Arabic Text Classification Process

Rasha Elhassan[1] and Mahmoud Ali[2]

[1] Information Systems Department, King Khalid University, Khamis Mushait, 61961 , Kingdom of Saudi Arabia

[2] Faculty of Mathematical Science, University of Khartoum, Khartoum, 11115, Sudan

[1]relhasan@kku.edu.sa, [2]mali@uofk.edu.sd

ABSTRACT

Due to the richness of the language, the ordinary Arabic Text classification of the content is very complex, difficult and challenging task. This resulted as unavailable benchmarking Arabic corpus, limitation of research and ambiguous processing phases in the field of Arabic text classification, This paper presents the complex nature of Arabic language, poses the problems of lacking free public Arabic corpora, explain the Classification phases throw the literature Arabic text classification.

Keywords: *Arabic Text Classification, Data Preprocess, Feature Selection, Classifier*.

1. INTRODUCTION

Text classification is a common technique that used to labeling documents with thematic categories from a categorized set or use specific criteria to belongs it to some category [1, 2]. Text categorization is taken to deal with sorting documents by content, while text classification is used to classify the documents by any kind of assignment of documents to classes [3].

The main goal of text classification is to extract the information with value from unstructured textual resources. It deals with the operations like, retrieval, classification and summarization [4]. There are two approaches to the text classification: rule based and machine learning [5]. In the rule based an expert system that automatically classified text is built with a knowledge engineer and a domain expert, and in the machine learning a general inductive process is fed with a set of “training” documents, preclassified according to the categories of interest [6]. Spam and e-mail filtering; news monitoring; and searching for interesting information on web sites used the technique of text classification [7]. The quality of the data source may affect the performance of a classification algorithm; the irrelevant and redundant features of data may reduce the quality of the result [4].

Arabic language has a complex orthography and rich morphology [8]. The size of feature vectors may increases and that make the task of feature selection more important to avoid the irrelevant data [9]. On contrary of Arabic text, there are many benchmarking corpus that can be used for Latin base language, Japanese and Chinese text classification [1, 10, 11]. For Arabic language there is a lack of the studies on the classification of Arabic text documents with limitation of free benchmarking dataset [11, 12]. On the other hand the richness of Arabic Language increases the size of the feature vectors and [13] that affected negatively the total number of research and studies in the field of text classification.

2. PREVIOUS WORK

Almost of the authors used an in-house Arabic which had different sizes and contents, which came from different web sites [14, 15]. The studies in the field of Arabic text classification used Naïve Bayes [16, 17], Support vector machine [18], Decision Trees [13] as classifier algorithm.

In [13], the author collected in-house Arabic corpus that consists of 1000 documents, were divided into 10 categories and used it to compared the performance of three Arabic text classifiers: Naïve Bayes, K-Nearest Neighbor and Distance–Based. The recall, precision, error rate and fallout are using to compare the accuracy of the classifiers. The data was preprocessed by removed the stop words and extracted the root of the words. The results of the experimentation show that the Naïve Bayes classifier outperforms the other two (more than 95%).

In [19] the authors used in house corpus consists of 242 documents which belong to six categories to evaluate the performance of two classifiers K-NN and Naive Bayes to classify the Arabic text. The k-fold cross-validation method is used to test the accuracy. They extracted feature set of keyword to improve the performance. The result showed that K-NN classifier had better performance.

[20] Investigated Naïve Bayesian and Support Vector Machine on different Arabic data sets. The Experimental results against different Arabic text categorization data sets reveal that SVM algorithm outperforms the NB. While in [21] which compared the Support vector

machine and C5.0 classifier and C5.0 classifier gives better accuracy.
By [12]. Support vector machine had shown the superiority in feature selections, weighting methods, and classification algorithms, followed by the decision tree algorithm (C4.5) and Naive Bayes. The best classification accuracy was 97% for the Islamic Topics dataset, and the least accurate was 61% for the Arabic Poems dataset.

3. ARABIC LANGUAGE STRUCTURE

Arabic language is richness and complex language that makes the analysis very complex task [5, 22]. Opposite to Latin-based alphabets, Arabic language written from right to left and has 28 characters. By additional shapes, marks, and vowels, the alphabets can be extended to ninety elements [13]. There are two genders: masculine (مذكر) and feminine (مؤنث) which present by attached the suffix (ة) to the end of the word. Numerically, there are singular (مفرد) , dual (مثنى), and plural (جمع) numbers. Plurals numbers are regular (جمع الصحيح) or broken جمع) (تكسير[23].
Grammatical there are three cases: nominative (الرفع) when it's subject, accusative (النصب) when it's the objective of a verb, and genitive (الجر) when it's the object of the preposition [22]. There are three parts of speech: verbs (افعال), adverb (ظرف) and nouns (اسماء) which including adjectives (صفات) and particles (احرف) [24]. Noun has the genitive case when it is the object of a preposition; nominative case when it is subject; and accusative when it is the object of a verb [5]. Arabic language doesn't distinguish the proper and common nous like English language.
The pattern Fa'ala (فعل) is the base of most of Arabic words and all words that following that pattern has common properties and states [5]. All words following the same pattern have common properties and states. For example, the pattern Faa'el (فاعل) indicates the subject of the verb, the pattern Maf'ool (مفعول) represents the object of the verb. Table 1 illustrated different derivations for the root word.

Table1: Different derivations for the root word

Pattern	Arabic words			
Fa'ala فعل	علم	فتح	درس	قراء
Faa'el فاعل	عالم	فاتح	دارس	قاري
Maf'ool مفعول	معلوم	مفتوح	مدروس	مقروء

The characters "ي", "و", and the "ا" are used as long vowels and the rest of the alphabet are constant. Arabic also contains short vowels, called "التشكيل", which give different pronunciation of the word by double the letter [16, 25]: بَ (Fathah),بً (Double Fathah), بُ (Dama), بٌ (Double Dama), بِ (Kasra), بٍ (Double Kasra), بْ (Sukon) and بّ (Shada). It can change e the meaning of the word and give other sense of the sentence.

For example, the word "مُدرسة" has a different manning of "مَدرسة" the first one means the female teacher and the second one means the school. Grammatically they are required but rarely used in written modern text. Table 2 explained the effect of the short vowels.

Table 2: Short Vowels Effect

Short Vowels	Word	Word With short vowels	Meaning
Fathah	ضعف	ضَعْفٍ	old
Kasra		ضِعْفٍ	Double
Kasra	يصدون	يَصِدُّونَ	They Laugh
Dama		يَصُدُّونَ	They Prevent
Shada	المصدقين	الْمُصّدِّقِين	Gifted Money
Fathah		الْمُصَدِّقِينَ	To Believe

Also there is a kashida, horizontal line in the middle of the word, which does not change the meaning of the word as shown in Figure 1.

السودان
الســـــودان
الســـــــــــودان
الســـــــــــــــودان

Fig. 1. Words with Kashida

According to the position of the letter, Arabic letters change its shape [26]. There are 22 out of 28 letters connected to the other letters before and after them and call connectors letters [27]. The remaining, six letters which are called non-connectors letters, connect by the right side.

Table 3 illustrated the letters connection and their position shapes.

Tables 3: Letters Connection and Position Shapes

Non-connector	Beginning		Medial		End	
ا	احمد	ا	حامد	- ا	سما	- ا
د	دم	د	يهدى	- د	شهد	- د
ذ	ذنب	ذ	الذين	- ذ	فولاذ	ذ
ر	رمال	ر	طرق	- ر	ايثار	ر
ز	زيت	ز	الاحزاب	- ز	إيجاز	ز
و	وارفة	و	نوح	- و	يوليو	- و

Speech is structured from noun, verb and particles. Noun is named things like a man, horse and wall. Verb is taken from the word names and describes what happened in past, what is happening now and what will happen in future. Particles are not noun or verb meaning. Figure 2 shows the parts of Arabic speech

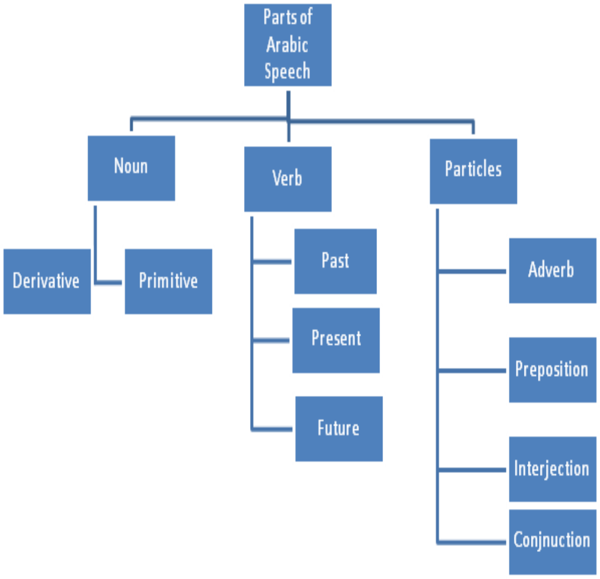

Fig. 2. Arabic Speech Parts

According to the position of the word, Arabic words change its meanings due to the whole meaning of the sentence. Table 4 explains some example of the different meaning of the word "قلب".

Tables 4: The Different Meaning of the Same Word

Word	Meaning	Sentence
كبد	Cindy	شرب المياه مفيد للكبد
	Truth	اصاب كبد الحقيقة
قلب	Heart	يضخ القلب الدم الي الجسم
	Core	شاهد في قلب الحدث

4. ARABIC TEXT CLASSIFICATION PROCESS

General the process of Arabic text classification contains three phases: Data pre-processing, Text classification and Classifier evaluation [10, 17]. Sometimes each phase can be divided into sequence steps as [12] divided it in more dilates: (a) data collection; (b) Text processing; (c) data division; (d) Feature extraction; (e) Feature selection; (f) Feature representation; (g) data representation; (h) classification algorithm training and testing; (i) classification model evaluation.
Simply it can be divided in five major phases which include all the other steps: data collection, data preprocessing, feature selection, text classification and the classifier evaluation. Figure 3 explained that phases.

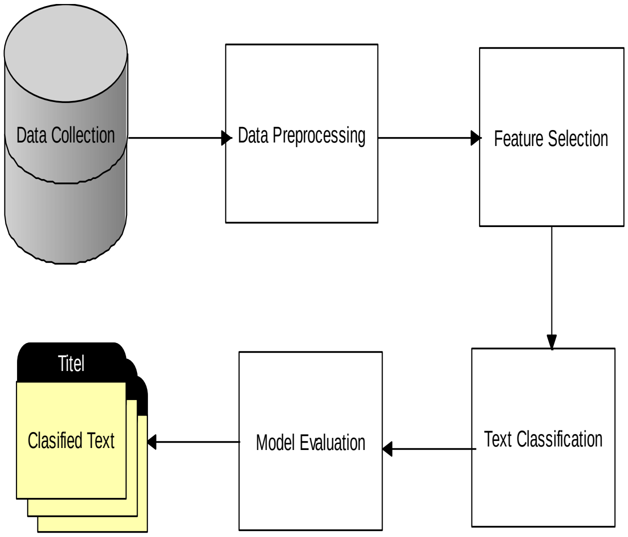

Fig. 3. Text Classification Phases

4.1 Data Collection

This step aims to build the suitable dataset with different instances that will be used in training and testing the model. There are many benchmarking datasets or corpus that can be used as initial corpora for text classification and it is publically available for experimental purposes for almost Latin base language, Chinese, Japanese and other international language [1] and the most widely used is the Reuters collection. There is a lack in the studies on the classification of Arabic text documents with limitation of benchmarking dataset [10, 11] and that led to low quality in the categorization of the Arabic online articles. Table 5 showed some available free Arabic corpus

Table 5: Free Arabic Corpus

Collector	Resource	Size
Dr Ahmed Abdelali	Journals	800 Mb
طه زروقي	Books from Al-Shamela library	122 Mb (zip)
Dr. Morad Abbas	Journals (AlWatan & Alkhaleej)	14 Mb
Dr. Latifa Al-Sulaitii	Journals	-
Dr Saad Motazz	Journals	15 Mb
Muayyed Al-Saadii	Thwab library	6 Mb (zip)

4.2 Pre-Processing

Data pre-processing aims to prepare a data in a good manner to be taken into classification task. Due to the richness of the Arabic language and complex language morphology as compared to the other languages [28], data pre-processing is a very challenge step in Arabic text classification [29]. The number of the terms that are extracted from the text is increases by the length of the documents and usually very big and contains unnecessary data.

How to prepare the data in a good manner to increase the classifier accuracy is the most important issue to classify the Arabic text. This phase contains: data preparation, stop words removal and word stemming.

The data preparation is done by extracting and normalizing the data. The document is extracted by converting it to a suitable form such as plain text and then tokenizing it into initial words. All non-Arabic base letters, digits, numbers, and punctuation marks are removed and the remaining character strings are considered as terms or tokens of the classification task [4]. For the richness of the Arabic language there is a need to normalizing some writing forms that include Hamza “ء”, Taa Marboutah “ة” to “ا” and “ي” to “ى” [5, 15].

Any words such as (the, or, and, a, an, for and such) are the words that don’t give any value to text classification process. They are called stop words. In Arabic language the stop words include nouns and particles that are common to all categories and do not give a full meaning in itself such as (الى, فوق, جيد, اجل) [13, 31].

There are different stop words lists with different sizes that are used in the experiments for Arabic text classification task. There are no standard stop words lists that are available and explained in details.

After removing unnecessary word from the documents the extraction of the word root [8] can done by using some stemming techniques. The root can be extracted by three ways: the root-based stemmer; the light stemmer; and the statistical stemmer [17].

Next the documents are indexed and represented as a vector of weighted terms [5, 18]. Some authors included the feature selection in the step of data preprocessing.

Table 6 summarized the step of data preprocessing in the Arabic text classification literature and give information about the corpus witch used by the authors:

Table 6: Data preprocessing

Ref.	Corpus Type.	Data preprocessing
Author [5]	In House	Remove digits, punctuation marks, non-Arabic letters, normalization some letters
Author [23]	In House	Stemming
Author [10]	Saudi Newspapers	Remove digits, punctuation marks, non-Arabic letters, normalization some letters.
Author [13]	Seven genres	Remove stop words.
Author [21]	In House	Remove stop words, digits, and punctuation marks.
Author [18]	In House	Stop word remove and stemming
Author [13]	In House	Remove stop words, punctuation marks and stemming word and store in feature vector.
Author [19]	In House	Remove stop words.
Author [21]	SNP	Remove digits, punctuation marks, non-Arabic letters, normalization some letters.
Author [12]	In House	Remove digits, numbers, and punctuation marks, non-Arabic letters, normalization some letters.
Author [23]	In House	tf.idf measure for feature selection

4.3 Feature Selection

One technique for dimensionality reduction is feature selection technique that aims to improve text classification model accuracy by removing irrelevant, redundant and noisy data that lead to avoid over fitting which improves the performance of the model and also provides faster and more cost-effective models [30, 31, 32, 33]. There are two approaches for feature selection, wrapper or filter approaches [34].

a. In the wrapper approach the selection of subset of features is process based on the learning algorithm that used to train the model.

b. In filter approach the selection of subset of features is independent in the characters of the data [34, 35, 36, 37, 38].

There are many methods that use for feature selection as: Document Frequency Thresholding (DFT), Information Gain (IG), Mutual Information (MI), Chi-square statistic (CHI), Term Strength (TS) and Odds ratio [8, 23].

4.4 Text Classification

In this step the training set is used to train the classification algorithm [5, 23, 10, 14]. Train set aims to learn the classifier how to classify the documents by using the feature of each one. A lot of algorithms can be used to design the model of the classifier: Naïve Bayes [3, 4], Support vector machine [5], KNN and Decision Trees [6]. Table 7 summarizes the algorithm that used to classify the text.

Table 7: Algorithm Type

References.	Classifier Used
Author [5]	SVM/NB/ KNN
Author [8]	KNN/ Rocchio
Author [12]	Cosine, Jaccard, Dice
Author [13]	SVM/ C5.0
Author [16]	(SMO), NB, J48.
Author [19]	SVM/NB/ KNN/ Rocchio
Author [21]	NB/KNN/ Distance-Based
Author [21]	KNN/NB
Author [23]	SVM/NB
Author [23]	SVM/C 4.5/NB
Author [24]	KNN/SVM

4.5 Evaluation

Precision, recall and F-measure are evaluation measures specified by Rijsbergen, (1975) as citied by [40] and widely used in the field of information retrieval and classification tasked to evaluated the performance in terms of computational of efficiency and effectiveness of categorization [8, 40, 41, 42]

To explain how the basic evaluation measures of classification task are defined, a table of actual and predicted results given by a classifier is considered. Table 8 explains the actual and predicted results.

Table 8: Actual and Predicted Results

	Classified Positive	Classified Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The evaluation measures are computed based on the following equation [43].

Precision (P) = TP/ (TP+FP)
(1)

Recall (R) = TP/ (TP+FN)
(2)

F-score = $\frac{(1+\beta^2)*P*R}{\beta^2*P+R}$
(3)

Most commonly, β is taken to be 1.

The authors had used different types of measures to test their classifiers and this may not justified the result. Table 9 summarizes the accuracy measures that were used to measure the evolution of the algorithm and the percentage of it:

Table 9: Evaluation Methods and their Percentage

Refer.	Accuracy Measure	Classifier/ Accuracy Percentage
Author [5]	Precision, recall, F1.	SVM / 90%
Author [8]	Precision, recall	Rocchio / 98%
Author [12]	Precision, recall, F1.	Cosine / 95%
Author [13]	Correct/total	C5.0/ 78%
Author [16]	Percentage split method.	SMO / 96%
Author [19]	Micro average	SVM / 90%
Author [21]	Precision, Recall, Fallout and error rate.	NB / -
Author [21]	K-fold	KNN / 84%
Author [23]	Recall, Precision, and F1.	SVM / 3.8%
Author [23]	Percentage spilt/10 folds	NB / 85%
Author [24]	Recall, Precision, and F1.	SVM / 95%

5. CONCLUSION

Rewritten the result of unavailable Arabic corpus, most of the authors used their own corpus. There is a need to build a benchmarking data set for Arabic text classification. There are different data pre-processing steps that used to prepare the text. There are a lot of text classifications algorithms with well performance. Support vector machine algorithm show superiority of it, followed by the Naïve Bayesian algorithm. Precision, recall and F-measure techniques are used to evaluate the classifier. The future work is to explain experimentally each phases of Arabic text classification process by details.

REFERENCES

[1] Sebastiani, F. "Machine learning in automated text categorization," ACM computing surveys, vol. 34, 2002.

[2] Khreisat, L. "A machine learning approach for Arabic text classification using N-gram frequency statistics," Journal of Informatics, Volume 3, 2009.

[3] Jackson, P. and Moulinier, I. Natural language processing for online applications: text retrieval, extraction, and categorization. John Benjamins Publishing Company, Amsterdam, 2002.

[4] Khan A., Baharudin B., Lee L.H., khan K. "A Review of Machine Learning Algorithms for Text-Documents Classification," Journal of Advances in Information Technology, vol. 1,2010.

[5] Mesleh, A.M.A. Support Vector Machine text Classifier for Arabic Articles: Ant Colony Optimization-based Feature Subset Selection., The Arab Academy for banking and financial Science, Ph.D.Thesis, 2008.

[6] Sebastiani, F. (2005). Text classification. Idea Group Inc 2005, [Online] pp. 683-687. Available at: http://nmis.isti.cnr.it/sebastiani/Publications/EDTA05.pdf. (Accessed 2015-08-20.)

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

[8] Mesleh, A.M., "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

[9] Caballero, Y., Bello, R., Alvarez, D., Garcia, M.M., " Two new feature selection algorithms with Rough Sets Theory," IFIP International Federation for Information Processing, a Springer Series in Computer Science, 2006.

[10] Khorsheed, M.S. and Al-Thubaity, A.O. "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," Springer Science + Business Media Dordrecht 2013.

[11] El-Halees, A.M. "Arabic Text Classification Using Maximum Entropy," The Islamic University Journal (Series of Natural Studies and Engineering), Vol. 15, 2007.

[12] Karima, A., Zakaria, E and Yamina, T.G. "Arabic Text Categorization: A Comparative Study of different Representation Modes," Journal of Theoretical and Applied Information Technology, Vol. 38, 2005.

[13] Duwairi, R. "Arabic Text Categorization," The International Arab Journal of Information Technology, Vol. 4, 2007.

[14] Ababneh, J., Almomani, O., Hadi, W., El-Omari, N.K.T., and Al-Ibrahim, A., "Vector Space Models to Classify Arabic Text," International Journal of Computer Trends and Technology (IJCTT), vol 7, 2014.

[15] Elhassan R. & Ahmed M. (2016B). Arabic Text Stemming: Comparative Analysis. In the Processing of: Conference of Basic Sciences and Engineering Studies (SGCAC), Khartoum. pp. 88-93. doi: 10.1109/SGCAC.2016.7458011.

[16] Saad, M.K. & Ashour W."Arabic Text classification using decision tree," Workshop on computer science and information technologies CSIT'2010, Russia, 2010

[17] El-Kourdi, M., Bensaid, A., and Rachidi, T., "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," 20th International Conference on Computational Linguistics, Geneva, 2004.

[18] Gharib, T.F., Habib, M.B. and Fayed, Z.T. "Arabic Text Classification Using Support Vector Machines," http://www.home.cs.utwente.nl/~badiehm/PDF/ISCA2009.pdf.

[19] Bawaneh, M.J., Alkoffash, M.S., and Al Rabea A.I."Arabic Text Classification using K-NN and Naive Bayes". Journal of Computer Science, vol. 4, 2008.

[20] Alsaleem, S., " Automated Arabic Text Categorization Using SVM and NB," International Arab Journal of e-Technology, Vol. 2, 2011

[21] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S. and Al-Rajeh, A. "Automatic Arabic Text Classification," Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, Lyon- France, 2008.

[22] Syiam, M. M., Z. T. Fayed & M. B. Habib. An intelligent system for Arabic text categorization. IJICIS, Vol.6, No. 1, 2006.

[23] Motaz K. Saadand Wesam Ashour OSAC: Open Source Arabic Corpora 6th International Conference on Electrical and Computer Systems (EECS'10), Nov 25-26, 2010, Lefke, North Cyprus.

[24] Wahbeh, A.H., Al-Kabi, M. "Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text," ABHATH AL-YARMOUK: "Basic Sci. & Eng." Vol. 21, 2012

[25] Al-Shargabi, B., AL-Romimah, W. and Olayah, F. "A Comparative Study for Arabic Text Classification Algorithms Based on Stop Words Elimination," ISWSA '11 Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, Article No. 11, 2011.

[26] Farghaly, A., & Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Language Information Processing (TALIP). 8(4). USA. ACM. pp. 1-22.

[27] Alhawary, M. (2011). Modern standard Arabic grammar: a learner's guide. USA. Wiley-Blackwell.

[28] Mustafa, Suleiman H. (2012). Word stemming for Arabic information retrieval: The case for simple light stemming. Abhath Al-Yarmouk : Science & Engineering Series 21. 21(1), pp. 123-144. Available at http://journals.yu.edu.jo/aybse/Issues/Vol21No1_2013/09.pdf. (Accessed 2015-06-5).

[29] Froud, H., Lachar, A. & Alaouiouatk, S. (2012). A comparative study of root-based and stem-based approaches for measuring the similarity between Arabic text mining applications. Advanced Computing: An International Journal (ACIJ). 3(6). pp. 55-67.

[30] Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research. 3, pp. 1157-1182.

[31] Liu, H., Abraham, A. and Li, Y. (2009). Nature inspired population-based heuristics for Rough set reduction. In: Abraham, A., Falcon, R. and Bello, R. (Eds.) Rough Set Theory: A True Landmark in Data Analysis. Studies in Computational Intelligence. (174). Berlin. Springer-Verlag Berlin Heidelberg. pp. 261-278.

[32] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics. 23(19), pp. 2507-2517.

[33] Wang, G., Lochovsky, F. & Yang, Q. (2004). Feature Selection with Conditional Mutual in MaxiMin in Text Categorization. In the Proceedings of the thirteenth ACM international conference on Information and knowledge management. USA. ACM. pp. 342-349.

[34] Witten, I. H., Frank, E. & Hall, M. A. (2011). Data Mining Practical Machine Learning Tools & Techniques. 3rd ed.

[35] Chantar, Hamouda K. (2013). New techniques for Arabic document classification. Ph.D Theses, Heriot-Watt University. Available at: http://www.ros.hw.ac.uk/bitstream/h&le/10399/2669/ChantarHKH_0913_macs.pdf?sequence=1&isAllowed=y.

[36] Abd-Alsabour, N. & R & Randall, M. (2010). Feature selection for classification using an ant colony system. In the processing of the E-Science 2010: Sixth IEEE International Conference on e.Science. Brisbane, Australia. IEEE computer society. pp. 86-91.

[37] Duch, W., Biesiada, J., Winiarski, T., Grudzinski, K. & Grabczewski, K. (2003). Feature Selection Based on Information Theory Filters. In: Rutkowski, L. & Kacprzyk, J. (eds.) Neural Networks and Soft Computing. pp. 173-178.

[38] Mesleh, A. (2011). Feature sub-set selection metrics for Arabic text classification. Pattern Recognition Letters. 32(14), pp. 1922–1929.

[39] Asch, V. (2013). Macro- and micro-averaged evaluation measures [BASIC DRAFT]. [Online]. Available at: http://www.cnts.ua.ac.be/~vincent/pdf/microaverage.pdf. (Accessed 2015-11-07).

[40] A. Jafar, et al. “Vector space models to classify Arabic text,” International Journal of Computer Trends & Technology. vol. 7, no.4, pp. 219-222, 4014.

[41] K. Tarek, & F. Edward, Automated Arabic text classification with Pstemmer machine learning, and a Tailored news article taxonomy, Journal of the Assoc. for Information Science and Technology [online]. Available: https://vtechworks.lib.vt.edu/handle/10919/51269.

[42] A. Saleh, (2011). Automated Arabic text categorization using SVM & NB, International Arab Journal of e-Technology. [Online]. Available: http://www.iajet.org/iajet_files/vol.2/no.2/Automated%20Arabic%20Text%20Categorization%20Using%20SVM%20&%20NB_doc.pdf.

L. Bing, Web data mining: Exploring Hyperlinks, Contents, and Usage Data. Berlin. Springer, 2007.