Feature Selection Method Based On Statistics of Compound Words for Arabic Text Classification

Aisha Adel, Nazlia Omar, Mohammed Albared, and Adel Al-Shabi
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

Abstract: *One of the main problems of text classification is the high dimensionality of the feature space. Feature selection methods are normally used to reduce the dimensionality of datasets to improve the performance of the classification, or to reduce the processing time, or both. To improve the performance of text classification, a feature selection algorithm is presented, based on terminology extracted from the statistics of compound words, to reduce the high dimensionality of the feature space. The proposed method is evaluated as a standalone method and in combination with other feature selection methods (two-stage method). The performance of the proposed algorithm is compared to the performance of six well-known feature selection methods including Information Gain, Chi-Square, Gini Index, Support Vector Machine-Based, Principal Components Analysis and Symmetric Uncertainty. A wide range of comparative experiments were conducted on three Arabic standard datasets and with three classification algorithms. The experimental results clearly show the superiority of the proposed method in both cases as a standalone or in a two-stage scenario. The results show that the proposed method behaves better than traditional approaches in terms of classification accuracy with a 6-10% gain in the macro-average, F1.*

Keywords: *Feature selection method, compound words, arabic text classification.*

Received March 15, 2015; accepted December 27, 2015

1. Introduction

With the rapid advance of Internet technologies, text classification has become the key technology for classifying the huge number of electronic documents available on the World Wide Web. More approaches based on statistical theory and supervised machine learning have been commonly used in recent years, such as k-Nearest Neighbor, decision tree, neural networks, support vector machines and Naïve Bayes.

One of the most indispensable pre-processing steps in the text categorization is the Feature Selection (FS) step, especially when dealing with a dataset that has a high dimensional space of features. Feature selection is an active research topic in many areas including text classification, pattern recognition, data mining and signal processing [3, 9, 12]. The feature selection process can be defined as a task in which a subset of features is selected from the original feature space according to some criteria of feature importance. The use of a high dimensional feature space often causes computational time and performance degradation of the classifier. Therefore, the main aim of feature selection is to simplify the dataset by selecting the most discriminative features from the original feature set to reduce the feature space dimensionality without sacrificing predictive accuracy [30].

Many selection techniques have been proposed by various authors of text categorization. Despite the massive number of proposed feature selection approaches, feature selection is still one of the most active research topics in many areas of machine learning, statistics and data mining [11, 29].

Researchers are still looking for new techniques to select distinguishing features so that the performance of the classifiers can be improved and the processing time can be reduced.

The main problem of most of the current feature selection methods in text categorization is that each word is treated as a feature, or what is called single feature (s-feature), resulting in high dimensionality [11]. More powerful algorithms have been proposed for more powerful and effective text classifiers. These algorithms use more sophisticated text representations than the traditional bag-of-words such as syntactic phrases, word n-grams and non-adjacent terms [4, 11]. The aim of these representations is to reduce the inherent ambiguity of the s-features. However, the main problem with these algorithms is that they mainly depend on the frequencies of the s-features to rank them and do not take the context of these features into account when ranking them.

This paper proposes and evaluates a novel method for feature selection aimed at extracting compound words from each class and then augmenting both the training and the testing data with these compound words. The proposed algorithm attempts to alleviate the problems associated with the existing algorithms that use compound features (c-features) and to produce better ranking algorithms that take into account both the frequencies of the c-features and their context when ranking these c-features.

The paper is organized as follows: section 2 provides a review of related work. Section 3 presents the proposed feature selection method. Section 4 gives a brief outline of the implementation and the used classifiers, FS methods, datasets and performance measures. Section 5 discusses the experimental results, where three distinct Arabic reference text collections are employed, and presents a characterization of the impact of the proposed method over different classification algorithms. Finally, section 6 presents the conclusion and outlines future work.

2. Related Work

Feature selection is a process in which a subset of important and distinctive features is selected. Such a selection can help in developing efficient and accurate text categorization models [5]. Traditional feature selection processes consist of four basic steps, namely feature subset selection, subset evaluation (feature ranking), stopping criterion, and result validation [25].

The subset of features that is selected by a feature selection algorithm is either a set of single features (individual terms or s-features) or a set of compound features (c-features) or both. Table 1 summarizes some of the work for feature selection that uses s-features. Even though most of the work in feature selection is based on the use of individual terms or s-features, there are several works that use c-features. The following works show that working with c-features is better than working with just s-features.

Tan *et al*. [26] used bi-grams (c-features) in conjunction with s-features. Their method selected only 2% of the c-features, and was aimed at using only discriminative n-grams. The results showed that the combination of both c-features and s-features produced more accurate classification results than just s-features, thus presenting statistically significant gains.

Mladenic and Grobelnik [18] studied the use of word n-grams (n adjacent words/terms) with lengths of up to 3 as features for text classification. They trained Naive Bayes classifiers using only frequent n-grams with lengths of up to 3 instead of using only s-features. The experimental results showed that classifiers using frequent n-grams with lengths of up to three s-features performed better than classifiers based on s-features. However, the highest improvement was achieved when n=2.

Figueiredo *et al*. [11] proposed an effective method that obtained high discriminative c-features consisting of pairs of single-features. Their approach has three steps. In the enumeration step, the top N s-features (individual terms) are selected and ranked using information gained method. These top N single terms are combined to form the c-features list. In the selection step, the generated c-features are ranked and the top ranked ones are selected. Finally, in the augmentation step, only c-features that have high dominance in a given class are inserted into all training documents in that class. The augmentation of a test document is done by inserting all high dominance c-features that appear in the document. Experiments were conducted over four data sets, using three different classification methods. The results showed that the new feature extraction method consistently improved the text classification.

3. The Proposed Method Compound Words Statistics

The proposed method Compound Words Statistics (CWS) does not use single features (s-feature) for example (Middle, company, markets) but it directly extracts the compound features (c-features) for example (Middle east, Stock markets) and ranks them. In addition, the proposed method not only depends on the frequencies of the c-features, but the method manages to take into account both the frequencies of the c-features and their actual uses in a corpus within one ranking function. Unlike Figueiredo *et al*. [11], where single terms are ranked using a traditional feature selection method, namely Information Gain, the ranked single terms are combined to form bigram terms (c-features) and after that these c-features are ranked. The method consists of the following steps: The candidate selection (listing) step: every two adjacent words are selected as candidate c-features (bigram terms). In the ranking step, the method selects the candidate c-features that will be used to augment the documents of the training and testing sets. As the selection and ranking criterion, a compound terminology extraction method is adopted to extract compound words. This method was used by [10, 19, 20]. The top N ranked c-features (bigram terms) from each category are selected to augment the training documents from that category. The following Equation is used to rank the features:

$$\begin{aligned} RANK(c-feature, c_j) \\ = GM(c-feature, c_j) \\ * f(c-feature, c_j) \end{aligned} \quad (1)$$

Where

$$\begin{aligned} GM(c-feature, c_j) \\ = \prod_{i=1}^{n} ((L(w_i, c_j)+1)(R(w_i, c_j)+1))^{\frac{1}{2n}} \end{aligned} \quad (2)$$

$$L(x, c_j) = \sum_{i=1}^{\#L(w, c_j)} \#L_i \quad (3)$$

$$R(x, c_j) = \sum_{i=1}^{\#R(w, c_j)} \#R_i \quad (4)$$

where f(*c-feature*, c_j) is the frequency of the c-feature in class c_j, $\#L(w, c_j)$ and $\#R(w, c_j)$ are the number of distinct simple words which directly precede or succeed the word w in the class c_j and L (w, c_j) and R (w, c_j) are the cumulative frequencies of the words that directly precede or succeed w in class c_j and n is the length of the c-feature. In this paper, only the c-features of size two (bigram terms) are used.

Finally, the training data which is generated after the augmentation process is used either directly to train a machine learning classifier (standalone feature selection method) or used in a two-stage scenario (two-stage feature selection method) in which a traditional feature method is also used before the training data is used to train a machine learning classifier.

Table 1. Summary of the feature selections that use single features.

Author	**Language**	**Classification method**	**Feature selection method**	**Data set**
[27]	English	KNN C4.5 DT	Information Gain, Genetic Algorithm and Principal Component Analysis	• Reuters-21,578. • Classic3 datasets
[22]	English	NB, Centroid SVM		• Reuters-21578 • 20 Newsgroups, and • RCV1-v2datasets
[16]	English	SVM	Two-stage methods (document frequency (DF), Information Gain (IG), Mutual Information (MI) Chi-square Statistic (CHI)	• LingSpam corpus • AndrewFarrugia corpus
[21]	English	K-NN NB	ALOFT (At Least One Feature).	• Reuters-21578, • 20 Newsgroup • WebKB
[17]	Arabic	SVM	Chi, GSS3 score, GSS square (GSSS), NGL4 Coefficient Odds Ratio, MI, IG, Bi-Normal separation, Document Frequency, Power	
[24]	English	Knn fkNN SVM	Gini Index, Cross Entropy, CHI, Weight of Evidence, Information Gain	• Reuters-21578
[28]	English	DT SVM NN	Chi Square, Gini index, Information Gain and deviation from Poisson distribution Distinguishing Feature Selector (DFS)	• Reuters-21578 • 20Newsgroups. • Short Message Service (SMS) • Enron1.
[14]	English Chinese	NB	Multi-class Odds Ratio and Class Discriminating Measure	• Reuters-21578 • Chinese text classification corpus
[1]	English	KNN	Chi Square, Information Gain, ant colony optimization based feature selection algorithm	• Reuters-21578 dataset
[15]	English	NB	Information gain (IG), Chi Square, an enhanced ACO algorithm	• 20Newsgroup

4. Experimental Work

In this section, an in-depth investigation is carried out to make a comparison between the compound words method of feature selection with the other well-known feature selection methods such as Information Gain, Chi-square, Gini index, Uncertainty, SVM-based and PCA. This comparison will take into consideration the classification accuracy. To achieve this, the process of comparison is carried out on three Arabic standard datasets and many performance measures will be used in order to test to what extent the proposed methods are effective under various conditions. The datasets, feature selection methods, classification algorithms and performance measures used in this paper will be described briefly.

4.1. Datasets

Three published Arabic datasets were used in the experiments: CNN, BBC and OSAC Arabic corpora. All the corpora used in this paper were collected and published by [23] and they are publically available at (http://sourceforge.net/projects/ar-text-mining/). The datasets used in this paper are described below.

4.2. Feature Selection Methods

As pointed out in section 2 and table 1, there are many feature selection techniques for the selection of distinctive features in text classification. Among all those techniques, the Information Gain, Principal Components Analysis, Gini Index, Uncertainty, Chi-square and SVM-based methods have been proven to be much more effective for text categorization [7, 13, 25]. Therefore, these feature selection methods have been selected for comparison with the proposed algorithm when it is applied as a standalone feature selection method and to show the effect of the proposed algorithm on their performance when the feature selection is done in two-stage scenario. Since the main concern here is to evaluate the performance of the proposed algorithm and to compare it with these algorithms, the performance of these feature selection algorithms in addition to the proposed method will be studied with three state-of-the-art machine learning algorithms.

4.3. Classification Methods

In the machine learning workbench, some classifiers like support vector machine, Naïve Bayes and K-NN have achieved great success in text categorization [2, 6, 8]. To evaluate the effects of the proposed feature selection method over existing methods, these three machine learning classification methods are used in three datasets.

4.4. Performance Measures

In order to assess the utility of the various feature selection methods, the F1-measure, that combines both precision and recall, is utilized. Precision is defined as the ratio of retrieved instances that are related, whereas recall is defined as the proportion of relevant instances that are retrieved. For ease of comparison, the Macro-averaged (Macro-F1) is used. Macro-averaging assigns equal weight to each class no matter what the class frequency, and it is often dominated by the performance of the system on most common categories. The Macro-average F1 is calculated by the following formula:

$$F_1^{\text{macro}} = \frac{1}{m}\sum_{i=1}^{m} F_1(\mathrm{c}_i) \qquad (5)$$

5. Results and Discussion

In order to demonstrate the performance of the proposed feature selection approach, two types of experiments were conducted. In the first type of experiments, the proposed method was used as the first step of a two-stage feature selection algorithm, where the second step was one of the six well-known feature selection methods (Chi-Square, Inf. Gain, Gini Index, SVM-based, PCA and Uncertainty). In the second type of experiments, the proposed method was used as a standalone feature selection method and its performance was compared to the best results achieved by one of the six well-known feature selection methods. All these experiments were carried out on the three datasets CNN, BBC and OSAC Arabic corpora.

Tables 2, 3, and 4 compare the behaviors of the six well-known feature selection methods (Chi-Square, Inf. Gain, Gini Index, SVM-based, PCA and Uncertainty) in both cases (standalone and two-stage) with the NB, K-NN and SVM classifiers on the three datasets, CNN Arabic corpus, BBC Arabic corpus and OSAC Arabic corpus, respectively. The features were selected from the feature space at different sizes (500, 1000, 2000, 3000, 4000, and 5000).

Table 2 shows that the performance of all six feature selection methods (Chi-Square, Inf. Gain, Gini Index, SVM-based, PCA and Uncertainty) in combination with the proposed method (two-stage scenario) was better than their original performance. As shown in Table 2, which is based on the CNN Arabic corpus, the two-stage feature selection method was superior to all the six feature selection methods. Generally speaking, the proposed method significantly improved the categorization performance.

Table 3 also shows that the performance of all the six feature selection methods (Chi-Square, Inf. Gain, Gini Index, SVM-based, PCA and Uncertainty) in combination with the proposed method (two-stage scenario) was much better than their original performance in all cases. The results shown in Table 3 are based on the BBC Arabic corpus.

Table 4 shows the results that were obtained based on the OSAC Arabic corpus. Table 4 shows that the two-stage feature selection method did slightly better than the original methods due to the very high results of the original methods. In fact, all the experiments on this dataset achieve a high classification performance, regardless of the classification method or feature selection method used. This means that the dataset itself was inappropriate for the task of evaluating the text classification because there was not much difficulty in classifying this dataset. Also this confirms that, the characteristics of the dataset have an important effect on performance of text classifiers. The second type of experiment was aimed at comparing the performances of the proposed method CWS as a standalone feature selection method with the best results obtained by the well-known feature selection methods. These experiments were also conducted on the three datasets; CNN, BBC and OSAC Arabic corpora, and three classifiers namely NB, K-NN and SVM. In what follows, the results of these experiments on the CNN, BBC and OSAC Arabic corpora are described in Figures 1, 2, and 3, respectively.

Table 2. Comparisons of the performance (Macro-F1) of one-stage and two-stage feature selection methods with the three classifiers on the CNN Arabic corpus.

		Macro-F											
		One Stage						**Two Stage**					
	Feature size	500	1000	2000	3000	4000	5000	500	1000	2000	3000	4000	5000
NB-CNN	Chi-Square	79.23	84.72	86.99	86.91	86.62	85.73	93.09	93.78	94.3	94.34	94.26	94.31
	Inf. Gain	79.06	85.03	87.16	87	86.57	86.01	92.47	94.08	94.4	94.2	94.47	94.21
	Gini Index	77.66	84.22	86.8	86	86.12	86.12	92.39	93.41	94.31	94.33	94.28	94.31
	SVM	86.23	87.72	88.6	87.45	86.97	85.73	94.09	94.47	94.33	94.34	94.21	94.23
	PCA	65.97	75.27	85.37	87.15	86.54	85.73	82.48	91.59	94.22	94.34	94.26	94.31
	Uncertainty	79.22	84.65	86.81	87.04	86.36	86.01	92.43	94.04	94.32	94.2	94.47	94.21
KNN-CNN	Chi-Square	79.51	82.54	80.88	79.35	84.69	89.65	83.56	85.4	80.4	91.56	91.5	91.31
	Inf. Gain	79.31	83.57	80.85	88.11	88.37	89.12	82.95	85.85	82.5	91.51	91.32	91.37
	Gini Index	77.86	82.2	84.36	87.94	87.43	89.48	82.99	86.05	82.61	91.52	91.32	91.22
	SVM	83.8	83.46	80.91	82.73	85.65	89.4	86.85	85.83	90.85	91.31	91.39	91.28
	PCA	66	73.17	76.56	77.99	83.57	89.22	73.31	76.12	78.42	91.38	91.51	91.44
	Uncertainty	79.91	79.91	84.36	83.14	85.82	89.48	82.06	80.39	79.61	91.52	91.32	91.22
SVM-CNN	Chi-Square	86.17	90.41	92.03	92.61	92.77	92.74	91.36	92.81	94.02	94.27	94.35	94.28
	Inf. Gain	86.59	90.29	92.15	92.95	92.97	92.51	92.3	93.7	94.29	94.33	94.14	94.26
	Gini Index	84.65	89.59	92.29	92.7	93	92.81	91.79	93.4	94.1	94.34	94.28	94.17
	SVM	92.11	93.73	94.19	93.92	93.4	92.74	93.81	94.55	94.39	94.27	94.35	94.28
	PCA	71.24	83.85	91.28	91.97	92.37	92.51	84.33	91.63	94.13	94.33	94.14	94.26
	Uncertainty	86.58	90.29	92.32	92.41	92.96	92.81	91.84	93.34	93.97	94.34	94.28	94.17

Table 3. Comparisons of the performance (Macro-F1) of one-stage and two-stage feature selection methods with the three classifiers on the BBC Arabic corpus.

		Macro-F											
		One Stage						**Two Stage**					
	Feature size	500	1000	2000	3000	4000	5000	500	1000	2000	3000	4000	5000
NB-BBC	Chi-Square	66.59	76.62	79.9	77.13	75.74	75.98	92.27	94.86	94.88	94.87	94.88	94.88
	Inf. Gain	77.04	81	81.38	79.02	75.77	75.82	95.21	94.82	94.88	94.87	94.86	94.87
	Gini Index	73.39	79.63	79.69	77.97	76.31	75.12	95.09	94.88	94.87	94.87	94.89	94.88
	SVM	76.08	77.5	80.03	77.13	75.74	75.98	94.97	94.86	94.88	94.87	94.88	94.88
	PCA	68.04	75.99	77.69	77.02	75.77	75.82	94.59	94.82	94.88	94.87	94.86	94.87
	Uncertainty	73.13	79.52	80.27	77.94	76.31	75.12	92.51	94.88	94.87	94.87	94.89	94.88
KNN-BBC	Chi-Square	74.68	76.06	72.21	68.07	81.09	80.57	83.48	87.07	87.2	87.11	86.92	87.36
	Inf. Gain	78.42	77.12	75.29	73.42	80.57	80.43	86.87	87.31	87.23	87.08	86.91	87.15
	Gini Index	77.98	76.39	74.98	73.71	80.68	79.66	87.22	86.99	86.94	86.93	87.36	87.05
	SVM	75.08	78.08	75.35	70.44	81.09	80.57	87.91	87.07	87.2	87.11	86.92	87.36
	PCA	76.14	76.49	74.52	75.19	80.57	80.43	88.27	87.31	87.23	87.08	86.91	87.15
	Uncertainty	77.01	76.31	72.64	69.43	80.68	79.66	83.96	86.99	86.94	86.93	87.36	87.05
SVM-BBC	Chi-Square	93.15	90.05	89.01	88.55	88.62	88.52	94.6	94.76	94.64	94.65	94.63	94.72
	In. Gain	90.94	88.56	88.66	88.36	88.59	87.33	94.77	94.7	94.79	94.74	94.64	94.6
	Gini Index	90.04	88.14	87.75	88.33	87.63	88.45	94.71	94.69	94.58	94.62	94.66	94.63
	SVM	93.76	91.28	90.32	89.22	88.62	88.52	94.92	94.76	94.64	94.65	94.63	94.72
	PCA	91.95	90.05	88.19	87.75	88.59	87.33	94.66	94.7	94.79	94.74	94.64	94.6
	Uncertainty	91.2	88.82	88.74	88.81	87.63	88.45	94.66	94.69	94.58	94.62	94.66	94.63

Table 4. Comparisons of the performance (Macro-F1) of one-stage and two-stage feature selection methods with the three classifiers on OSAC Arabic corpus.

		Macro-F											
		One Stage						**Two Stage**					
	Feature size	500	1000	2000	3000	4000	5000	500	1000	2000	3000	4000	5000
NB-OSAC	Chi-Square	96.92	97.46	97.58	97.64	97.61	97.6	97.22	97.7	97.78	97.82	97.83	97.82
	Inf. Gain	96.23	97.36	97.6	97.62	97.6	97.59	97.09	97.76	97.81	97.83	97.82	97.82
	Gini Index	96.55	97.19	97.6	97.61	97.58	97.59	97.54	97.7	97.81	97.83	97.82	97.82
	SVM	97.55	97.67	97.68	97.67	97.66	97.59	97.78	97.8	97.81	97.82	97.83	97.82
	PCA	96.22	97.44	97.54	97.62	97.61	97.61	97.1	97.75	97.8	97.83	97.82	97.82
	Uncertainty	96.76	97.37	97.61	97.65	97.61	97.62	96.96	97.65	97.8	97.83	97.82	97.82
KNN-OSAC	Chi-Square	97.55	97.57	97.55	96.97	96.13	97.15	97.71	97.72	97.69	97.7	97.71	97.7
	Inf. Gain	97.55	97.46	97.42	96.88	97.22	95.94	97.67	97.73	97.67	97.7	97.72	97.7
	Gini Index	97.29	97.51	97.44	96.9	96.27	95.34	97.71	97.73	97.71	97.72	97.68	97.7
	SVM	97.6	97.5	97.33	96.84	97.52	97.44	97.76	97.73	97.7	97.7	97.71	97.7
	PCA	97.64	97.6	97.47	97.4	97.26	92.73	97.72	97.76	97.58	97.7	97.72	93.7
	Uncertainty	97.59	97.46	97.49	96.91	97.37	95.58	94.69	97.76	97.69	97.72	97.68	97.7
SVM-OSAC	Chi-Square	97.54	97.68	97.7	97.74	97.76	97.75	97.64	97.71	97.74	97.8	97.79	97.79
	In. Gain	97.35	97.67	97.72	97.76	97.76	97.76	97.6	97.71	97.79	97.8	97.8	97.8
	Gini Index	97.18	97.64	97.71	97.76	97.74	97.76	97.6	97.69	97.8	97.8	97.79	97.8
	SVM	97.73	97.76	97.76	97.76	97.76	97.75	97.79	97.8	97.8	97.8	97.79	97.79
	PCA	97.54	97.68	97.7	97.72	97.74	97.74	97.66	97.73	97.73	97.8	97.8	97.8
	Uncertainty	97.49	97.66	97.71	97.75	97.76	97.77	97.59	97.69	97.74	97.8	97.79	97.8

According to the results shown in Figure 1, it can be observed that the CWS performed better than the other feature selection methods in all classifiers when classifying the CNN Arabic corpus.

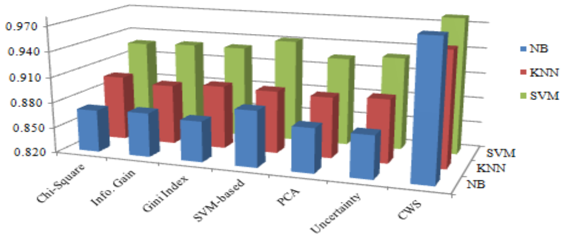

Figure 1. Comparisons of the performance of the different feature selection methods with the three classifiers on the CNN Arabic corpus.

As an illustration, for the CWS, the obtained accuracies in the NB, K-NN and SVM classifiers were 98%, 95%, and 98% respectively, which were much higher than other feature selection methods. The results shown in Figure 2 indicate that the performance of the three classifiers with the CWS was comparable to other feature selection methods based on the classification of the BBC Arabic corpus. From Figure 2, it can be seen that the CWS method achieved the best classification accuracies in all cases. In terms of the classification algorithm, Figures 1 and 2 shows that the highest accuracies were obtained when the SVM classifier was used.

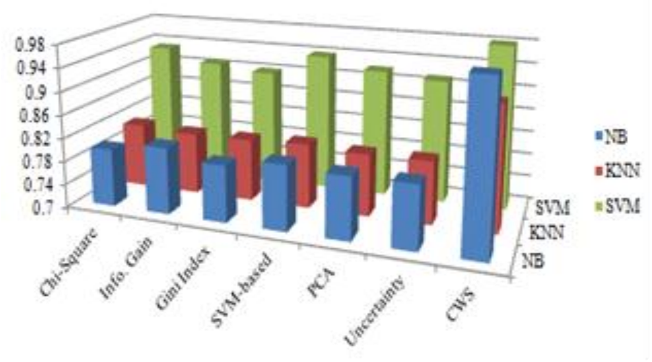

Figure 2. Comparisons of the performance of the different feature selection methods with the three classifiers on the BBC Arabic corpus.

Figure 3 shows the accuracy of the classification process that was performed on the OSAC Arabic corpus. It can be seen that the CWS method performed better than other methods when the NB and SVM classifiers were used.

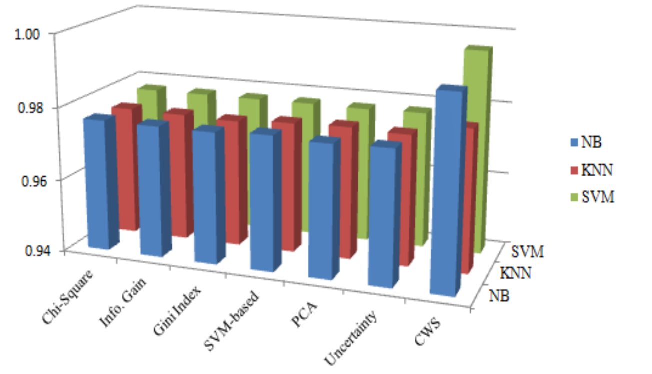

Figure 3. Comparisons of the performance of the different feature selection methods with the three classifiers on OSAC Arabic corpus.

Figures 1, 2, and 3 show that when the proposed feature selection method (CWS) was used, the classification performance was obviously superior in all cases, regardless of the classification method used. To compare the behaviors of NB, SVM and K-NN classifiers, the experiments were performed on three datasets namely BBC, CNN and OSAC Arabic corpora. As shown in Figures 1 and 2, the SVM algorithm showed a higher performance than the NB and K-NN algorithms. This outcome proves that the effect of different representation strategies on the performance of text classification algorithms. In our case, the results confirm this observation where representation of the class features as compound terms produced more desirable performance than single features produced.

6. Conclusions and Future Work

In literature, two major factors can be distinguished that complicate the classification of documents. The first factor is how to define the set of features that better identify the class to which each document belongs. The second factor is concerning the best method to be used in order to learn effective document classifiers once the set of features has been defined. In this paper an attempt was made to address the first issue, as well as to evaluate how traditional classification algorithms can benefit from the proposed method in order to obtain the best classification results. More specifically, this paper focused on extracting new features called compound words that are relevant to the classification task. The compound words are sets of terms that co-occur in any part of a document. An effective method was proposed that allowed high distinctive c-features to be obtained. The method has three steps. The first step is the listing step that selects the best terms to be used to form the c-features. In the ranking step, the method selects the candidate c-features that will be used to augment the documents of the training and testing sets. Finally, the extracted features are used as the representation features of the document.

For future work, it is suggested that other text mining techniques be explored to take advantage of other relations among terms. For example, using closed term sets to elicit new discriminative features. Also, using sizes greater than two for c-features could be explored.

References

[1] Aghdam M., Ghasem-Aghaee N., and Basiri M., “Text Feature Selection using Ant Colony Optimization,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 6843-6853, 2009.

[2] Alhutaish R. and Omar N., “Arabic Text Classification using K-Nearest Neighbour Algorithm,” *The International Arab Journal of Information Technology*, vol. 12, no. 2, pp. 190-195, 2015.

[3] Baccianella S., Esuli A., and Sebastiani F., “Feature Selection for Ordinal Text Classification,” *Neural Computation*, vol. 26, no. 3, pp. 557-591, 2014.

[4] Bespalov D., Bai B., Qi Y., and Shokoufandeh A., “Sentiment Classification based on Supervised Latent N-Gram Analysis,” *in Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, Glasgow, pp. 375-382, 2011.

[5] Chen Y., Sun Y., and Han B., “Improving Classification of Protein Interaction Articles using Context Similarity-based Feature Selection,” *Biomed Research International*, 2015.

[6] D’orazio V., Landis S., Palmer G., and Schrodt P., “Separating the Wheat from the Chaff: Applications of Automated Document

Classification using Support Vector Machines," *Political Analysis*, vol. 22, no. 2, pp. 224-242, 2014.

[7] Dai J. and Xu Q., "Attribute Selection based on Information Gain Ratio in Fuzzy Rough Set Theory with Application to Tumor Classification," *Applied Soft Computing*, vol. 13, no. 1, pp. 211-221, 2012.

[8] Dai Y. and Sun H., "The Naive Bayes Text Classification Algorithm based on Rough Set in the Cloud Platform," *Journal of Chemical and Pharmaceutical Research*, vol. 6, no. 7, pp. 1636-1643, 2014.

[9] De Stefano C., Fontanella F., Marrocco C., and Di Freca A., "A GA-based Feature Selection Approach with an Application to Handwritten Character Recognition," *Pattern Recognition Letters*, vol. 35, pp. 130-141, 2014.

[10] Dias G. and Kaalep H., "Automatic Extraction of Multiword Units for Estonian: Phrasal Verbs," *Languages in Development*, vol. 41, pp. 81-89, 2003.

[11] Figueiredo F., Rocha L., Couto T., Salles T., Gonçalves M., and Meira W., "Word Co-Occurrence Features for Text Classification," *Information Systems*, vol. 36, no. 5, pp. 843-858, 2011.

[12] Ganapathy S., Vijayakumar P., Yogesh P., and Kannan A., "An Intelligent CRF based Feature Selection for Effective Intrusion Detection," *The International Arab Journal of Information Technology*, vol. 13, no. 1, pp. 44-50, 2016.

[13] Gao Z., Xu Y., Meng F., Qi F., Lin Z., "Improved Information Gain-based Feature Selection for Text Categorization," *in Proceedings of 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace and Electronic Systems*, Aalborg, pp. 1-5, 2014.

[14] Li S., Xia R., Zong C., and Huang C., "A Framework of Feature Selection Methods for Text Categorization," *in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, pp. 692-700, 2009.

[15] Meena M., Chandran K., Karthik A., and Vijay A., "An Enhanced ACO Algorithm to Select Features for Text Categorization and its Parallelization," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5861-5871, 2012.

[16] Meng J., Lin H., and Yu Y., "A Two-Stage Feature Selection Method for Text Categorization," *Computers and Mathematics with Applications*, vol. 62, no. 7, pp. 2793-2800, 2011.

[17] Mesleh A., "Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study," *in Proceedings of Advances in Computer and Information Sciences and Engineering*, pp. 11-16, 2008.

[18] Mladenic D. and Grobelnik M., "Feature Selection for Unbalanced Class Distribution and Naive Bayes," *in Proceedings of the 16th International Conference on Machine Learning*, San Francisco, pp. 258-267, 1999.

[19] Nakagawa H., "Automatic Term Recognition based on Statistics of Compound Nouns," *Terminology*, vol. 6, no. 2, pp. 195-210, 2001.

[20] Nakagawa H. and Mori T., "A Simple but Powerful Automatic Term Extraction Method," *in Proceedings of 2nd International Workshop on Computational Terminology*, Stroudsburg, pp. 1-7, 2002.

[21] Pinheiro R., Cavalcanti G., Correa R., and Ren T., "A Global-Ranking Local Feature Selection Method for Text Categorization," *Expert Systems with Applications*, vol. 39, no. 17, pp. 12851-12857, 2012.

[22] Ren F. and Sohrab M., "Class-Indexing-based Term Weighting for Automatic Text Classification," *Information Sciences*, vol. 236, pp. 109-125, 2013.

[23] Saad M., the Impact of Text Preprocessing and Term Weighting on Arabic Text Classification, Theses, Master of Science, Computer Engineering, the Islamic University, 2010.

[24] Shang W., Huang H., Zhu H., Lin Y., Qu Y., and Wang Z., "A Novel Feature Selection Algorithm for Text Categorization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 1-5, 2007.

[25] Singh B., Kushwaha N., and Vyas O., "A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty," *Journal of Data Analysis and Information Processing*, vol. 2, no. 4, pp. 95-105, 2014.

[26] Tan C., Wang Y., and Lee C., "The Use of Bigrams to Enhance Text Categorization," *Information Processing and Management*, vol. 38, no. 4, pp. 529-546, 2002.

[27] Uğuz H., "A Two-Stage Feature Selection Method for Text Categorization by using Information Gain, Principal Component Analysis and Genetic Algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024-1032, 2011.

[28] Uysal A. and Gunal S., "A Novel Probabilistic Feature Selection Method for Text Classification," *Knowledge-Based Systems*, vol. 36, pp. 226-235, 2012.

[29] Vege S., Ensemble of Feature Selection Techniques for High Dimensional Data, Theses, Western Kentucky University, 2012.

[30] Wang J., Zhou S., Yi Y., and Kong J., "An Improved Feature Selection based on Effective Range for Classification," *The Scientific World Journal*, pp. 1-8, 2014.

Aisha Adel is PhD candidate in UKM, Malaysia. She earned her MSc degree in 2014 in computer science from UKM, Malaysia. BSc degree in 2009 UST Yemen. Her research interests are on machine learning and optimization algorithms.

Nazlia Omar is currently an Associate Professor at the Center for AI Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia. She holds her PhD in Computer Science from the University of Ulster, UK. Her main research interest is in the area of Natural Language Processing and Computational Linguistics.

Mohammed Albared obtained his BSc in Computer Science from Yarmouk University, Jordan. He obtained his master degree in Computer Science from Yarmouk University, Jordan. He did his PhD in Computer Science at Universiti Kebangsaan Malaysia. Now, he is working as an Assistant professor at Sana'a University. His research interest falls under Natural Language Processing (NLP), Machine Learning, Text and Web Mining, and Sentiment Analysis.

Adel Al-shabi earned his PhD degree in 2018 and MSc degree in 2013 in computer science at Universiti Kebangsaan Malaysia. He obtained his BSc degree in 2006 at National University, Yemen. His research interests are on machine learning and sentiment analysis.