; there are two basic remarkable contributions in the TC area. First, the results were obtained without the need of careful preprocessing tasks at the morphological level. So, we have escaped a difficult task for several languages, in particular for the high inflectional languages including Arabic. Second, features that were used in the classification process are standard and separate from contents of the documents, so it does not increase with increasing volume of data; this is an important issue to keep the memory requirements at minimum.

V. CONCLUSION

In this paper a novel approach bi-gram alphabet for text classification was presented. Term frequency of bi-gram alphabet was used as the weighting scheme to represent document content. The proposed approach have proved high accuracy results and outperformed other well-known Arabic TC systems to classify Arabic and English text documents using SVM-SMO classifier. The current approach has two main contributions. First, the adopted features are standard and separate from contents of the documents; this helps to reduce the high dimensionality with increasing the volume of data. Second, the classification process performs well without many preprocessing tasks especially at the morphological level, which is difficult task for several languages including Arabic.

REFERENCES

[1] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), 34(1), 1-47, 2002.

[2] S. Khorsheed, O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," Language resources and evaluation, 47(2), 513-538, 2013.

[3] T. Kanan, A. Fox, "Automated Arabic text classification with PStemmer," machine learning, and a tailored news article taxonomy. Journal of the Association for Information Science and Technology, 67(11), 2667-2683, 2016.

[4] M. M. Syiam, Z. T. Fayed, M. B. Habib, "An intelligent system for Arabic text categorization," International Journal of Intelligent Computing and Information Sciences, 6(1), 1–19, 2006.

[5] J. Diederich, J. L. Kindermann, E. Leopold, G. PAAß, Authorship attribution with support vector machines. Applied Intelligence, 19(1/2), 109–123, 2003.

[6] A. Mesleh, "Chi square feature extraction based Svms Arabic language text categorization system," Journal of Computer Science, 3(6), 430–435, 2007.

[7] F. Thabtah, M. Eljinini, M. Zamzeer, W. Hadi, "Naı¨ve Bayesian based on Chi Square to categorize Arabic data," In Proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on

Innovation and knowledge Management in Twin Track Economies, 2009, pp. 930–935.

[8] L. Khreisat, "Arabic text classification using N-gram frequency statistics a comparative study," In Proceedings of the 2006 International Conference on Data Mining, 2006, pp. 78–82.

[9] H. Sawaf, J. Zaplo, H. Ney, "Statistical classification methods for Arabic news articles," Arabic Natural Language Processing Workshop, ACL'2001, 2001, pp. 127–132.

[10] M. M. Zahran, G. Kanaan, M. B. Habib, "Text feature selection using particle Swarm optimization algorithm," World Applied Sciences Journal, 7 (Special Issue of Computer , IT), 69–74, 2009.

[11] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval", Information processing management, 24(5), 513-523, 1988.

[12] A. El-Halees,"A comparative study on Arabic text classification," Egyptian Computer Science Journal, 30(2), 2008.

[13] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection. Journal of machine learning research," 3(Mar), 1157-1182, 2003.

[14] H. K. Chantar, D. W. Corne, "Feature subset selection for Arabic document categorization using BPSO-KNN", In Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on (pp. 546-551). IEEE

[15] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," In European conference on machine learning, 1998, pp. 137-142, Springer, Berlin, Heidelberg.

[16] S. Bahassine, A. Madani, M. Kissi, "Arabic Text Classification Using New Stemmer for Feature Selection and Decision Trees," Journal of Engineering Science and Technology, 12(6), 1475-1487, 2017.

[17] S. Oraby, Y. El-Sonbaty, M. A. El-Nasr, "Exploring the effects of word roots for Arabic sentiment analysis," In Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013, (pp. 471-479).

[18] S. Khoja, "APT: Arabic part-of-speech tagger," In Proceedings of the Student Workshop at NAACL, 2001, pp. 20-25.

[19] K. Taghva, R. Elkhoury, J. Coombs, "Arabic stemming without a root dictionary," In Information Technology: Coding and Computing, 2005, ITCC 2005. International Conference on (Vol. 1, pp. 152-157). IEEE.

[20] Tashaphyne (2010) Arabic light stemmer, [Online]. Available: http://tashaphyne.sourceforge.net/.

[21] E. Al-Thwaib, "Text summarization as feature selection for Arabic text classification," World of Computer Science and Information Technology Journal (WCSIT), 4(7), 101-104, 2014.

[22] R. Al-Shalabi, R. Obeidat, "Improving KNN Arabic text classification with n-grams based document indexing," Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, 108-112, 2008.

[23] S. A.Yousif, V. W. Samawi, I. Elkabani, "Arabic Text Classification: The Effect of the AWN Relations Weighting Scheme," In Proceedings of the World Congress on Engineering Vol. 2 , 2017.

[24] M. M. Al-Tahrawi, S. N. Al-Khatib, "Arabic text classification using Polynomial Networks,"Journal of King Saud University-Computer and Information Sciences, 27(4), 437-449, 2015.

[25] N.Anitha, B. Anitha, S. Pradeepa, "Sentiment Classification Approaches," International Journal of Innovation Engineering and Technology, 3(1), pp. 22-31, 2013.

[26] N. Cristianini, J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods," Cambridge university press, 2000.

[27] I. H. Witten, E. Frank, M.A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[28] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[29] Rapid Miner Project RM (2013). The Rapid Miner Project for Machine Learning. Available: http://rapid-i.com. Last access on December 2017

[30] Arabic Corpora - Mourad Abbas. (2004). Available: https://sites.google.com/site/mouradabbas9/corpora. Last access on January 2018.

[31] Arabic Corpora - Alj-News.(2004). Available: https://filebox.vt.edu/users/dsaid/Alj-News.tar.gz. Last access on January 2013.

[32] Saad, M. K., Ashour, W. Osac: Open source Arabic corpora. In 6th ArchEng Int. Symposiums, EEECS (Vol. 10) , 2010.

[33] Open-source BBC Dataset, available at: http://mlg.ucd.ie/datasets/bbc.html.