

# An Efficient Feature Selection Method for Arabic Text Classification

Bilal Hawashin  
Department of Computer  
Information Systems,  
Alzaytoonah University of Jordan,  
Amman 11733, Jordan,

Ayman M Mansour  
Department of Electrical and  
Computer Engineering,  
Tafila Technical University  
Tafila 66110, Jordan,

Shadi Aljawarneh  
Department of Software  
Engineering,  
Al-Isra University,  
Amman 11622, Jordan,

## ABSTRACT

This paper proposes an efficient, Chi-Square-based, feature selection method for Arabic text classification. In Data Mining, feature selection is a preprocessing step that can improve the classification performance. Although few works have studied the effect of feature selection methods on Arabic text classification, limited number of methods was compared. Furthermore, different datasets were used by different works. This paper improves the previous works in three aspects. First, it proposes a new efficient feature selection method for enhancing Arabic text classification. Second, it compares extended number of existing feature selection methods. Third, it adopts two publicly available datasets to encourage future works to adopt them in order to guarantee fair comparisons among the various works. Our experiments show that our proposed method outperformed the existing methods in term of accuracy.

## Keywords

Data Mining, Arabic Text Retrieval, Feature Selection, CHI Square.

## 1. INTRODUCTION

Text Classification is a data mining application that automatically assigns one or more predefined labels to free text items based on their content[9]. Currently, the amount of the available text data in the web is increasing daily. This huge size makes the process of classifying it manually a very difficult and time consuming task. Therefore, the trend of automatically classifying text data has been introduced. Text Classification is used in many fields, such as filtering emails, digital libraries, online databases, and online news.

Although a lot of works have studied classification of English texts, few works have studied the classification of Arabic texts. For example, [13] studied the performance of C5.0 and Support Vector Machines classifiers on Arabic texts, where the latter outperformed the former with accuracies 0.78 and 0.69 respectively. [12] evaluated Naïve Bayes on classifying Arabic web texts, and its accuracy was 0.68. [4] investigated the performance of CBA, Naïve Bayes, and SVM on classifying Arabic texts. The results showed that CBA outperformed NB and SVM and its accuracy was 0.8. [10] compared the performance of SVM and KNN on Arabic texts. SVM outperformed KNN.

As text items are represented using a term document matrix, where every row represents a term, and every column represents a text item, the original number of terms could be huge, which could negatively affect the classification performance. Therefore, one of the important preprocessing steps in the text classification, and in data mining applications generally, is the feature selection.

In this step, only the important terms are selected, which would reduce the space consumption and can improve classification accuracy by eliminating the noisy terms.

Few works have studied the effect of feature selection on Arabic text classification. For example, [4] studied the effect of the maximum entropy method in classifying Arabic texts, and its accuracy was 0.80. [2] showed that SVM classifier in combination with Chi-square-based feature selection is an appropriate method to classify Arabic texts. [5] evaluated the effect of Ngram frequency statistics on classifying Arabic texts. [14] compared TF.IDF, DF, LSI, Stemming, and Light Stemming. Their work showed that the former three methods outperformed the latter two stemming methods.

In most of the previous works, limited number of methods was used. Besides, different works used different datasets, which made the comparison of their methods difficult. Furthermore, the sizes of the used datasets were rather small, which could affect the experiment results. In this paper, the previous works are extended by comparing more feature selection methods. Two publicly available datasets were used in order to make the different works comparable. Furthermore, an improved Chi-square -based method will be proposed and its effect in the Arabic text classification performance will be analyzed.

In this paper, our proposed method will be compared with the regular Chi-square statistics[16], Information Gain[15], Mean TF.IDF [12], DF[16], Wrapper Approach with SVM Classifier[11], Feature Subset Selection[7], and a CHI square variant. Most of these methods have strong theoretical foundations and have proved their superiority in feature selection of English texts. In order to evaluate their performance, two publicly available datasets, Akhbar Alkhalij and Alwatan datasets[1] were used. SVM classifier is used to classify the texts after the feature selection process.

The contributions of this work are as follows.

- Proposing a new improved Chi-square-based method.
- Extending previous works by comparing more existing feature selection methods according to their performance in classifying Arabic texts.
- Adopting the use of two publicly available datasets in an attempt to make different works comparable.

In what follows, various existing feature selection methods will be described to be compared (Compared Feature Selection Methods section), our proposed improved method (Improved Chi-square-Based Feature Selection Method section), illustrates phase one of the experimental part (Comparing Regular And

Improved Chi-square Methods section), which compares this method with the regular Chi-square method and another Chi-square variant, phase two of the experimental part (Comparing Our Method With Existing Feature Selection Methods section), which compares our method with various existing feature selection methods according to their effect in classifying Arabic texts, and conclusion (Conclusion section).

## 2. COMPARED FEATURE SELECTION METHODS

Various feature selection methods which will be compared with our method in the experimental parts are described in this section. In this context, the words feature and term are used interchangeably.

### 2.1 Chi-square

Chi-square is a well known statistic measurement that has been used in feature selection [16]. This method assigns a numerical value for each term that appears at least once in any document. The value of a term  $w$  is calculated as follows:

$$\text{Val}(w) = \sqrt{\frac{(n_{pt+} + n_{nt+} + n_{pt-} + n_{nt-})(n_{pt+}n_{nt-} - n_{pt-}n_{nt+})^2}{(n_{pt+} + n_{pt-})(n_{nt+} + n_{nt-})(n_{pt+} + n_{nt+})(n_{pt-} + n_{nt-})}}, \quad (1)$$

Where  $n_{pt+}$  and  $n_{nt+}$  are the number of text documents in the positive category and the negative category respectively in which term  $w$  appears at least once. The positive and negative categories are used to find the accuracy measurements per class when multiple classes are used such that the positive category indicates a class and the negative category indicates the remaining classes.  $n_{pt-}$  and  $n_{nt-}$  are the number of text documents in the positive category and the negative category respectively in which the term  $w$  does not occur. The value of each term represents its importance. The terms with the highest values are the most important terms.

### 2.2 Mean TF.IDF

According to this method [12], term document matrix is constructed for the training set, and TF.IDF weighting method is used to weight each term in each training document. TF.IDF of the term  $w$  in document  $d$  is calculated as follows.

$$\text{TF.IDF}(w,d) = \log(\text{tf}_{w,d} + 1) \cdot \log(\text{idf}_w), \quad (2)$$

where  $\text{tf}_{w,d}$  is the frequency of the term  $w$  in document  $d$ ,  $\text{idf}_w$  is  $\frac{N}{n_w}$ , where  $N$  is the number of training documents, and  $n_w$  is the number of training documents that contains the term  $w$ .

Later, Mean TF.IDF is calculated for each term using the following equation.

$$\text{Val}(w) = \sum_d \frac{\text{TF.IDF}(w,d)}{\text{Count}(d)} \quad (3)$$

Where  $\text{Count}(d)$  is the total number of documents in the dataset. The features with higher Mean TF.IDF values are selected.

### 2.3 Document Frequency (DF)

Here, each term is valued according to the number of documents that contains this term. The more the documents that contain this term, the more the  $DF$  value, and the more its importance.

$$DF(w) = n_w, \quad (4)$$

where  $n_w$  is the number of training documents that contains the term  $w$ .

### 2.4 Information Gain (IG)

Information Gain [15] is a probability based feature selection method that uses the following formula.

$$IG = H(\text{Class}) - H(\text{Class}|\text{Feature}), \quad (5)$$

$$\text{where } H(\text{Class}) = - \sum_{Class_i \in \text{Class}} P(\text{Class}_i) \log P(\text{Class}_i) \quad (6)$$

and  $H(\text{Class}|\text{Feature}) =$

$$- \sum_{F_i \in \text{Feature}} P(F_i) \sum_{Class_i \in \text{Class}} P(\text{Class}_i | F_i) \log(P(\text{Class}_i | F_i)), \quad (7)$$

Where  $P(\text{Class}_i)$  is the probability of Class  $i$ ,  $P(F_i)$  is the probability of Feature  $i$ , and  $P(\text{Class}_i|F_i)$  is the probability of Class  $i$  given Feature  $i$ .

### 2.5 Feature Subset Selection(FSS)

This method [3] evaluates the importance of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

### 2.6 Wrapper Approach

This method [8] evaluates feature sets by using a learning method. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes. This method could improve the classification accuracy but with a significant increase in the feature selection time.

## 3. IMPROVED CHI-SQUARE-BASED FEATURE SELECTION METHOD

This section presents our Chi-square-based improved feature selection method.

**Algorithm1: \_\_CHI-SQUARE BASED EQUAL CLASS FEATURE SELECTION METHOD**

**Input:** Term Document matrix TD represents the training set of D documents, T terms, and C classes.

The requested number of reduced features  $R < T$ .

**Output:** Reduced Term Document Matrix RD.

**Algorithm:**

```

01 //Find the Chi-square value for each attribute in the
    training set using the regular Chi-square method.
02 //Sort the attributes according to their Chi-square values
    in a descending order and store in a vector ORDERED.
03 Num_Selected_Features = 0;
04 Cntr = 0;
05 FPC = R/C; //The number of features per class.
06 Int featuresPerClass[C] = {0};
07 While((Num_Selected_Features < R) && (Cntr < T))
08     {Current = ORDERED[Cntr];
09     L ← LabelOf(Current);
10     If featuresPerClass[L] < FPC
11         {featuresPerClass[L]++;
12         S = S U Current;
13         Num_Selected_Features++;}
14     Cntr++;
15 }// End While
16 Return RD, a SxD matrix.
    
```

**Algorithm2: \_\_CHI-SQUARE BASED EQUAL CLASS FEATURE SELECTION WITH COSINE SIMILARITY**

**Input:** Reduced Term Document matrix RD represents the training set of D documents, R reduced terms, and C classes.

**Output:** A vector RESULT, which will be sent to the classifier.

**Algorithm:**

```

01 Find the pairwise cosine similarity among documents
    in RD. Store the output in Result, which is DxD.
02 Return Result.
    
```

Our improved Chi-square-based feature selection method is described below. Our method is composed of two algorithms; Algorithm 1 and Algorithm 2. In Algorithm 1, the input is a term document matrix, where every row represents a term, and every column represents a document. TF.IDF weighting [12] was used to weight the features in that matrix. This method was already described in (Compared Feature Selection Methods section), equation 2. Another input is the user defined number of reduced features. Algorithm1 applies the regular Chi-square method to evaluate each feature according to its importance. Next, the algorithm selects the features that belong to different classes equally. The resulting reduced features represent the various classes in the training set equally. Algorithm1 outputs the reduced term document matrix RD, where every record represents a reduced feature, and every column represents a document. This matrix serves as an input to Algorithm2, where the pairwise cosine similarity among the documents in RD is calculated to eliminate the noise. The output of this algorithm is a document by document matrix, which is the input to the classifier.

**4. COMPARING REGULAR AND IMPROVED CHI-SQUARE METHODS**

In order to evaluate the previous Chi-square-based improved method in Arabic Text Classification, two datasets were used, Akhbar Alkhalij News and Alwatan News. These datasets are publicly available from [1]. These two datasets were adopted here to encourage the future works to use them in order to make the various methods comparable. In our experiments, a subset from each dataset will be used. Table 1 below describes the use of these datasets in phase1.

The following is a brief description of each dataset.

**4.1 Akhbar Alkhalij**

This dataset is based on Akhbar Alkhalij news, and it is publicly available on [1]. A subset of 5692 texts will be used, each of which has one of four classes. 200 records were selected for the training dataset, and the remaining 5492 records for the testing datasets as displayed in Table 1. The distribution of the classes in the used portion of the dataset is represented in Table 2.

**4.2 Alwatan News**

This dataset is based on Alwatan news, and it is publicly available on [1]. A subset of 5250 texts were used, each of which has one of five classes. 250 records were selected for the training dataset, and the remaining 5000 records for the testing datasets as displayed in Table 1. The distribution of the classes in the used portion of the dataset is represented in Table 3.

For our experiments, An Intel® Xeon® server of 3.16GHz CPU and 2GB RAM was used, with Microsoft Windows Server 2003 Operating System. Also, Microsoft Visual Studio 6.0 was used to read the datasets, and Weka 3.6.2 for both the implementations of the feature selection methods and the SVM classifier.

**Table 1. Datasets Description.**

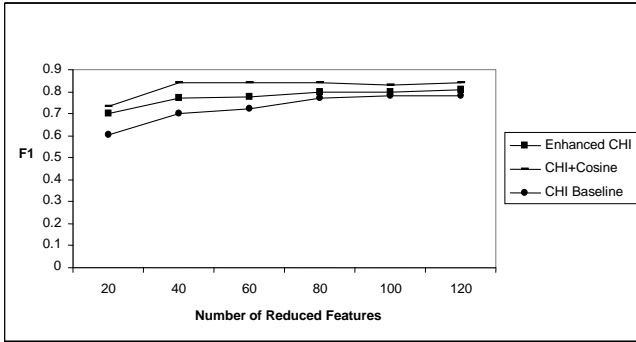
Dataset	Number of Training Records	Number of Testing Records	Number of Classes
Akhbar Alkhalij	200	5492	4
Alwatan	250	5000	5

**Table 2. Distribution of Classes In Akhbar Alkhalij.**

Dataset	Number of Training Records	Number of Testing Records
Sport	50	1380
Economy	50	859
Local News	50	2348
International News	50	905

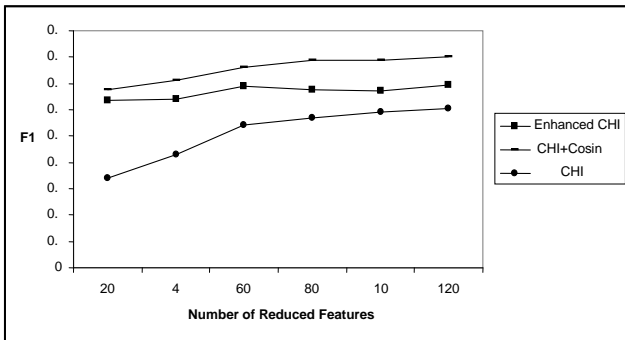
**Table 3. Distribution of Classes In Alwatan.**

Dataset	Number of Training Records	Number of Testing Records
Religion	50	1000
Economy	50	1000
Local News	50	1000
International News	50	1000
Sport	50	1000



**Fig 1: Comparing our proposed method (CHI + Cosine) with the regular Chi-square method (CHI Baseline) and another Chi-square variant (Enhanced CHI) according to their effect on the accuracy of Classifying Akhbar Alkhalij dataset. Results showed that our improved method outperformed the other two methods.**

In order to compare the performance of the previously mentioned feature selection methods according to their effect on Arabic document classification, SVM classifier was used to classify the documents in the reduced space.  $F_1$ , Feature Selection Time, Classifier Training Time, and Classification Time were used. They are described as follows.



**Fig 2: Comparing our proposed method (CHI + Cosine) with the regular Chi-square method (CHI Baseline) and another Chi-square variant (Enhanced CHI) according to their effect on the accuracy of Classifying Alwatan dataset. Results showed that our improved method outperformed the other two methods.**

- Classifier  $F_1$  rating is the harmonic mean of the classifier recall and the precision. It is given as

$$F_1 = \frac{2 * R * P}{R + P}, \quad (8)$$

where  $R$  represents the recall, which is the ratio of the relevant data among the retrieved data, and  $P$  represents the precision, which is the ratio of the accurate data among the retrieved data. Their formulas are given as follows.

$$R = \frac{TP}{TP + FN}, \text{ if } TP + FN > 0,$$

Otherwise undefined. (9)

$$P = \frac{TP}{TP + FN}, \text{ if } TP + FN > 0,$$

otherwise undefined. (10)

In order to find these measurements, a two-by-two contingency table is used for each class. Table 4 below represents the contingency table.

To assess the global performance over all the classes, the macro average  $F_1$  measurement was used in our experiments. It is found by averaging the per-class  $F_1$  values.

- Feature Selection time is the time needed to perform the feature selection method on the dataset.
- Classifier Training Time is the time needed by the classifier to learn using the training set after applying the feature selection method.
- Classifier Testing Time is the time needed to classify the testing documents.

**Table 4. The Contingency table to describe the components of the performance measurements.**

Actual Class	Predicted Class	
	Class = Yes	Class = No
Class = Yes	TP	FN
Class = No	FP	TN

**Table 5. Feature selection time (in seconds) for the compared CHI Variants on both Akhbar Alkhalij and Alwatan datasets.**

	Akhbar Alkhalij	Alwatan
Enhanced CHI	197	280
CHI + Cosine	201	370
CHI Baseline	196	277

The proposed method, which is composed of both Algorithm 1 and Algorithm 2, was compared with the regular Chi-square method, and another Chi-square variant, which represents Algorithm 1 only. Figure1 and Figure2 present the results on Akhbar Alkhalij and Alwatan datasets respectively. The experiment results showed that our method outperformed the two latter methods according to the effect on classification accuracy. Regarding the feature selection time, as displayed in Table 5, our improved method was slower than the regular Chi-square method in Alwatan dataset, and similar to it in Akhbar Alkhalij dataset. This could be due to the cosine similarity that increases the number of used dimensions to represent each document.

For example, if a training set of 500 documents was used, and the user defined number of reduced features  $R$  is 50, the original Chi-square method will represent each document using 50 dimensions, while our method will represent using 500 dimensions, which would increase the feature selection time.

According to the Classifier Training Time and Classifier Testing Time, no significant differences were detected.

As feature selection is a preprocessing step that is done once only in most of the applications, the relatively large running time of our method can be ignored. Furthermore, some solutions can be used, such as using parallel computing, to improve the running time. Therefore, our method was selected to compare it with the existing commonly used feature selection methods.

## 5. COMPARING OUR METHOD WITH EXISTING FEATURE SELECTION METHODS

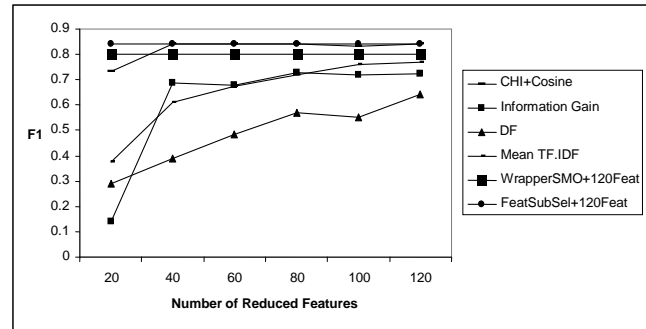
This section illustrates phase two of the experimental part, which compares our method from phase one with various existing feature selection methods. Mainly, our method was compared with Information Gain, DF, Mean TF.IDF, Wrapper Approach with SVM classifier, and Feature Subset Selection. The best First Search method was used in both the Wrapper approach and feature subset selection method. Weka 3.6.8[6] was used for the implementations of the IG, Wrapper, and Feature Subset Selection, and Visual Studio 6.0 was used to implement DF and Mean TF.IDF using C++.

Figure 3 and Figure 4 illustrate the classification accuracies of SVM in Akhbar Alkhalij and Alwatan respectively after using each feature selection method, while Table 6 illustrate the feature selection time for each method in both datasets.

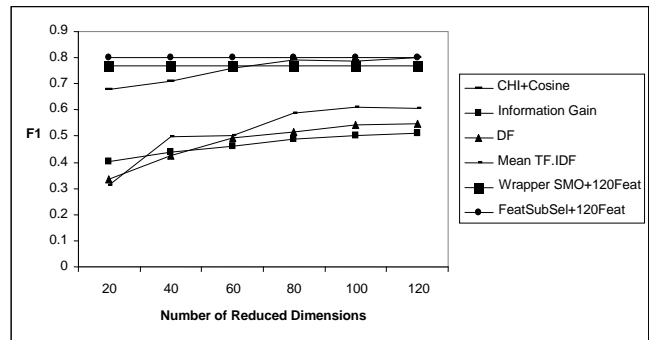
First, both Wrapper approach and Feature Subset Selection failed to work using the previously described system environment. This was due to the large memory that is needed by these two methods when the original number of attributes is large. Regarding the two datasets that were used, Akhbar Alkhalij has 20607 attributes originally, and Alwatan has 21962 attributes. Therefore, Wrapper approach was applied and Feature Subset Selection on a 120 reduced features selected by our method from phase one. For other methods, mainly IG, DF, and Mean TF.IDF, the original number of features was used. The experimental results showed that our improved Chi-square method outperformed the other well known methods in  $F_1$  measurement. Feature Subset Selection with 120 reduced features inherited the accuracy of our method, as the 120 features were previously selected by our method, but it failed to improve it further. Furthermore, the Wrapper approach method was the third in order and outperformed DF, Mean TF.IDF, and IG. Furthermore, with the increase in the number of dimensions, Mean TF.IDF showed better performance than DF and IG.

Regarding Classification time, no significant differences were noticed among the various compared methods in the two datasets, and it was around 2 seconds on average using the previous datasets. Similarly, there were no significant differences in the classifier training times, which were on average between 0.6 and 1 second.

Regarding the feature selection time, Table 6 presents the results. There is a clear difference between feature selection time of Wrapper approach and that of other methods. It is reasonable as wrapper methods select the features that improve the classification accuracy by spending more feature selection time. For example, Wrapper approach spent 9000 seconds in the feature selection process with 120 features, while other methods spent a maximum of 201 seconds on the original space of features. The second largest feature selection time belongs to our proposed method. This is due to the extensive amount of calculations needed for selecting each attribute in this Chi-square-based method.



**Fig 3: Comparing our improved method (CHI + Cosine) with the IG, DF, Mean TF.IDF, Wrapper, and Feature Subset according to their effect on the accuracy of Classifying Akhbar Alkhalij dataset. Results showed that our improved method outperformed other methods, while Feature Subset Selection failed to improve it further.**



**Fig 4: Comparing our improved method (CHI + Cosine) with the IG, DF, Mean TF.IDF, Wrapper, and Feature Subset according to their effect on the accuracy of Classifying Alwatan dataset. Results showed that our improved method outperformed the other methods. Feature Subset Selection failed to improve it further.**

**Table 6. Feature selection time (in seconds) for the compared methods on both akhbar alkhalij and watan datasets.**

	Feature Space	Akhbar Alkhalij	Watan
Information Gain	Original	0.5	0.62
CHI + Cosine	Original	201	370
DF	Original	1.5	1.6
Mean TF.IDF	Original	1	1.1
Wrapper(SMO)+Best Search	120 Reduced	9000	6600
Feature Subset Selection + Best Search	120 Reduced	3	2.5
Wrapper(SMO)+Best Search	Original	Failed to Work	
Feature Subset Selection + Best Search	Original	Failed to Work	

## 6. CONCLUSION

In this work, a new efficient feature selection method based on Chi-square Statistics was proposed. This method outperformed various existing feature selection methods according to its effect on classifying Arabic text items. Mainly, the proposed method outperformed Information Gain, DF, Chi-square, Mean TF.IDF, Wrapper Approach with SVM and Best Search, and Feature Subset Selection with Best Search. Furthermore, two publicly available, sufficient sized datasets were used to encourage future works to use them in order to make various works comparable. Future work could be done to optimize this method in order to improve its performance. Moreover, feature subset selection could be studied further in an attempt to enhance the output of this method.

## 7. REFERENCES

- [1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. . Akhbar Alkhalij and Alwatan-datasets, <https://sites.google.com/site/mouradabbas9/corpora>.
- [2] Al-Harbi, S., Al-Muhareb, A., Al-Thubaity, M., Khorsheed, S., and Al-Rajeh, A. 2008. Automatic Arabic Text Classification. JADT: 9es, Journées internationales d'Analyse statistique des Données Textuelles, 77-83.
- [3] Al-Saleem, S. 2010. Associative Classification to Categorize Arabic Data Sets. Internationsl Journal of ACM Jordan (Jan. 2010), 118-127.
- [4] El-Halees, A. 2007. Arabic Text Classification Using Maximum Entropy. The Islamic University Journal (Jan. 2007), 157-167.
- [5] El-Kourdi, M., Bensaid, A., and Rachidi, T. 2004. Automatic Arabic text categorization based on the Naive-Bayes Algorithm. Workshop on computational approaches to Arabic script-based languages.
- [6] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009. The Weka Data Mining Software: An Update, SIGKDD Explorations (Jun. 2009), 10-18.
- [7] Hall, M.A. 1999. Correlation-based Feature Subset Selection for Machine Learning. Thesis. Department of Computer Science, The University of Waikato.
- [8] Harrag, F., El-Qawasmah, E., and Al-Salman, A. S. 2010. Comparing Dimension Reduction Techniques for Arabic Text Classification using BPNN algorithm. First International Conference on Integrated Intelligent Computing, 6-11.
- [9] Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of 10th European Conference on Machine Learning, 137-142.
- [10] Khreisat, L. 2009. A Machine Learning Approach For Arabic Text Classification Using N-gram Frequency Statistics, Journal of Informatics (Sep. 2011) , 72-77.
- [11] Kohavi, R., John, G. H. 1997. Wrappers for Feature Subset Selection. Artificial Intelligence (Dec. 1997), 273-324.
- [12] Lam, S.L.Y. and Lee, D.L. 1999. Feature Reduction for Neural Network Based Text Categorization. In Proceedings of the 6<sup>th</sup> International Conference on Database Systems for Advanced Applications, 195-202.
- [13] Liu, H. and Setiono, R 1996. A Probabilistic Approach to Feature Selection - A Filter Solution. In Proceedings of the 13<sup>th</sup> conference on Machine Learning, 319-327.
- [14] Mesleh, A. 2007. Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System. Journal of Computer Science, (Jun. 2007), 430-435.
- [15] Quinlan, J.R. 1986. Induction of Decision Trees. Machine Learning (Jan. 1986), 81-106.
- [16] Yang, Y. and Pedersen, J.O. 1997. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14<sup>th</sup> International Conference on Machine Learning, 412-420.

## AUTHORS

**Dr. Bilal Hawashin** received his Ph.D in Computer Science, College of Engineering, from Wayne State University in 2011. Also, he worked in the Department of Computer Information Systems at Jordan University of Science and Technology from 2003-2007. His current research interests include Similarity Join, Text Mining, Information Retrieval, and Database Cleansing. Dr. Hawashin received his B.S. in Computer Science from The University of Jordan in 2002, and his M.S. in Computer Science from New York Institute of Technology in 2003.

**Dr. Ayman M Mansour** received his Ph.D. degree in Electrical Engineering from Wayne State University in 2012. Dr. Mansour received his M.Sc degree in Electrical Engineering from University of Jordan, Jordan, in 2006 and his B.Sc degree in Electrical and Electronics Engineering from University of Sharjah, UAE, in 2004. He graduated top of his class in both Bachelor and Master. His areas of research include communication systems, multi-agent systems, fuzzy systems, data mining and intelligent systems. He conducted several researches in his area of interest. Dr. Mansour is a member of IEEE, IEEE Honor Society (HKN), Tau Beta Pi Honor Society, Sigma Xi and Golden Key Honor Society.