Journal of Theoretical and Applied Information Technology
31st March 2018. Vol.96. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPARING TWO FEATURE SELECTIONS METHODS (INFORMATION GAIN AND GAIN RATIO) ON THREE DIFFERENT CLASSIFICATION ALGORITHMS USING ARABIC DATASET.

ADEL HAMDAN MOHAMMAD

Computer Science Department
The World Islamic Sciences and Education University
adel.hamdan@wise.edu.jo
adel_hamdan@yahoo.com

ABSTRACT

Text classification is a very important topic. Nowadays there are a huge amount of available data. This data need to be classified and categorized. Number of researches applied on Arabic dataset still need more investigation. There are several available methods and techniques to classify data. Also, there are several feature selection methods used in pre-processing stage. Experiments in this paper done using two feature selections (information gain and gain ratio) on three classification methods (Naïve Bayes, Decision tree C4.5 and Support Vector Machine SVM. Experiments done using Arabic dataset. Results shows that feature selection done in pre-processing stage is a key success factor for success. Also results demonstrate that gain ratio is a little bit better than information gain and SVM is approximately very closed to Naïve Bayes and both of SVM and Naïve Bayes is more accurate than C4.5.

Keyword: *Text classification, information retrieval, Naïve Bayes, Support Vector Machine, Decision tree, C4.5.*

1. INTRODUCTION

No doubt that text categorization (TC) is a very important topic. Nowadays there are a huge amount of available data. This data required an emergent text categorization or classification techniques. In addition, the massive amount of data available online makes the process of classification and categorization an active research area for all researchers [1]. The aim of Text classification researchers is categorizing document into a group of predefined documents based on its content. [2] There are Several researchers talk about text classification for both English and Arabic languages. Also, no doubt that the number of researches which talk about English text classification is more than the number of researches which talk about Arabic text classification. [3,4,5]

Text classification for Arabic language is a hot topic for Arabic researchers since Arabic language is the main language in more than 25 countries. And actually a few number of researchers talk about Arabic text categorization. So Arabic text categorization still needs more research and investigation. One of the key success for text classification is feature selection which can be considered as the main success factor in dataset preprocessing stage in text categorization process.

1.1 Feature selection

Feature selection is a very significant process in text categorization. The objective of Feature selection is to select subcategory of extremely discriminant features. Also, the focus of feature selection is selecting subset of terms in training data and using it as features in the next stages of the process. No doubt that feature selection is expected to classify samples that belong to diverse classes and create features is the key success factor for text classification. Besides that, feature selection helps researchers to understand and reduce diversity of the dataset. [6].

Feature selection is a pre-processing technique used in text classification. Feature selection aims at removing irrelevant attributes which can enhance and increase accuracy. This means that feature selection is a significant factor for the success or failure of the whole process. Based on that, Selecting or discarding attributes is a key point for success or failure in feature selection. [7]

Several researchers talk about feature selection methods and techniques. Also, some researchers categorize feature selection techniques into several

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

methods and techniques depending on some criteria, such as filter methods, wrapper methods, embedded methods and hybrid methods. [6, 8, 9, 10]. Most of researchers are using English dataset to apply feature selection. But, there are several researchers talks about features selection based on Arabic dataset, but no doubt that the number of researches based on English dataset is greater than Arabic dataset.

Feature selection techniques involve information gain, Gain ratio, Symmetric uncertainty, correlation based feature selection, Markov blanket filter, Fast correlation based feature selection and minimum redundancy maximum relevance. [7, 11, 12, 13]

1.2 Information Gain

Information gain is also known as mutual Information which can be considered as a measure of the mutual dependence between two variables. Information gain used to obtain the amount of useful information obtained from one random variable, through using another variable. Information gain, in other words, is a symmetrical measure of dependency.

Information gain in a simple form can be considered as the amount of information gained about X after observing Y is equal to the information gained about Y after observing X. information gain tells us the importance of given attribute of the feature vectors and it used to decide the ordering of the attributes in decision tree nodes. [14] One of the main Weaknesses of information gain method is that its preference for the features with the highest values even when these features less informative.

Information Gain formula defined as: [9, 10,11]

IG (X;Y)= H(X)-H(X|Y)

1.3 Gain Ratio

Information gain ratio can be defined as a ratio of information gain to the intrinsic information. Beside that, Gain ratio is an adjustment of information gain that decreases its bias when the number of branching features is high. Gain ratio takes in consideration the size and the number of branches to pick a feature.

Gain Ratio (GR) formula defined as:

GR= IG/H(X).

Gain ratio values are always between [0,1]. If the value of GR = 1, this value tells us that the knowledge of X completely leads to Y, and if the value of GR=0, this means that there is no relation between X and Y. [14]

1.4Naïve Bayes

Naïve Bayes(NB) is an effective method of classification. Naïve Bayes are, relatively, simple probabilistic and great method used in text classification. Naïve Bayes formulas are defined as the following:

P (class | Document) = P (class). P (Document | class) / P (Document)

P (class | document): The probability that a given document D belongs to a given class C.

P (document): The probability of a document.

P (class): The probability of a class (or category).

P (document | class) represents the probability of document given class. [15,16,17,18].

NB technique (See figure1) depend on representing each class with a probabilistic summary then it finds the most expected class for each example it is asked to classify. One of the most drawbacks of NB is that if two or more attributes in the data set are extremely correlated, their weight will be high in the final decision. [15,16,19]

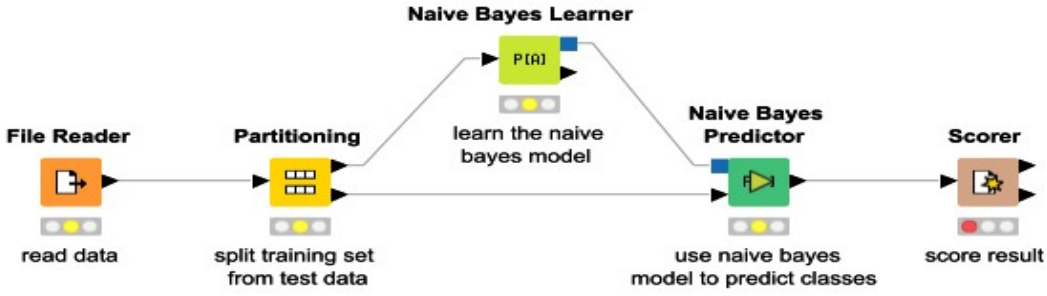

Figure 1 (Naïve Bayes Classifier) [20]

Journal of Theoretical and Applied Information Technology
31st March 2018. Vol.96. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1.5 C4.5 (Decision Tree)

C4.5 is one of the most important statistical classifier methods used in text classification (See Figure 1). C4.5 used to generate a decision tree. C4.5 becomes quite widespread after ranking #1 in the Top 10 Algorithms used in Data Mining pre-eminent paper published by Springer[21]. C4.5 used features which builds during training in decision and classification. One main merit for C4.5 is that if there is a set of records and each record has the same features and structure and consist of several numbers of attributes then only one of these attributes will represent the category of the document. [22,23,24]

1. C4.5: classification
2. K-Means: clustering
3. SVM: classification
4. Apriori: association analysis
5. EM: statistical learning
6. PageRank: link mining
7. AdaBoost: bagging and boosting
8. kNN: classification
9. Naive Bayes: classification
10. CART: classification

Figure 2: Top 10 Data Mining Algorithm

1.6 Support Vector Machine

Support vector machines (SVM) or support vector networks is one the most successful methods used in text classification. SVM is a machine learning technique. SVMs is considered as a supervised learning model used for analyses and classification. In SVM the training algorithm used to construct a new model which used to assign new documents into a set of predefined groups. SVM can be found in two forms (See figure 3, 4) Linear classifier and Non-Linear classifier. Linear classifier based on linear function [25, 26]. In some cases, data is not linearly divided. But with some modifications linear SVM can be generalized to adapt Non-linear problems. [27, 28]

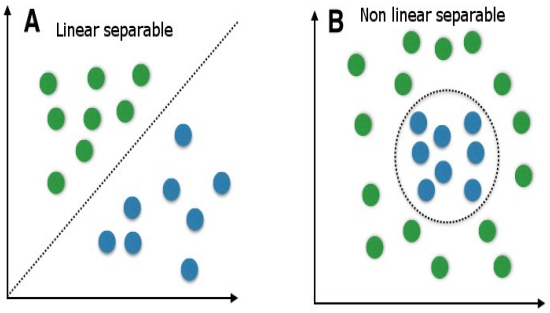

Figure 3: Linear And Non-Linear Separable [29]

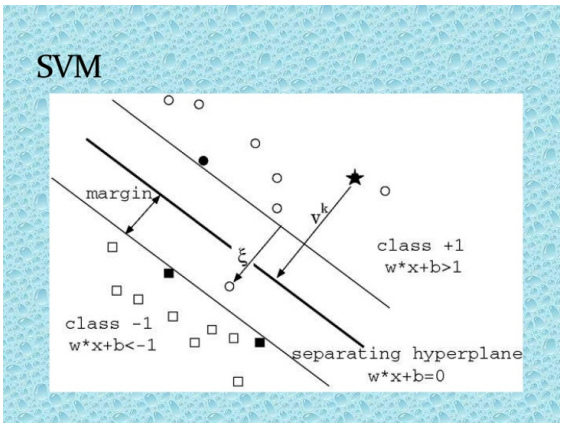

Figure 4: SVM

2. RELATED STUDIES

Ghazi Raho, Riyad Al-Shalabi [30] in this research authors demonstrate how feature selection is an effective and major factor for the success of text classification. they compare the performance between difference classifiers in different situation using feature selection with stemming. Arabic dataset is used to evaluate their experiments. Evaluation in their research done using decision tree, K-nearest neighbor Naive Bayesian and naïve Bayes Multinomial. They used precision, recall, F-Measure and accuracy to evaluate their results. Their results show that the accuracy for decision tree, Naïve Bayesian method and Naïve Bayes multinomial better than k-nearest neighbors.

Mary Walowe [31] in his research showed the importance and the need to make researcher aware of feature selection methods. He observed that obtaining an optimal subset of related and non-related features is not an easy task. He mentioned that the existing methods depend on univariate ranking that does not consider relations and interactions between selected subsets and the remining one.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Adel Hamdan , Omar Al-Momani, Tariq Alwada'n [32,33] in this research authors applied several methods for classification such as k-nearest neighbor, C4.5 decision tree, Rocchio classifier, Support vector machine, Naive Bayes and neural network. Authors experiments applied on Arabic dataset. Also, authors experiments show that Rocchio and K-NN are better than C4.5. In addition, experiments displayed that Support vector machine gives a good result. Besides that, authors experiments demonstrate that MPL-NN will give best results with 600 input layers.

Ashraf Odeh, Aymen Abu-Errub [34] in this research authors propose a good way for Arabic text classification. their method based on using vector evaluation. Also, they apply their experiment on Arabic dataset. The experiments determine the key words of the tested document by weighting each word, and then they compared these key words with the key words of the testing corpus categorizes. authors say that this algorithm prepare the document to ensure a better selection.

Bilal Hawashin, Ayman M Mansour [35], they propose an efficient feature selection method. Experiments in this research done using Arabic dataset. They propose a new efficient feature selection method. Also, they says that their method outperformed several feature selection methods.

Arabic Language and Dataset

Arabic language is the formal language for 25 countries. Also, Arabic language spoken by 250 million. Arabic language consists from 28 letters plus hamaz(ء) which considered a letter by some Arabic linguistics. Majority of Arabic words has its root and representing words to their root is very important to reduce number of words. [32, 36, 37]

Dataset used in this research is the same dataset used in two related research for the author. Data set collected from several web site such as Al-Jazeera news web site, Saudi Press Agency (http://www.spa.gov.sa/index.php) and finally Al-Hayat web site (http://www.alhayat.com/). Dataset used in this research consists of 1600 Arabic documents which belong to dissimilar classes (see table 1).

Table 1: Dataset

Category	Total Number of documents	Number of documents used for training	Number of documents used in testing.
Computer	200	130	70
Economics	200	130	70
Education	200	130	70
Law	200	130	70
Medicine	200	130	70
Politics	200	130	70
Religion	200	130	70
Sports	200	130	70
Total	1600	1040	560

3. EXPERIMENTS AND RESULTS

Author in this research use three evaluation measures. These measures are (recall, Precision, and F1). precision (called positive predictive value) is the fraction of relevant instances among the retrieved instances, recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. To demonstrate precision, recall and F1 measures formula in more detail (see table 2)

Table 2: Document Possible Sets Based On A Query

Iteration	Relevant	Irrelevant
Document retrieved	a	b
Document not retrieved	c	d

F1= (2* Precision * Recall) / (Recall + Precision)
Precision = a / (a + b)
Recall = a / (a + c)

In this research, the number of feature selected is applied from 200 to 1000 but author notice that the best results from 500, 600 and 700, So the author select 600 features in this study and all results displayed below are based on 600 features.

Table 3, 4 and 5 shows results for Naïve, C4.5 and SVM text classification using information gain as feature selection technique. Mentioned tables demonstrate that precision, recall and F1 measure. After analyzing results author find that SVM and Naïve gives the most promising results.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Naïve Bayes (Precision, Recall, F1) (Info gain)

Naive/ Info Gain			
Category (Naïve)	Precision	Recall	F1
Computer	0.871	0.883	0.876
Economics	0.865	0.881	0.872
Education	0.795	0.798	0.796
Law	0.695	0.923	0.792
Medicine	0.821	0.874	0.846
Politics	0.908	0.798	0.849
Religion	0.789	0.699	0.7412
Sports	0.912	0.693	0.7875
Average	**0.832**	**0.8186**	**0.8205**

Table 4: C4.5 (Precision, Recall, F1) (Info gain)

C4.5/ Info Gain			
Category (C4.5)	Precision	Recall	F1
Computer	0.698	0.852	0.767
Economics	0.621	0.874	0.726
Education	0.598	0.862	0.706
Law	0.712	0.798	0.752
Medicine	0.623	0.741	0.676
Politics	0.698	0.752	0.723
Religion	0.499	0.698	0.581
Sports	0.485	0.808	0.606
Average	**0.616**	**0.798**	**0.692**

Table 5: SVM (Precision, Recall, F1) (Info gain)

SVM / Info Gain			
Category (SVM)	Precision	Recall	F1
Computer	0.874	0.741	0.802
Economics	0.821	0.753	0.7855
Education	0.789	0.874	0.829
Law	0.855	0.856	0.855
Medicine	0.789	0.891	0.836
Politics	0.881	0.851	0.865
Religion	0.908	0.791	0.845
Sports	0.912	0.897	0.904
Average	**0.8536**	**0.831**	**0.840**

Experiments using information gain demonstrate that naïve Bayes method when we apply categorization on sport category shows the best results with precision 0.912. and the worst result arise when using law category with precision 0.695. using C4.5, as a classification method, the best result is shown when we applied categorization on law category with precision 0.712 and the worst result arise when using sports with precision 0.485. besides that, experiments using SVM demonstrate that sport category shows best results (0.912) and the worst with education category (0.789).

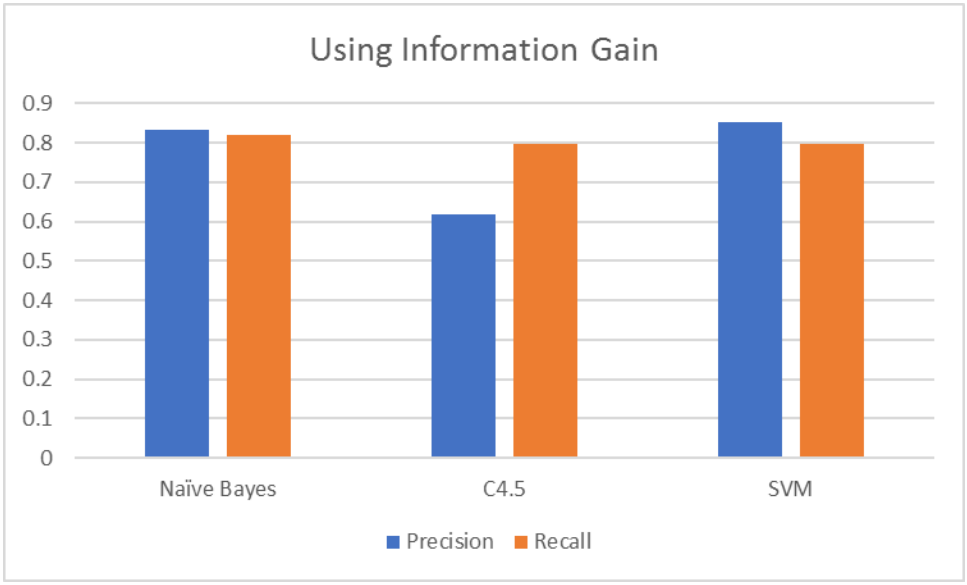

Figure 5: Precision, recall using Information Gain

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 5 shows precision and recall for three classification methods used in this research with information gain as feature selection method. Naïve Bayes and SVM give better results than C4.5 in our experiments.

Table 6, 7 and 8 shows results for Naïve, C4.5 and SVM text classification using Gain ratio as feature selection technique. Besides that, tables demonstrate results for all category used in our experiments. After analyzing results author find that SVM gives the best results.

Table 6: C4.5 (Precision, Recall, F1) (Gain Ratio)

Naive/ Gain Ratio			
Category (Naïve)	Precision	Recall	F1
Computer	0.901	0.883	0.891
Economics	0.902	0.881	0.891
Education	0.897	0.798	0.844
Law	0.879	0.923	0.900
Medicine	0.885	0.901	0.892
Politics	0.889	0.963	0.924
Religion	0.799	0.951	0.868
Sports	0.912	0.898	0.904
Average	**0.883**	**0.899**	**0.889**

Table 7: SVM (Precision, Recall, F1) (Gain Ratio)

C4.5/ Gain Ratio			
Category (C4.5)	Precision	Recall	F1
Computer	0.712	0.691	0.701
Economics	0.695	0.874	0.774
Education	0.621	0.891	0.731
Law	0.821	0.809	0.814
Medicine	0.796	0.819	0.807
Politics	0.741	0.798	0.768
Religion	0.598	0.589	0.593
Sports	0.741	0.719	0.729
Average	**0.7156**	**0.773**	**0.740**

Experiments using gain ratio demonstrate that naïve Bayes method when we apply categorization on sport category shows the best results with precision 0.912. and the worst result arise when using religion category with precision 0.799. using C4.5, as a classification method, the best result is shown when we applied categorization on law category with precision 0.821 and the worst result arise when using Religion with precision 0.598. besides that, experiments using SVM demonstrate that sport and computer category shows best results (0.912) and the worst with politics category (0.881).

Table 8: SVM (Precision, Recall, F1) (Gain Ratio)

SVM / Gain Ratio			
Category (SVM)	Precision	Recall	F1
Computer	0.912	0.881	0.896
Economics	0.901	0.887	0.893
Education	0.897	0.963	0.928
Law	0.891	0.871	0.880
Medicine	0.889	0.891	0.889
Politics	0.881	0.928	0.903
Religion	0.908	0.921	0.914
Sports	0.912	0.897	0.904
Average	**0.898**	**0.904**	**0.901**

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

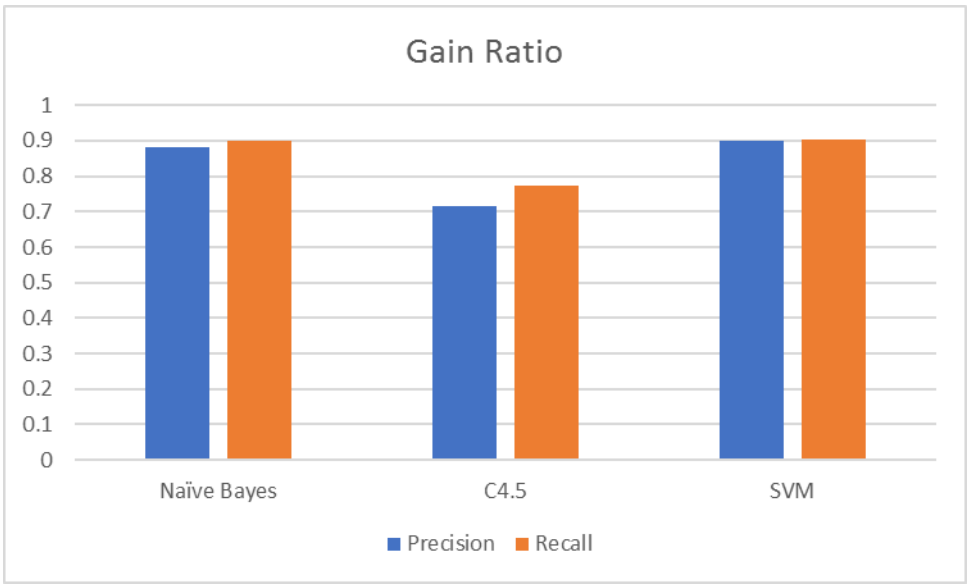

Figure 6: Precision, Recall Using Gain Ratio

Figure 6 shows precision and recall for three classification methods used in this research with gain ratio as feature selection method. Results demonstrate that SVM is the Best one in our experiments with a little bit different from Naïve Bayes and C4.5 gives an average precision 0.715 and average recall 0.733.

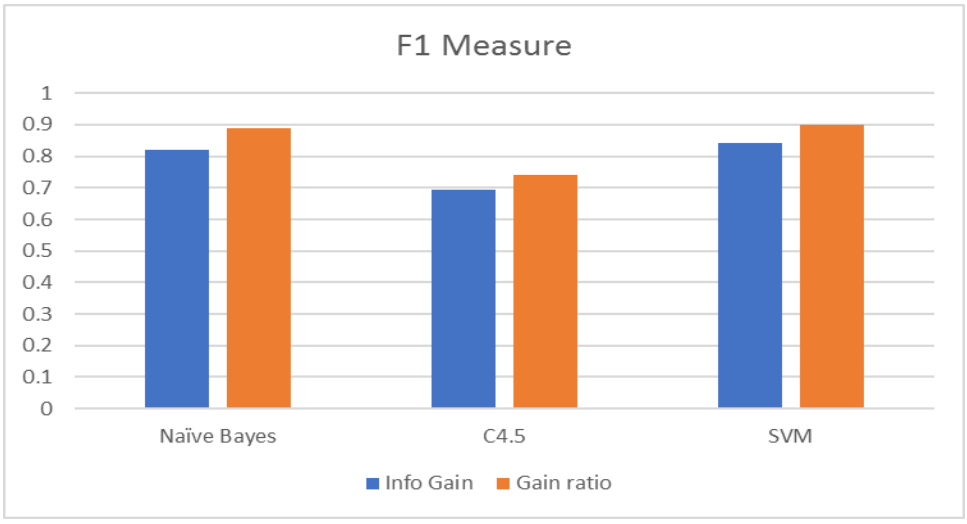

Figure 7: F1 Measure Using Information Gain And Gain Ratio

Figure 7 displays F1 measures using both information gain and gain ration with text classification process. Figures demonstrate that gain ratio results are little bit better than information gain.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. CONCLUSION

In this research author talks about the problem of text categorization. A lot of researchers talk about English text classification, but related to Arabic language still the numbers of research need more investigation and examination. Author of this paper used in-house developed Arabic Dataset. In addition, in this paper author investigate three main classification methods. Methods used are Naïve Bayes, C4.5 and SVM. Author used two feature selections in preparing dataset which are information gain and gain ratio. The results demonstrate that when using information gain as feature selection the SVM and Naïve Bayes approximately gives same results and the results of SVM and Naïve Bayes is better than C4.5. Finally, the results demonstrated that SVM giving a little bit better results than naïve Bayes and C4.5 with gain ratio as feature selection. Future work for author are considering more feature selection methods and technique for Arabic dataset.

REFERENCE

[1] Motaz K Saad, Wesam Ashour, (2010)Arabic Text Classification Using Decision Trees(2010) proceedings of the 12th international workshop on computer science and information technologies CSIT'2010, Moscow – Saint-Petersburg, Russia, 2010

[2] Mofleh Al-diabat ,(2012) Arabic Text Categorization Using Classification Rule Mining, Applied athematical Sciences, Vol. 6, 2012, no. 81, 4033 – 4046

[3] Adel Hamdan,, Raed Abu-Zitar (2011) Spam Detection Using Assisted Artificial immune System, Volume: 25, Issue: 8(2011) pp. 1275-1295, International Journal of Pattern Recognition and Artificial Intelligence

[4] Raed Abu-Zitar ,Adel Hamdan (2011) , Application of Genetic Optimized Artificial Immune System and Neural Networks in Spam Detection ,Applied Soft Computing, Volume 11, Issue 4, June 2011, Pages 3827-3845 ,Elsevier, 2011.

[5] Rasha Elhassan, Mahmoud Ahmed (2015), Arabic Text Classification Review International Journal of Computer Science and Software Engineering (IJCSSE), Volume 4, Issue 1, January 2015

[6] Girish Chandrashekar, Ferat Sahin, (2014). "A survey on feature selection methods". Computers and Electrical Engineering.

[7] Mary Walowe Mwadulo, A Review on Feature Selection Methods For Classification Tasks, International Journal of Computer Applications Technology and Research Volume 5– Issue 6, 395 - 402, 2016, ISSN:-2319–8656

[8] Yvan Saeys, Inak Inza, Pedro Larranaga, (2007). "A review of Feature Selection techniques in bioinformatics". Bioinformatics, Oxford University press.

[9] Feng Tan, Xuezheng Fu, Yanqing Zhang, Anu G. Bourgeois, (2008). "A genetic algorithm-based method for feature subset selection". Soft Comput.

[10] Muhammad Shakil Pervez, Dewan Md. Farid ,(2015). "Literature Review of Feature Selection for mining Tasks".International Journal of Computer Application, Vol 116, No. 21.

[11] Zakaria elberrichi and karima abidi, Arabic text categorization : a comparative study of different Representative Modes, the International Arab Journal of Information Technology , Vol 9, No5, September 2012.

[12] Zena M. Hira, Duncan F. Gillies, (2015) "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data". Hindawi Publishing Corporation Advances in Bioinformatics

[13] B.Azhagusundari, Antony Selvadoss Thanamani, (2013). "Feature Selection based on Information Gain".International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol 2, issue 2.

[14] Hall, M. A. & Smith, L. A. (1998). Practical feature subset selection for machine learning. In C. McDonald (Ed.), Computer Science 98 Proceedings of the 21st Australasian Computer Science Conference ACSC 98, Perth.

[15] Elkan, C. 1997. Boosting and Naive Bayesian Learning. Technical Report No. CS97-557, Department of Computer Science and Engineering, University of California, San Diego

[16] Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955

[17] Saleh Alsaleem, Automated Arabic Text Categorization Using SVM and NB, International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011

[18] Rish, Irina (2001). An empirical study of the naïve Bayes classifier . IJCAI Workshop on Empirical Methods in AI.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[19] Chotirat ann ratanamahatana and dimitrios gunopulos, FEATURE SELECTION FOR THE NAIVE BAYESIAN CLASSIFIER USING DECISION TREES, Applied Artificial Intelligence, 17:475–487, 2003

[20] https://stackoverflow.com /questions/ 10059594/ a-simple-explanation-of-naive-bayes- classification

[21] XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, Top 10 algorithms in data mining, Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007 Published online: 4 December 2007, © Springer-Verlag London Limited 2007

[22] Motaz K Saad, Wesam Ashour, (2010)Arabic Text Classification Using Decision Trees(2010) proceedings of the 12th international workshop on computer science and information technologies CSIT'2010, Moscow – Saint-Petersburg, Russia, 2010

[23] Mofleh Al-diabat ,(2012) Arabic Text Categorization Using Classification Rule Mining, Applied athematical Sciences, Vol. 6, 2012, no. 81, 4033 – 4046

[24] Abdullah H. Wahbeh* and Mohammed Al-Kabi (2012), Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text, Abhath Al-Yarmouk: Basic Sci. & Eng. Vol. 21, No. 1, 2012, pp. 15- 28

[25] Mesleh, A.M. (2008), "Support Vector Machines Based Arabic Language Text Classification System: Feature Selection Comparative Study," Advances in Computer and Information Sciences and Engineering,
Springer Science + Business Media B.V., 2008

[26] Thorsten Joachims. "Text categorization with support vector machines: learning with many relevant features". InProceedings of the 10th European Conference on Machine Learning ECML-98, Chemnitz, Germany. Pages 137–142. 1998.

[27] Vladimir, N., Vapnik. 1995. The Nature of Statistical Learning Theory. Springer-Verlag Berlin.

[28] Cristianini, N., and J. Shawe-Taylor. 2000 An Introduction to Support Vector Machines (and other kernel-based learning methods). Cambridge University Press

[29]http://sebastianraschka.com /Articles/2014_kernel_pca.html

[30] Ghazi Raho , Riyad Al-Shalabi, , Ghassan Kanaan, Asma'aNassar Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 2, 2015

[31] Mary Walowe , A Review on Feature Selection Methods For Classification, International Journal of Computer Applications Technology and Research Volume 5– Issue 6, 395 - 402, 2016, ISSN:-2319–8656

[32] Adel Hamdan Mohammad, Omar Al-Momani , Tariq Alwada'n, Arabic Text Categorization using k-nearest neighbour, Decision Trees (C4.5) and Rocchio Classifier: A Comparative Study, International Journal of Current Engineering and Technology, Accepted 03 December 2015, Available online 10 March 2016, Vol.6, No.2 (April 2016)

[33] Adel Hamdan Mohammad, Omar Al-Momani , Tariq Alwada'n, Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network, GSTF Journal on Computing (JOC), Volume 5, Issue 1; 2016 pp. 108-115

[34] Ashraf Odeh, Aymen Abu-Errub, Qusai Shambour and Nidal Turab, ARABIC TEXT CATEGORIZATION ALGORITHM USING VECTOR EVALUATION METHOD, International Journal of Computer Science & Information Technology (IJCSIT) Vol 6, No 6, December 2014

[35] Bilal Hawashin, Ayman M Mansour, Shadi Aljawarneh, An Efficient Feature Selection Method for Arabic Text Classification, International Journal of Computer Applications (0975 – 8887), Volume 83 – No.17, December 2013.

[36] Rehab Duwairi (2005) ,Machine learning for Arabic text categorization ," Journal of American society for information science and technology (JASIST), Vol57, No8,pp1005-1010, 2005.

[37]Eldos T . (2003), "Arabic Text Data Mining" A root Based Hierarchical Indexing Model", International Journal of Modeling and Simulation, vol23, no3,pp158-166,2003.