International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Arabic Text Categorization using k-nearest neighbour, Decision Trees (C4.5) and Rocchio Classifier: A Comparative Study

Adel Hamdan Mohammad†*, Omar Al-Momani‡ and Tariq Alwada'n†

†Computer Science Department, The world Islamic Sciences and Education University, Amman-Jordan
‡Network Department, The World Islamic Sciences and Education University, Amman-Jordan

Accepted 03 December 2015, Available online 10 March 2016, **Vol.6, No.2 (April 2016)**

Abstract

No doubt that text classification is an important research area in information retrieval. In fact there are many researches about text classification in English language. A few researchers in general talk about text classification using Arabic data set. This research applies three well known classification algorithm. Algorithm applied are K-Nearest neighbour (K-NN), C4.5 and Rocchio algorithm. These well-known algorithms are applied on in-house collected Arabic data set. Data set used consists from 1400 documents belongs to 8 categories. Results show that precision and recall values using Rocchio classifier and K-NN are better than C4.5. This research makes a comparative study between mentioned algorithms. Also this study used a fixed number of documents for all categories of documents in training and testing phase.

Keywords: *Text Categorization, k-nearest neighbour, Decision tress, C4.5, Rocchio classifier*

1. Introduction

Text categorization (TC) is a very important area in information retrieval. Text categorization or text classification is a method used in data mining aims to extract valuable information from large amount of data [Rasha Elhassan]. Also the huge number of documents available on line makes the process of text classification essential and important, moreover, the huge number of documents available on line make the process of text classification an active research area [Motaz K Saad, (2010)]. Text classification aims at categorizing documents into a set of predefined documents based on their content [Mofleh Al-diabat (2012), Sebastiani, F. (2002)]. Text classification has been used in many applications such as email filtering (spam or legitimate) [Adel Hamdan (2011), Raed Abu-Zitar(2011)] , indexing of scientific articles and news monitoring [Sebastiani, F. (2002), Dharmadhikari, C.S.(2011)].

A lot of researchers talk about the process of text classification in English language. And actually there are many algorithms developed and used for English text classification such as Support Vector Machines (SVM), Bayesian Classification, Hidden Markov model, neural network, k-nearest neighbour algorithm and others. Unfortunately, only few researches talk about Arabic text classification. Arabic language is the main language in the Arab world and the secondary language in the Islamic countries.[Mofleh Al-diabat,(2012), Duwairi, R. (2007)]. Arabic language has a rich morphology and a complex orthography [Mesleh, A.M. (2008].

This paper used k-nearest neighbour, Decision Trees (C4.5) and Rocchio Classifier in Arabic text classification. K-NN and Rocchio are two classifiers frequently used for TC, and they are both similarity based. This paper aims at making a comparative study between mentioned algorithms on an Arabic data set. Actually and up to this paper date authors did not find any research that make a comparative study between mentioned algorithms on a large Arabic data set.

The rest of this paper organized as follow: section 2 talks about Arabic language, section 3 talks about document pre-processing steps, Section 4 explains Term Selection and Weighting methods, section 5 talks about k-nearest neighbour, section 6 talks about C4.5 (Decision Tree), section 7 talks about Rocchio classifier, section 8 demonstrate related studies, section 9 show data set used in experiments, and finally section 10 show our experiments and results.

2. Arabic Language

Arabic language is the main language in 25 countries. Beside that Arabic language is spoken from more than 250 million. Arabic language consists of 28 letters plus hamza (ء) which is considered a letter by some Arabic linguistics. Arabic language is written from right to left. Also the letters of Arabic language have different

*Corresponding author: **Adel Hamdan Mohammad**

shapes when appearing in a word depending on the position of the letter (at the beginning, at the end, at the middle) [Rehab Duwairi (2005)]. The majority of Arabic words have their roots, beside that over 80% of Arabic words can be mapped into 3-letter root. Fortunately, Arabic language has its own building mechanism which means different words can be mapped into their roots. Representing words according to their roots is very important and definitely will reduce the number of words. [Eldos, (2003)]

3. Document Pre-processing

Document processing and representation is essential step to clean the text by removing useless data such as removal of preposition, pronouns, conjunctions words, auxiliary verbs, digits, numbers, stop words and formatting tags.

Arabic language is a very rich language. This richness increases the size of feature vectors created. Fortunately, Arabic language has it built-in filtering mechanisms. Most words in Arabic language can be mapped into their roots using stemming. Root in Arabic language available in three, four, five and six letters. Also over 80% of Arabic words can be mapped into three-letter root. [Rehab Duwairi (2007)]

4. Term Selection and Weighting

Data pre-processing is an essential step in text classification. Date pre-processing aims to reduce the complexity of text documents. Data pre-processing can be classified into Feature Extraction (FE) and Feature Selection (FS). [Liu, H. (1998)].

Feature Extraction aims to clean text documents. Feature Extraction includes stemming and presenting text in a clear format [Wang, Y and Wang X (2005)].

Feature Selection used after Feature Extraction to build vector space. FS also aims to select suitable features from the original document. Beside that FS aim at keeping words that have the highest scores according to a set of predefined measures.[Zi-Qia ng,(2005), Montanes, E,(2003)].

No doubt that representing all document words into their roots is very important. [Eldos T. (2003)]. Root extraction or stemming for Arabic language can be divided into two types.[Al-Shalabi R(1998), El-Sadany T. A(1989), Gheith M(1987), Hilal Y(1989)]. First, getting the root based on the idea of removing prefixes, infixes, suffixes. Second, getting the root based on the weight of letters embedded within the text. [Al-Shalabi, R., Kanaan, (2003)].

Obviously, There are many term evaluation function used for English text categorization such as Chi square, information gain, Document Frequency Threshold [Franca Debole (2003), Al-Zaghoul F (2013), Sebastiani F. (2002)]. Also after selecting the most important terms in each document, each document must be weighted as vector based on the words found in it. In fact there are many weighting techniques such Term Frequency (TF), Inverse Document Frequency, Term Frequency Inverse Document Frequency (TFIDF), and Normalized-TFIDF weighting. [Johannes Furnkranz (1998), Abu-Errub A (2014)].

5. K-Nearest Neighbour

Key nearest neighbour is one of the statistical learning algorithms that have been used for text classification [Y. Yang and X. Liu (1999), Riyad Al-Shalabi (2006)]. K-NN is known as one of the top classification algorithms for English language. The k-nearest-neighbour classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. The idea of K-NN can be explained as follow: when the algorithm tests a document, first the algorithm tries to find the k nearest neighbours among the training documents. Also the K-NN uses the categories of the k Neighbours to weight the candidate category. In K-NN the similarity score of each document to the test document is used. If several of the k nearest neighbours shares a category then the resulting weighted sum is used as the likelihood score of candidate categories. One of the major disadvantages of K-NN algorithm is its difficulty to determine the best value of K and the complexity of computation time, as it requires to compare a test document with all samples in the training set generally, the value of k is usually selected based on many trials of the training and validation sets. [Riyad Al-Shalabi (2006), Gongde Guo(2005), Li Baoli(2003)]

In this paper the cosine similarity measures is used to calculate the similarity between documents. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. If we assume that A and B are vectors representing documents J and K respectively then we can calculate the similarity between vectors A and B using the following formula (Formula 1)[Riyad Al-Shalabi (2006)]. Also in this paper we used Normalized-TFIDF weighting technique.

$$\text{SIM (A, B)} = \frac{\sum_{i=1}^{r} W_{ij} * W_{ik}}{\sqrt{\sum_{i=1}^{r} W_{ij}^2} * \sqrt{\sum_{i=1}^{r} W_{ik}^2}} \quad (1)$$

6. C4.5 (Decision Tree)

C4.5 is one of statistical classifiers which can be used to generate decision tree. C4.5 developed by Ross Quinlan .Decision tree generated by C4.5 can be used for text classification. [Motaz K saad (2010), Abdullah H. Wahbeh(2012), Mofleh Al-diabat(2012)].

A decision tree like tree structures, in decision tree each internal node represents a test to the document, also each branch in the tree represents an outcome of the test, beside that each leaf node holds a class label. Basically it is a top-down technique which recursively constructs a decision tree classifier [Nidhi(2011)].

First C4.5 developed to overcome limitation of ID3. One major limitation of ID3 is too sensitive to the features with large numbers of values. C4.5 builds his

decision trees from a set of training data. [Badr hssina (2014)]. In C4.5 if you have a set of records and each record has the same structure and consists of a number of attributes then one of these attributes will represent the category of the document.

C4.5 works as follow: Given a set S of cases, C4.5 will first grow an initial tree using divide and conquer algorithm. After that C4.5 will evaluate the cases. If all the cases in S belong to the same class or S is small. Then the tree is a leaf and will be labelled with the most frequent class in S. otherwise. C4.5 will select a test based on a single attribute which has two or more outcomes. In addition C4.5 will make this test the root of the tree with one branch for each outcome of the test. Then partition S into equivalent subsets S1, S2,… according to the outcome for each case. Then apply the same procedure recursively [(XindongWu(2007)]. C4.5 uses information Gain and Gain Ration to rank possible tests.

7. Rocchio Classifier

The rocchio algorithm idea based on relevance feedback found in information retrieval. Rocchio is a linear classifier. Rocchio feedback approach developed using vector space model (vector space model is an algebraic model for representing text documents). Rocchio algorithm is created based on the assumption that most users have a common conception of which documents should be indicated as relevant or non-relevant. In rocchio classifier a prototype vector is built for each class Ci, and a document vector d is classified by calculating the distance between *d* and each of the prototype vectors. Then prototype vector for class Ci is calculated as the weighted average vector over all training document vectors that belong to class Ci. [(F. Sebastiani(1999), F. Sebastiani(2002), M. M. Syiam(2006)].

Rocchio method is efficient and easy. In rocchio classifier classifying new instance requires computing the inner product between the new instance and the generalized instances. Also Rocchio classifier can summarize the contribution of the instances belonging to each category. Beside that one major advantage of rocchio classifier is its capability of filter out certain irrelevant features [Gongde Guo(2006)].

The Rocchio classifier is a linear classifier. Given a training dataset Tr, it directly computes a classifier C_i (W_{i1},W_{i2},….W_{in}) for category C_i. The weighted average of a category is computed as follow [Tarek Fouad Gharib(2009), Gongde Guo(2006)]:

$$W_{ik} = \beta \sum_{dj \in POSi} \frac{Wjk}{|POSi|} - \gamma . \sum_{dj \in NEGi} \frac{Wjk}{|NEGi|}$$

Where W_{ik} is the weight of term t_k in document d_j. POS_i and NEG_i means document d_j belonging to (or not belonging to) category *ci.* β and γ are control parameters that allow setting of positive and negative examples [Gongde Guo(2006)].

8. Related Studies

Wail Hamood [Wail Hamood (2014)] applied traditional K-NN and Naïve Bayesian using Weka toolkit on Arabic data set. Also he proposed modified K-NN algorithm to skip the classes that less similar and identify the right class from k nearest neighbours which increases the accuracy. He concludes that his modification is promising. Also in his paper an improved K-NN implemented on Arabic data set improves the performance.

Riyad al-Shalabi [Riyad al-Shalabi (2006)] implemented K-NN on Arabic data set. He has reached 0.95 micro-average precision and recall stores. Also he uses 621 Arabic text documents belong to six different categories. He has used a feature set consist of 305 keywords and another one of 202 keywords. Selection of keywords based on Document Frequency threshold (DF) method.

Tarek Fouad [(Tarek Fouad(2009)] applied SVM model for classifying Arabic text documents. he compare his results with Bayes and Rocchio classifier. His experiments applied on 1132 documents. Rocchio classifier gave the best results when the size of feature set is small while SVM outperform the other classifiers when the size of the feature set is large enough.

M. M. Syiam [M. M. Syiam(2006)] had used k-nearest neighbour and Rocchio classifiers are used for classification process. Experiments done on self-collected Arabic data set. In his experiments he showed that Rocchio classifier has the advantages over k-nearest neighbour classifier in the classification process. The experimental results demonstrate that the proposed model is an efficient method and gives generalization accuracy of about 98%.

Majed Ismail Hussien [Majed Ismail Hussien(2011)] implemented the Sequential Minimal Optimization (SMO), Naïve Bayesian (NB) and J48 (C4.5) Algorithms using Weka program, these algorithms implemented on Arabic data set. His experiments demonstrate that the (SMO) classifier achieves the highest accuracy and the lowest error rate, followed by J48 (C4.5), then the (NB) classifier.

Al-Harbi [Al-Harbi(2008)] presents the results of classifying Arabic text document using seven different Arabic data set. He used SVM and C4.5 algorithms. A tool for feature extraction and selection was implemented. Is his experiment C5.0 classifier gives better accuracy.

9. Data set

Dataset set used in our experiments collected from Aljazeera news web site (http:// www.aljazeera.net), Saudi Press Agency (SPA) (http://www.spa.gov.sa/ index.php) and Al-hayat (http:// www.alhayat. com/).Dataset consists of 1400 Arabic documents belong to different categories (see table 1).

Table 1: Data set

Category	Total Number of documents	Number of documents used for training	Number of documents used in testing
Computer	175	115	60
Economics	175	115	60
Education	175	115	60
Law	175	115	60
Medicine	175	115	60
Politics	175	115	60
Religion	175	115	60
Sports	175	115	60
Total	1400	920	480

Approximately, dataset is divided into two parts. Dataset used for training is 920 documents (66%) and 480 documents is used for testing (34%).

10. Experiments and Results

In this paper authors evaluate K-NN, C4.5 and Rocchio classifier on an Arabic data set. Data set used in our experiments is in-house developed data set which consists of 1400 documents. Also authors in this paper use recall, precision and F1 as evaluation measures. Precision are also called (positive predictive value). In information retrieval precision is the number of true positive which means the number of items correctly labelled as belonging to the positive class divided by the total number of elements labeled as belonging to the positive class. Also Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class. And F1 is a measure of a test's accuracy. Consider table 2 and suppose that a: relevant document retrieved, b: irrelevant document retrieved, c: relevant document not retrieved, d: irrelevant document not retrieved. (See table 2).

Table 2: Document possible sets based on a query

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

$F1 = 2.\frac{precision*recall}{precision+Recall}$

Precision= $\frac{a}{a+b}$

Recall= $\frac{a}{a+c}$

Goal of these experiments is to evaluate K-NN, C4.5 and Rocchio algorithm on an Arabic data set. As mentioned previously there are a few researches using proposed methods on Arabic data set are done. And no doubt that applying most of classification algorithm on Arabic data set still need more investigation. In addition applying a comparative study on the mentioned algorithms is one of the goals of this research.

Table 4 and 5 demonstrates our experiments using Rocchio and C4.5 algorithms.

Table 4: Rocchio Results

Category(Rocchio)	Precision	Recall	F1
Computer	0.78	0.87	0.82
Economics	0.88	0.89	0.88
Education	0.79	0.78	0.78
Law	0.81	0.91	0.85
Medicine	0.9	0.89	0.89
Politics	0.81	0.79	0.79
Religion	0.79	0.71	0.74
Sports	0.81	0.83	0.81
Average	0.82	0.833	0.82

Table 5: C4.5 Results

Category (C4.5)	Precision	Recall	F1
Computer	0.66	0.59	0.62
Economics	0.74	0.71	0.72
Education	0.59	0.57	0.57
Law	0.61	0.63	0.61
Medicine	0.68	0.61	0.64
Politics	0.69	0.68	0.68
Religion	0.61	0.63	0.61
Sports	0.78	0.71	0.74
Average	0.67	0.641	0.65

Experiments demonstrated in table 4 and 5 show that the average precision and recall using Rocchio classifier is better than C4.5 on selected data set.

The value of K is very important on K-NN algorithm. Small value of K means low performance of recall and precision. Also high value of K means high computation.

Table 6,7,8 and 9 demonstrates our result using different values ok K.

Table 6: K-NN Results (K=1)

Category (K-NN 1)	Precision	Recall	F1
Computer	0.33	0.41	0.36
Economics	0.22	0.21	0.21
Education	0.19	0.18	0.18
Law	0.15	0.17	0.15
Medicine	0.19	0.23	0.20
Politics	0.32	0.31	0.31
Religion	0.31	0.23	0.26
Sports	0.41	0.39	0.39
Average	0.26	0.26	0.26

Table 7: K-NN Results (K=3)

Category (K-NN 3)	Precision	Recall	F1
Computer	0.35	0.36	0.35
Economics	0.37	0.38	0.37
Education	0.33	0.33	0.33
Law	0.41	0.39	0.39
Medicine	0.38	0.7	0.49
Politics	0.41	0.42	0.41
Religion	0.45	0.43	0.43
Sports	0.45	0.47	0.45
Average	0.39375	0.435	0.40

Table 8: K-NN Results (K=12)

Category (K-NN 12)	Precision	Recall	F1
Computer	0.55	0.56	0.55
Economics	0.58	0.57	0.57
Education	0.58	0.57	0.57
Law	0.61	0.71	0.65
Medicine	0.59	0.62	0.60
Politics	0.63	0.64	0.63
Religion	0.59	0.87	0.70
Sports	0.55	0.66	0.60
Average	0.585	0.65	0.61

Table 9: K-NN Results (K=18)

Category (K-NN 18)	Precision	Recall	F1
Computer	0.78	0.73	0.75
Economics	0.77	0.81	0.78
Education	0.75	0.76	0.75
Law	0.91	0.89	0.89
Medicine	0.85	0.84	0.84
Politics	0.84	0.81	0.82
Religion	0.87	0.79	0.82
Sports	0.91	0.84	0.87
Average	0.83	0.80	0.82

The results using different values of k using K-NN show that when k=1, the average results for almost all types of categories are bad. The results are promising when k values are 12 and 18. As shown in table 9 the best results are shown in table 9 when the value of k is 18. Authors made experiments for k =24 but the results is not encouraging so it is not demonstrated in out tables.

In summary our results demonstrate that K-NN and Rocchio can work well on Arabic data set. C4.5 results are not promising in our experiments.

Reference

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

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

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

F. Sebastiani. (2002) Machine learning in automated text categorization. ACM Computing Surveys, volume 34 number 1. PP 1-47. 2002.

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.

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.

Dharmadhikari, C.S., Ingle, M. and Kulkarni, P.(2011) Empirical Studies on Machine Learning Based Text Classification Algorithms, Advanced Computing: An International Journal (ACIJ), Vol.2, 2011.

Rehab Duwairi (2007), Arabic c text categorization, the international Arab journal of information technology , Vol4, No2,April 2007.

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

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

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

Liu, H. and Motoda,(1998) ., Feature Extraction, constraction and selection: A Data Mining Perpective., Boston, Massachusetts(MA): Kluwer Academic Publishers.

Wang, Y., and Wang X.J.(2005), A New Approach to feature selection in Text Classification, Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, Vol.6, pp. 3814-3819, 2005.

Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li (2005)An Optimal Svm-Based Text Classification Algorithm Fifth International Conference on Machine Learning and Cybernetics, Dalian,pp. 13-16 , 2006.Barizal, , pp.122-129, 2005.

Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J.,(2003) Measures of Rule Quality for Feature Selection in Text Categorization, 5th international Symposium on Intelligent data analysis, Germeny-2003, Springer- Verlag 2003, Vol2810, pp.589-598, 2003.

Al-Shalabi R, and Evans M.,(1998)A Computational Morphology System For Arabic, in proceedings of Computational approaches to semitic languages workshop (COLING'98), Montreal, Canada, pp.58-65,1998.

El-Sadany T. A. and Hashish M. A. ,(1989) An Arabic Morphological System . IBM Systems Journal, vol. 28, no 4,pp 600-612,1989.

Gheith M. and El-Sadany T., (1987)Arabic Morphological Analyzer on a personal Computer, in proceedings of the Arabic Morphology Workshop Stanford University, California, USA, pp55-65, 1987.

Hilal Y, (1989)Automatic Processing of Arabic Language and application [in Arabic] , in proceedings of the 1st Kuwaiti Computer conference, Kuwait, pp. 145-171, 1989.

Al-Shalabi, R., Kanaan, G., and Muaidi, H. (2003). New Approach for Extracting Arabic Roots. Proceeding of the International Arab Conference on Information Technology. Alexandria, Egypt

Franca Debole et al.,(2003) Supervised Term Weighting for Automated Text Categorization, proceedings of SAC-03,

18th ACM Symposium on Applied Computing, Melbourne, 2003, USA

Al-Zaghoul F., Al-Dhaheri S.,(2013) Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks, UKSim, pp. 485-490. 2013.

Johannes Furnkranz.,(1998) A Study Using n Gram Features For Text Categorization, Technical Report OEFAI-TR-1998-30.

Abu-Errub A., (2014) Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements, International Journal of Computer Applications 93 (6), 40-45, 2014.

Y. Yang and X. Liu. (1999). A re-examination of text categorization methods. In proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99),42-49, 1999

Riyad Al-Shalabi, Ghassan Kanaan, Manaf H. Gharaibeh (2006), Arabic Text CategorizationUsing kNN Algorithm, Proceedings of The 4th International Multiconference on Computer Science and Information Technology, Vol 4, P 5-7,2006

Gongde Guo, Hui Wang, David Bell , Yaxin Bi , and Kieran Greer (2006), Using kNN Model-based Approach for Automatic Text Categorization, Soft Computing , March 2006, Volume 10, Issue 5, pp 423-430

Li Baoli, Yu Shiwen, and Lu Qin (2003), An Improved k-Nearest Neighbor Algorithm for Text Categorization, Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China, 2003.

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

Nidhi, Vishal Gupta (2011), Recent Trends in Text Classification Techniques, International Journal of Computer Applications (0975 – 8887) Volume 35– No.6, December 2011

Badr Hssina, Abdelkarim Merbouha,Hanane Ezzikouri,Mohammed Erritali, (2014) A comparative study of decision tree ID3 and C4.5, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications (2014)

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, (2007) 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, Knowl Inf Syst (2008) 14:1–37, DOI 10.1007/s10115-007-0114-2

Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar (2013), A review paper on algorithms used for text classification, International Journal of Application or Innovation in Engineering & Management (IJAIEM), ISSN 2319 - 4847Volume 2, Issue 3, March 2013.

F. Sebastiani.(1999) A Tutorial on Automated Text Categorisation. Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence. PP 7-35. 1999.

M. M. Syiam, Z. T. Fayed & M. B. Habib,(2006) An intelligent system for arabic text categorization, International Journal of Intelligent Computing and Information Sciences IJICIS, Vol.6, No. 1, January 2006

Tarek Fouad Gharib, Mena Badieh Habib , and Zaki Taha Fayed,(2009) Arabic Text Classification Using Support Vector Machines, International Journal of Computers and Their Applications, Vol (16), Issue(4), (2009).

Wail Hamood Khaled, Haytham Saleem AL-Sarrayrih, Lars KNIPPING (2014) Arabic Text Categorization Using Improved k-Nearest neighbour Algorithm, Journal of Applied Computer Science & Mathematics, no. 18 (8) /2014, Suceava

Majed Ismail Hussien, Fekry Olayah, Minwer AL-dwan & Ahlam Shamsan,(2011) Arabic text classification using smo, naïve bayesian, J48 Algorithms, International Journal of Research and Reviews in Applied Sciences ,Vol 9 Issue 2/IJRRAS_9_2_15.pdf, November 2011

Al-Harbi S., Almuhareb A., Al-Thubaity A., Khorsheed M. S., and Al-Rajeh A (2008), Automatic Arabic Text Classification, 9es Journées internationals, France, pp. 77-83, 2008.