The Impact of Feature Selection Methods for Classifying Arabic Textual Data

Mohammad Abu-Arqoub, Ghassan F. Issa, Wael M. Hadi

Abstract: ***Text classification is a vital process due to the large volume of electronic articles. One of the drawbacks of text classification is the high dimensionality of feature space. Scholars developed several algorithms to choose relevant features from article text such as Chi-square (x^2), Information Gain (IG), and Correlation (CFS). These algorithms have been investigated widely for English text, while studies for Arabic text are still limited. In this paper, we investigated four well-known algorithms: Support Vector Machines (SVMs), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree against benchmark Arabic textual datasets, called Saudi Press Agency (SPA) to evaluate the impact of feature selection methods. Using the WEKA tool, we have experimented the application of the four mentioned classification algorithms with and without feature selection algorithms. The results provided clear evidence that the three feature selection methods often improves classification accuracy by eliminating irrelevant features.***

Keywords **: *Feature selection; Text classification; Arabic textual data; Classical algorithms.***

I. INTRODUCTION

As the Internet has rapidly developed, so has the amount of news and information accessible online. This volume of information poses a particular challenge to manual analysis and processing. Therefore, hierarchical organization of documents has become more and more reliant on text classification. Text classification is a process that aims to organize texts in accordance with certain groupings.

Text classification or categorization is the process of assigning a document to one or more labels or categories [1]. The first is known as single-label classification. The latter is addressed in the literature as multi-label classification [2]. Text categorization is a supervised process since the set of labels is known a priori. It has many applications such as spam filtering, topic identification, and tailored news or advertisement delivery.

Text classification is not without difficulties. One barrier to use is the large number of available features. This vast array of choice takes time and thereby reduces text classification performance. One method used to mitigate this difficulty is feature selection, which only examines the most relevant aspects [3]. Proposed improvements to English text classification have included a number of feature selection methods. In comparison, there are far less feature selection methods on offer for Arabic language texts. In addition majority of research in this area in regards to Arabic texts is concentrated on algorithmic efficiency, and does not apply appropriate concern to the manner in which classification accuracy can facilitate feature selection [4]–[6].

Feature selection has become importance to several research which deal with data mining and machine learning communities, because it makes the algorithms to be fast, cost-effective, and more accurate. Feature selection is the task of eliminating irrelevant (useless) or redundant features from the original dataset. Feature selection, therefore, reduces the processing time of the learning algorithm, while enhancing the classification accuracy through the elimination of useless features which may hold noisy data [7]. With feature selection the understandability can be improved and cost of data handling becomes smaller [8].

Feature selection methods are split into three types; embedded selectors, wrappers, and filters. Embedded methods conduct feature selection during learning phase like artificial neural networks. Filters assess each feature independent from the algorithm model, rank the features after assessment and take the greater ones [9]. This assessment may be done using entropy for example [10]. Wrappers on the other hand selects a subset of the feature set, assesses the model's performance on this subset, and then another subset is assessed on the model. The subset for which the model has the maximum performance is selected. Thus wrappers are dependent on the selected model. In fact wrappers are more consistent since classification algorithm affects the accuracy, while the selection of the subset is an NP-hard problem [10]. Thus it needs substantial memory and processing time. Several heuristic procedures can be utilized for subset selection such as best first, genetic algorithm, random search, or greedy stepwise. Hence, the filters are more efficient when compared to wrappers, but they don't take into account that selecting the better features may appropriate to data mining techniques [11].

The aim of the study is to comparatively investigate feature selection methods and produce benchmark results for feature selection in published Arabic datasets. To investigate the impact of feature selection methods, four well-known, top-performing and scalable algorithms have been selected: K-Nearest Neighbors (KNN), Naive Bayes (NB), Support Vector Machines (SVMs),

Revised Manuscript Received on November 15, 2019
***** Correspondence Author
Ghassan F. Issa*, Faculty of Information Technology, University of Petra, Amman, Jordan.
Mohammad Abu-Arqoub*, Faculty of Information Technology, University of Petra, Amman, Jordan.
Wael M. Hadi, Faculty of Information Technology, University of Petra, Amman, Jordan.

and Decision Tree. We want to answer following questions with empirical results:

- Are the feature selection methods improve the classification accuracy of the algorithms positively?
- Which combination of classifier and feature selection method perform best across algorithms and feature selection methods?

The answers of the above questions and others can help data mining experts as well as computational scientists in improving the design process of text classification models. The rest of the paper is organized as follows: Section two for related works. Section three feature selection methods are presented. Section four gives the results and discussion of the experiments. Finally, section five concludes the paper.

II. RELATED WORK

Several feature selection have been used on text classification. The majority of feature selection study has been investigated extensively for English text classification and other applications such as phishing classification.

Thabtah and Abdelhamid (2016) [12] presented a new way of clustering different attributes or features for websites in solving the important problem of phishing categorization for the online community. The authors presented a new mathematical solution based on figuring out sharp lines that may differentiate influential and uninfluential features during the process of phishing classification. Different data mining methods have been applied to measure the success and failure of the phishing detection using a variety of features sets derived by three common features filtering techniques: Chi-square (X^2) [13], correlation features [14], and Information Gain [15]. Experimental analysis has been performed to evaluate the filtering methods against a large security data set download from UCI for phishing classification [16]. Data mining based on rule methods have been used to generate error rates and knowledge-based classifiers. The results showed that there are three primary groups of features that work well together in reducing the risk of phishing for classification. Moreover, the data mining algorithm using a small feature set of only 2 features have shown good predictive performance when compared with the 30 features data set and using the same data mining algorithms.

Prusa et al. (2015) [17] evaluated the impact of ten feature selection methods on the problem of tweets classification. Experimental results showed that the feature selection methods could have an excellent effect on the classification accuracy of data mining algorithms used for sentiment classification. All experimental results conducted using popular WEKA data mining software [18] and carried out using a 5-folds cross-validation method using Artificial Neural Network as Multilayer Perceptron (MLP), KNN, J48 Decision Tree, and Logistics Regression.

In [20], the authors studied the effect of five well-known feature selection methods: Correlation, X^2, Information Gain [15], GSS Coefficient [21], and Relief-F [22]. Moreover, this paper proposed an approach of combination of feature selection methods based on the average weight of the features. All experiments are employed using SVMs and NB algorithms to classify Arabic textual datasets. The experimental results indicated that the Information Gain outperformed all other feature selection methods. Also, the combination of multiple feature selection methods obtained the highest classification accuracy produced by individual methods.

Karabulut et al. (2012) [23] investigated the effect of using feature selection methods on classification accuracy of Naïve Bayes (NB), MLP, and J48 Decision Tree algorithms. These three known algorithms compared against 15 real life datasets from the UCI data repository [24] All experimental results conducted using popular WEKA data mining software and carried out using a 10-folds cross-validation method. In addition, the experiments conducted using six feature selection methods: Information Gain, Gain Ratio [25], Symmetrical Uncertainty [26], Relief-F, One-R [27], and X^2. The results indicated that the classification accuracy improved for NB and MLP algorithms.

The authors of [28] investigated the impact of feature selection on the problem of web spam detection. This study proposed a new feature selection method called Imperialist Competitive Algorithm and implemented it with Genetic algorithm.

Experimental results against WEBSPAM-UK2007 datasets [29] showed that dropping the number of features reductions the classification cost and improves the classification accuracy.

Duwairi in 2013 [3] studied the impact of weight by statistical feature selection methods on the Arabic text classification accuracy such as deviation, correlation, uncertainty, and x2. All tests conducted using NB algorithm and carried out using a 3-folds cross-validation approach against Arabic data contains of 4000 articles that belong to two classes.

The experimental results demonstrate that weight by correlation produces the highest classification accuracy when compared with other methods.

In [5], the authors proposed a hybrid approach, namely Binary PSO-KNN that select the best subset of the relevant features. Also, demonstrated the proposed approach with three known machine learning algorithms: Support Vector Machines (SVMs), NB, and J48 Decision Tree. All experimental results conducted using WEKA software. The results indicated that the proposed approach outperformed other methods against Arabic textual datasets and utilizing feature selection on text classification increases the classification accuracy of classifiers.

III. FEATURE SELECTION METHODS

A. *Chi-Square (x^2)*

According to Forman in 2003 [30], the chi-square (x2) statistical test is frequently employed to assess the divergence from the anticipated distribution based on the assumption of the null hypothesis, that is, that there is no relationship between the class value and the

measured feature. This test is widely noted to behave unpredictably when the anticipated count is extremely small, as it is in the context of text classification due to the scarcity of word features and to conceptualizations based upon various positive training examples. The X^2 feature selection method is computed by the following equation.

$$x^2(c,t) = \frac{N*((A*D)-(B*C))}{(A+C)*(B+C)*(A+B)*(C+D)} \quad (1)$$

Where:

A = The frequency of t and c occurrences, B = The frequency of t occurrences without c, C = The frequency of c without t, D = The frequency of non-occurrence of both c and t and N is the quantity of document.

B. *Correlation*

Suganya & Rajaram in 2012 [14] described correlation as an example of a widely accepted methodology for assessing and grading the pertinence of various characteristics by evaluating the relationships between one factor and another and between factors and classes. The Correlation Feature Selection (CFS) assesses the pertinence of various subsets of characteristics by application of Pearson's correlation equation based on the quantity of characteristics (k) and classes (C) as:

$$Merits = \frac{kr_{kc}}{\sqrt{k+(k-1)r_{kk}}} \quad (2)$$

Where, Merits is considered as the relevance of feature subset, r_{kc} is the average linear correlation coefficient between these features and classes and r_{kk} is the average linear correlation coefficient between different features.

C. *Information Gain*

Another noted technique is that of Information Gain (IG), which has found use as a learning tool in various contexts, such as in the preliminary stages of data processing for machine learning operations, including text classification [31], and in the construction of decision trees [32]. An example of the latter application is in evaluating the decrease in the level of uncertainty involved in using a characteristic to establish a specific class label or, equivalently, in judging how revealing a specific characteristic will be when using Equations 3-4 to derive a specific class label.

IG assesses how a specific feature divides the input training dataset samples with respect to the available set of class labels. Given an input training data D of P outcomes, for each feature X is possible to calculate its entropy as:

$$E(D) = -\sum_i p(x_i)\log_2 p(x_i) \quad (3)$$

Where p(xi) is the probability that x belongs to class c. The IG of feature X in the input data (D) is

$$Gain(D,X) = E(D) - \frac{|Dx|}{|D|} * E(Dx) \quad (4)$$

Where D is input training data, Dx is the subset of D for which X has value x, |Dx| = the subset data size having Dx from D, |D| =The input training data size.

IV. EXPERIMENTAL ANALYSIS

All experimental analysis have been evaluated using the WEKA software tool [18]. The 10-fold validation method has been utilized in the training SVMs, KNN, Decision Tree, and NB in order to produce the classifiers from the Arabic datasets. This study uses Saudi Press Agency datasets (SPA) which is collected by [33]. The SPA dataset consists of 1,526 text articles; each of them belongs to one of the six categories as shown in Table 1.

Table 1: Classes in SPA datasets

Class Name	# of Documents
Cultural	258
Economic	250
General	255
Political	250
Social	258
Sports	255
Total	**1526**

Finally, all experimental investigations have been run on a PC with 3 Ghz processor. The three feature selection methods (Information Gain, Correlation, X^2) are built within WEKA. The main purpose behind our selection of these three feature selection methods is they often produce significant features in multiple real applications such as text classification and detecting phishing websites.

The experimentation is divided into two parts as follows:
A) Evaluation of predictive power of classifiers using the entire dataset (SPA). In other words, each of the data mining algorithms (SVMs, KNN, Decision Tree, and NB) is applied using all the features of the entire dataset.
B) Effect of applying feature selection algorithms on predictive power of classifiers: In this part of the experiment, feature selection algorithms are applied to the SPA dataset first (IG, X^2 and CSF), then the classification algorithms are tested.

Table 2: Results of algorithms on the complete SPA data

Classifier	Accuracy	Recall	Precision	F1
NB	71.6252	0.716	0.719	0.709
KNN	42.3984	0.424	0.605	0.438
SVMs	73.0668	0.731	0.732	0.730
Decision Tree	56.7497	0.567	0.566	0.566

After analyzing Table 2, we found that the SVMs algorithm reaches the best results for all measures (Accuracy, Recall, Precision, and F1). The second best algorithm is NB, and the KNN obtained the worst results for all measures. In particular,

the SVMs produced 1.44%, 16.32%, and 30.67% higher accuracies than NB, Decision Tree, and KNN algorithms, respectively. Figures 1, 2, 3, 4, 5, and 6 depict the classification accuracies results generated by all considered algorithms using three different feature selection methods, when the number of features varying from 500 to 1000, respectively. Figure 1 shows that the NB algorithm outperformed all other algorithms using three different feature selection, when the number of features is 500. In particular, the NB algorithm produces 71.30%, 71.82%, and 71.95% classification accuracies using CFS, IG, and X2, respectively. In general, the results produced by NB, Decision tree, and KNN using 500 features are higher than their original results on the complete dataset.

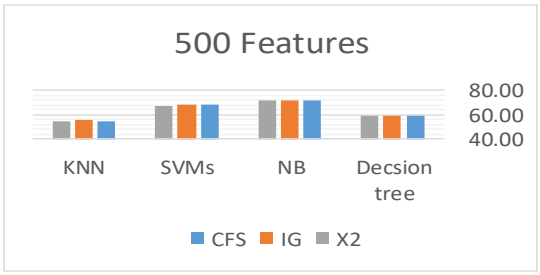

Figure 1: The results of algorithms on 500 Feature

Figure 2 shows that the NB algorithm outperformed all other algorithms using three different feature selection, when the number of features is 600. In particular, the NB algorithm produces 71.17%, 70.71%, and 71.10% classification accuracies using CFS, IG, and X^2, respectively. In general, the results produced by all algorithms using 600 features are comparable with their original results on the complete dataset.

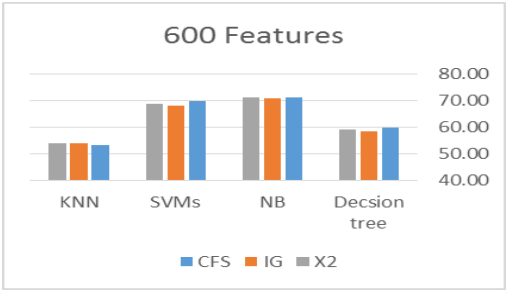

Figure 2: The results of algorithms on 600 Features

Figure 3 displays that the NB algorithm outperformed all other algorithms using three different feature selection, when the number of features is 700. In particular, the NB algorithm produces 71.43%, 70.97%, and 71.23% classification accuracies using CFS, IG, and X^2, respectively. In general, the results produced by all algorithms using 700 features are comparable with their original results on the complete dataset. Also, all classifiers produce results higher than 50%.

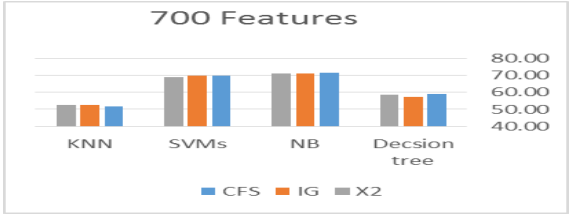

Figure 3: The results of algorithms on 700 Features

Figure 4 displays that the NB algorithm outperformed all other algorithms using three different feature selection, when the number of features is 700. In particular, the NB algorithm produces 71.63%, 71.43%, and 71.17% classification accuracies using CFS, IG, and X^2, respectively. In addition, the SVMs algorithm produces results greater than 70%. In general, the results produced by all algorithms using 800 features are comparable with their original results on the complete dataset. Also, all classifiers produce results higher than 50%.

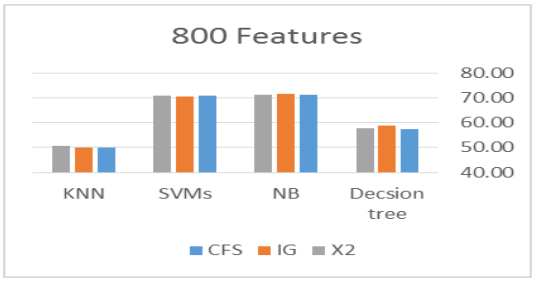

Figure 4: The results of algorithms on 800 Features

After analyzing figure 5, we found that the NB algorithm and SVMs achieved the best results using three deferent feature selection, when the number of features is 900. In particular, the NB algorithm and SVMs produce classification accuracies above 70% using CFS, IG, and X^2.

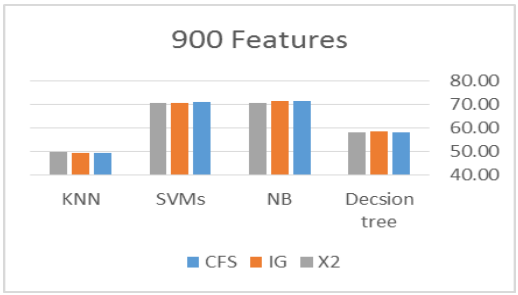

Figure 5: The results of algorithms on 900 Features

In addition, figure 6 depicts that the NB algorithm and SVMs achieved the best results using three different feature selection, when the number of features is 1000. In particular, the NB algorithm and SVMs produce classification accuracies above 70% using CFS, IG, and X^2

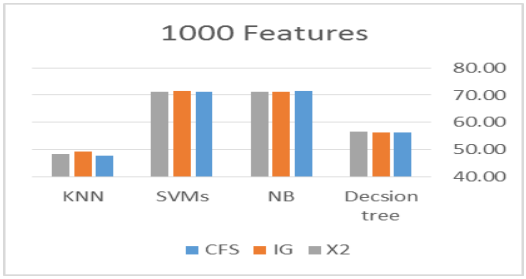

Figure 6: The results of algorithms on 1000 Features

From our experimental results, we conclude that the feature selection often increased classification accuracy by removing irrelevant features. In particular, the NB using IG and X^2 on 500 features produce 0.002

and 0.0033 higher accuracies than the original complete dataset. This reduce the run time and memory usage. In addition, the feature selection methods improve the performance of NB algorithm because its learning strategy depends on the probability of features

V. CONCLUSION

Text classification is one of the popular problems in information retrieval and machine learning fields. Given a large number of articles in a dataset (training data) where each article is belonged with its corresponding class. Classification procedure involves building a model from classified articles (learning phase), in order to predict previously unseen articles as accurately as possible (testing phase). One of the main drawbaks of text classification is the high dimensionality of feature space, which requires more time and more memory usage. Feature selection is an important issue in classification, because it may have a considerable impact on classification accuracy of the algorithm. It reduces the number of features of the dataset, so the memory usage and run time reduce; the dataset becomes easier and more understandable to investigate on.

The experimental results against SPA datasets provided evidence that the three feature selection methods often increased classification accuracy by removing irrelevant features. In general, the results produced by all algorithms using three feature selection methods are comparable with their original results on the complete dataset. Also, all classifiers produce results higher than 50%. In addition, the NB and SVMs worked well when the number of features reduced by all feature selection methods.

REFERENCES

1. W. Hadi, "ECAR : A New Enhanced Class Association Rule," Adv. Comput. Sci. Technol., vol. 8, no. 1, pp. 43–52, 2015.
2. F. Thabtah, W. Hadi, N. Abdelhamid, and A. Issa, "PREDICTION PHASE IN ASSOCIATIVE CLASSIFICATION MINING," Int. J. Softw. Eng. Knowl. Eng., vol. 21, no. 6, pp. 855–876, Sep. 2011.
3. R. M. Duwairi, "Statistical Feature Selection Techniques for Arabic Text Categorization," in The Fourth International Conference on Information and Communication Systems, 2013, pp. 1–4.
4. B. Al-Salemi and M. Ab Aziz, "Statistical Bayesian Learning for Automatic Arabic Text Categorization," J. Comput. Sci., vol. 7, no. 1, pp. 39–45, Jan. 2011.
5. H. K. Chantar and D. W. Corne, "Feature subset selection for Arabic document categorization using BPSO-KNN," Proc. 2011 3rd World Congr. Nat. Biol. Inspired Comput. NaBIC 2011, pp. 546–551, 2011.
6. B. Hawashin, A. Mansour, and S. Aljawarneh, "An Efficient Feature Selection Method for Arabic Text Classification," Int. J. Comput. Appl., vol. 83, no. 17, pp. 1–6, Dec. 2013.
7. S. Doraisamy, S. Golzari, N. M. Norowi, N. B. Sulaiman, and N. I. Udzir, "A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music," ISMIR 2008 Proc. 9th Int. Conf. Music Inf. Retr., pp. 331–336, 2008.
8. A. Arauzo-Azofra, J. L. Aznarte, and J. M. Benítez, "Empirical study of feature selection methods based on individual feature evaluation for classification problems," Expert Syst. Appl., vol. 38, no. 7, pp. 8170–8177, Jul. 2011.
9. P. Yildirim, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease," Int. J. Mach. Learn. Comput., vol. 5, no. 4, pp. 258–263, Aug. 2015.
10. U. Stańczyk, "Feature Evaluation by Filter, Wrapper, and Embedded Approaches," 2015, pp. 29–44.
11. R. Wald, T. Khoshgoftaar, and A. Napolitano, "Comparison of Stability for Different Families of Filter-Based and Wrapper-Based Feature Selection," in 2013 12th International Conference on Machine Learning and Applications, 2013, pp. 457–464.
12. F. Thabtah and N. Abdelhamid, "Deriving Correlated Sets of Website Features for Phishing Detection: A Computational Intelligence Approach," J. Inf. Knowl. Manag., vol. 15, no. 4, p. 1650042, Dec. 2016.
13. Huan Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, 1995, pp. 388–391.
14. R. Suganya and S. Rajaram, "Content Based Image Retrieval of Ultrasound Liver Diseases Based on Hybrid Approach," Am. J. Appl. Sci., vol. 9, no. 6, pp. 938–945, Jun. 2012.
15. T. M. Cover and J. A. Thomas, Elements of Information Theory. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2005.
16. R. Mohammad, F. Thabtah, and L. McCluskey, "Phishing Websites Dataset," 2015. Online.. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites. [Accessed: 01-Nov-2015].
17. J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, "Impact of Feature Selection Techniques for Tweet Sentiment Classification," in The Twenty-Eighth International Flairs Conference, 2015, pp. 299–304.
18. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," ACM SIGKDD Explor. Newsl., vol. 11, no. 1, p. 10, 2009.
19. I. Qabajeh and F. Thabtah, "An Experimental Study for Assessing Email Classification Attributes Using Feature Selection Methods," in 2014 3rd International Conference on Advanced Computer Science Applications and Technologies, 2014, pp. 125–132.
20. A. Adel, N. Omar, and A. Al-Shabi, "a Comparative Study of Combined Feature Selection Methods for Arabic Text Classification," J. Comput. Sci., vol. 10, no. 11, pp. 2232–2239, 2014.
21. G. Uchyigit and M. Y. Ma, Personalization Techniques and Recommender Systems, vol. 70. WORLD SCIENTIFIC, 2008.
22. K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in Machine Learning Proceedings 1992, Elsevier, 1992, pp. 249–256.
23. E. M. Karabulut, S. A. Özel, and T. İbrikçi, "A comparative study on the effect of feature selection on classification accuracy," Procedia Technol., vol. 1, pp. 323–327, 2012.
24. M. Lichman, "UCI Machine Learning Repository," 2013.
25. R. Hierons, "Machine learning. Tom M. Mitchell. Published by McGraw-Hill, Maidenhead, U.K., International Student Edition, 1997. ISBN: 0-07-115467-1, 414 pages. Price: U.K. £22.99, soft cover.," Softw. Testing, Verif. Reliab., vol. 9, no. 3, pp. 191–193, Sep. 1999.
26. M. C. Seiler and F. A. Seiler, "Numerical Recipes in C: The Art of Scientific Computing," Risk Anal., vol. 9, no. 3, pp. 415–416, Sep. 1989.
27. R. C. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets," Mach. Learn., vol. 11, no. 1, pp. 63–90, 1993.
28. J. Karimpour, A. A. Noroozi, and A. Abadi, "The Impact of Feature Selection on Web Spam Detection," Int. J. Intell. Syst. Appl., vol. 4, no. 9, p. 61, 2012.
29. C. Castillo et al., "A reference collection for web spam," ACM SIGIR Forum, vol. 40, no. 2, pp. 11–24, Dec. 2006.
30. G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Mach. Learn. Res., vol. 3, pp. 1289–1305, 2003.
31. A. K. Uysal, "An improved global feature selection scheme for text classification," Expert Syst. Appl., vol. 43, pp. 82–92, Jan. 2016.
32. J. Quinlan, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," Mach. Learn., vol. 16, no. 3, pp. 235–240, Sep. 1993.
33. S. Al-Harbi, a Almuhareb, and a Al-Thubaity, "Automatic Arabic text classification," 9es Journées Int. Anal. Stat. des Données Textuelles, pp. 77–84, 2008.

AUTHORS PROFILE

Mohammed Abu-Arqoub is the Chair of Computer Science at the University of Petra, Amman, Jordan. He received his Masters and Ph.D. Degrees from Saint Petersburg State Electrotechnical University, Russia. His research interests includes Simulation, Machine Learning, and e-Learning.

Ghassan Issa is a professor of Computer Science at the University of Petra in Amman, Jordan. He received his BET from Toledo University, Ohio, BSEE from Trine University, Indiana, M.S and PhD. from Old Dominion University, Virginia. His research interests include the areas in Artificial Intelligence, Machine learning, and e-Learning System.

.

Wael Hadi Author is currently the Chair of Computer Information Systems at the University of Petra. He Holds a Master and Ph.D degrees from the Arab Academy for Banking and Financial Sciences. His research interest in Data Mining, Machine Learning, and Big Data.