I.J.Modern Education and Computer Science, **2013, 3, 33-41**
Published Online April 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2013.03.05

EMCAR: Expert Multi Class Based on Association Rule

Wa'el Hadi
MIS Dept, Petra University, Airport Rd. ,Amman, Jordan
Email: whadi@uop.edu.jo

***Abstract*—** Several experimental studies revealed that expert systems have been successfully applied in real world domains such as medical diagnoses, traffic control, and many others. However, one of the major drawbacks of classic expert systems is their reliance on human domain experts which require time, care, experience and accuracy. This shortcoming also may result in building knowledge bases that may contain inconsistent rules or contradicting rules. To treat the abovementioned we intend to propose and develop automated methods based on data mining called Associative Classification (AC) that can be easily integrated into an expert system to produce the knowledge base according to hidden correlations in the input database. The methodology employed in the proposed expert system is based on learning the rules from the database rather than inputting the rules by the knowledge engineer from the domain expert and therefore, care and accuracy as well as processing time are improved. The proposed automated expert system contains a novel learning method based on AC mining that has been evaluated on Islamic textual data according to several evaluation measures including recall, precision and classification accuracy. Furthermore, five different classification approaches: Decision trees (C4.5, KNN, SVM, MCAR and NB) and the proposed automated expert system have been tested on the Islamic data set to determine the suitable method in classifying Arabic texts.

***Index Terms*—** Associative Classification, Arabic Text Classification, Data Mining

I. Introduction

In today's information-oriented society, the number of online documents has been growing exponentially. These documents conceal useful information that can be utilized by decision makers in their key business activities. The process of finding and generating the hidden and useful information from online documents manually by domain experts is hard and time consuming. This is due to the fact that the available amounts of online textual data are massive and having large dimensionality. Consequently, employing expert system tools to discover essential information automatically from textual documents grants companies to make right decisions that work for improving their competitive advantages.

Text mining (TM) is such a discipline that has been emerging to respond to the need of discovering useful knowledge from unstructured text data format [1]. Unlike, data mining which is mainly concerned about finding valuable patterns from highly structured data format [2]. Nevertheless, text mining derives a great deal of its inspiration from seminal studies on data mining. As a result, text mining and data mining applications demonstrate several high level architectural similarities [3]. There are many domains surrounding text mining including information retrieval, information extraction, summarization, clustering and classification [4].

Text Classification (TC) is one of the most significant domains of text mining that is responsible about understanding, recognizing, and organizing various types of textual data collections [5]. TC is concerned about predicting a specific category of incoming textual document. This type of prediction is called "supervised" learning in which the class category is predetermined within the input text collection [6]. TC is a multi phase process that includes, processing the textual documents, learning the document by an algorithm and evaluating the output models [7]. There are many classification approaches towards TC that have been adopted from data mining and machine learning, e.g. Decision trees, Naïve bayes, Support Vector Machine and Neural Network [3]. These approaches have been mainly investigated on classifying English documents [8]. But, researchers paid little interest for applying these approaches to other languages such as Arabic.

A few researchers have applied a number of classification approaches which are solely applicable for the problem of Arabic text classification, i.e. Naïve Bayes [9], Support Vector Machine (SVM) and Decision Tree [10]. However, researchers conclude that Arabic text classification is a very challenging task due to language complexity. For instance in Arabic morphology, words have affluent meanings and contain a great deal of grammatical and lexical information. In syntactic structure, Arabic sentence formation differs from English. In this regard, the Arabic text documents are required extensive pre-processing routines to build accurate classification model.

Most of the previous works on Arabic text classification attempt only to achieve the classification accuracy from the above mentioned learning approaches.

Some of domain experts have also interests in understanding the produced models that are formed as "IF-Then" patterns [6] [11]. This is since such models are easy to interpret and can be modified by users manually [4]. Thus, this paper is devoted to build an expert system on the problem of Arabic text classification. Primarily, Expert Multi Class based on Association Rule (EMCAR), SVM, NB, C4.5, and KNN learning methods are applied to Islamic Arabic data collection to measure their performance and effectiveness with reference to different text evaluation metrics such as error rate, precision and recall.

The ultimate aim of this research is to compare various rule based classification data mining algorithms using WEKA business intelligence tool for Arabic text documents and EMCAR was implemented using VB.net under MS windows platform. Text classification is one of the most significant problems in text mining as well as knowledge discovery. This problem can be considered as a large and complex because of the data being enormous and having a large dimensionality. Another primary aim along with the experimentations and evaluation is a comprehensive literature review on the problem of text classification within the context of text mining including related works to Arabic text classification.

The literature review is introduced in section II. The proposed algorithm main steps are presented in section III, and the experimental results are given in Section IV. Finally the conclusions are depicted in Section V.

II. Literature Review

As [12] pointed out that there are over 320 millions Arabic native speakers in 22 countries located in Asia and Africa. Due to the enormous energy resources, the Arab world has been developing rapidly in almost every sector especially in economics. As a result, a massive number of Arabic text documents have been increasingly arising in public and private sectors, where such documents contain useful information that can be utilized in a decision making process. Therefore, there is a need to investigate new intelligent methods in order to discover useful hidden information from these Arabic text collections.

Reviewing the existing related works proved that there are several methods which have been proposed by researchers towards Arabic text classification. For classifying Arabic text sources the N-Gram Frequency Statistics technique is investigated by [13]. This method is based on both Dice similarity and Manhattan distance measures in classifying an Arabic data set. For this research the Arabic data set was obtained from various online Arabic newspapers. The data is associated with four categories. After performing several pre-processing on the data, and experimentation, the results indicated that the "Sport" category outperformed the other categories with respect to recall evaluation measure. The least category was "Economy" with around 40% recall. In general the N-gram Dice similarity measure figures outperformed that of Manhattan distance similarity.

A modified version of Artificial Neural Network (ANN) method is proposed for classifying Arabic texts by [14]. The authors have used a Singular Value Decomposition (SVD) for data representation, which is a new representation space of the observations. A collection of Prophet Mohammad's 'Peace Be Upon Him' Hadeeth was collected from the "Nine Hadeeth Book". The data consists of 453 documents that are associated with fourteen categories. A comparison between the proposed method (ANN with SVD) and the original ANN was carried out against the Arabic data set with reference to F1 evaluation measure. The results revealed that (ANN with SVD) outperformed the classic ANN when the dimensionality increased.

Decision trees classification approach and the effect of feature selection on the predictive accuracy were applied to the problem of Arabic text mining in the work of [10]. Specifically, the authors have compared ID3 algorithm with the known statistical method of Naive Bayes for two Arabic data sets collected from Arabian scientific encyclopedia (Hal Taalam) and Prophet Mohammad's 'Peace Be Upon him' Hadeeths "Nine Hadeeth Book". The initial results indicated an improvement on average 10% and 26% on the "Hadeeth" and "Scientific" data sets respectively when employing feature selection instead of the whole data set. Moreover, the F1 has also improved by around 2.5% when using decision trees over that of Naïve Bayes.

In [15] a probabilistic supervised learning method called the Maximum Entropy method has been applied on real Arabic data set collected from Aljazeera news website. The author has tested the proposed method with and without pre-processing phase with regards to F1 evaluation measure. The results have shown that the F1 accuracy has increased from 68% to 84% proving that removing noise (stopwords, tokenization, stemming, etc) definitely improve the classification accuracy in the selected Arabic data set.

The performance of three classification data mining algorithms (KNN, Naive Bayes, Distance based learning) have been evaluated with respect to error rate, Recall, and Precision measures on Arabic data collection by [16]. The benchmark used in the experiment consists of 1000 documents and 10 categories. The author has processed the data before performing the training where punctuation marks and stopwords have been removed. The obtained results have shown a variety of the performance of the categories. For instance, "Internet" category achieved only 22% Recall, whereas "Economic" category achieved 98% Recall. Overall, the author has concluded that Naïve Bayes algorithm outperformed the other considered classifiers with respect to the above mentioned measures.

In [17] the performance of two common data mining approaches mainly Support Vector Machine and

decision trees (C5) have been evaluated on different Arabic benchmarks. The data which consists of 17658 documents have been gathered from different sources, especially Saudi Press Agency, Saudi newspapers, Internet articles, and discussion forums. The experimental tests have revealed that C5 algorithm achieved on average 10% more in accuracy than SVM on the considered data sets.

Most scholars [16] [14] [13] considered classifying Arabic text documents as a hard task. This is due to complexity and richness of the Arabic morphological analysis [15].

In general, most of the current research works conducted on Arabic text mining are just simple comparison studies. These researches mainly focused on adapting some of the data mining and machine learning approaches such as probabilistic, decision trees and SVM which are solely designed for English text collections to that of Arabic data sets. For instance, [17] have tested only two data mining methods and [15] has examined the performance of one probabilistic method with and without pre-processing.

Moreover, the lack of standardized published Arabic data sets is also unavailable or rare. Such works can be used as key data sets for researchers in related fields to compare the results. In fact, most of the related research articles have obtained data from online newspapers and websites. Such works usually do not publish their data for other researchers to utilize. Consequently the confidence in the results derived from such experimental studies is not high enough. Furthermore, the performance of the adopted data mining approaches is biased to such data sets and sometimes ambiguous. For example, the research conducted by [17] showed that decision tree algorithms outperformed the SVM with respect to classification accuracy. However, the majority of international research on English text mining proved that SVM is the best machine learning approach [18].

In general, comprehensive experimental and critical research studies that cover most of the common business intelligence techniques are rare. This is one of key motivations for this research work. In addition, this paper aims to investigate four different learning methods in data mining and machine learning. Furthermore, the work proposes to establish the performance on published Islamic Arabic data set collection. Future researchers might get access and utilize it in their experiments and compare their derived results with this work.

III. The Proposed Model

AC mining is a new classification approach in data mining that have been studied extensively in the last decade by many scholars in real world domains including medical diagnoses, bioinformatics, website and email phishing, English text categorization and others. The main reasons behind the popularity of this approach are due to a) the high predictive rate of the resulting classifiers and b) The simplicity of the rules contained within the classifier which are represented in a simple chunks of knowledge "If-Then" rules. Nevertheless, there is a drawback associated with AC mining approach and in particular the exponential growth of rules which can be resolved when appropriate pruning are plugged in the classifier construction step. Overall, the ultimate goal of an AC mining algorithm is to build a classification system, known as a classifier, from a labeled historical data set known as training data set in order to forecast the type / label of unseen data set known as test data set.

There have been a number of common AC algorithms in the literature that have been developed in several real world domains. Some of these common algorithms are MAC [19] CPAR [20], CACA [21], MCAR [22], Lazy [23] and others. Though, AC mining has not yet been explored in research field of Arabic TC. In fact, there are some attempts tackling the problem of Arabic textual classification mainly employed association rule mining [15]. It is the firm believe of the authors that AC mining will be high promising approach to the complex and hard problem of Arabic TC.

In general, an AC algorithm must go through two main steps where in the first step frequent ruleitems are discovered. A ruleitem is simply an attribute value plus the class attribute. The AC algorithm finds frequent ruleitems based on a threshold known as minimum support which is inputted by the end-user. A frequent ruleitem is simply a ruleitem that has a frequency in the training data set above the minimum support threshold. Once all frequent ruleitems are discovered, the AC algorithm evaluates their confidence values and produces ruleitems that hold enough confidence into Class Association Rules (CARs). Meaning, any frequent ruleitem that has a confidence value greater than or equal to the inputted minimum confidence is produced as a rule.

Once the complete sets of rules are derived, the algorithm ranks them according to particular parameters mainly confidence and support. Then, the algorithm chooses the most predictive rules as a classification system (classifier) from the complete set of discovered rules using pruning procedures. Lastly, the classifier is evaluated on an independent data set to measure its predictive rate and this step is called the prediction step and the output of this step is the prediction accuracy or error rate.

Arabic is considered a popular language in the United Nation since over twenty countries worldwide speak it as a native language and several countries speak it as a seconded language. Since the massive development in several sectors including commercial, manufacturing, oil, etc in the Arab world, very large numbers of offline and online documents are now available. These documents contain important knowledge that decision makers can benefit from in management related decisions. Therefore, applying data mining techniques are vital to discover

this hidden knowledge which latterly can be utilized in planning decisions. Though, the nature Arabic especially the morphological analysis requires an intelligent technique that can be competitive to complex mathematical approaches that are currently in use such as SVM and NN. Therefore, AC mining seems to be promising approach in mining Arabic text data sets because of its simplicity and high predictive power of its outputted classifiers.

In this paper, an AC algorithm named "Expert Multi-Class Association Rule" (EMCAR) is proposed which firstly allow the end-user ends up with controllable numbers of rules which he/s can better understand and maintain them. Secondly, unlike MCAR algorithm which uses a single rule for prediction, the proposed algorithm employs a new class assignment method which ensures that only high quality rules are used to predict test cases. This class assignment is based on a group of rules prediction rather than single rule, and therefore multiple rules are used to contribute to class assignment. This may enhance the classification accuracy of the resulting classifiers.

MCAR algorithm is presented in section 3.1. The expert algorithm (EMCAR) is presented in Section 3.2 where details about rule discovery, rule pruning (Knowledge base builder) and class assignment of test cases are discussed.

A. Multi-class Classification based on Association rule (MCAR)

In this section, we explain MCAR algorithm [22] in details since the proposed Arabic text categorizer is based on it. MCAR is an AC classification algorithm that was developed in 2005 by [22], and is considered the first AC algorithm that uses fast intersection method for rule discovery based on the concept of vertical mining. This algorithm constitutes of multiple phases where the first phase is optional and the algorithm checks whether the input training data set contain continuous attribute and if so MCAR invokes Entropy based discretization method. Once this done, MCAR utilizes TID-List intersection method for frequent ruleitems discovery. A TID-List of a ruleitem (attribute value, class) contains the locations of ruleitem in the training data set along with the locations of its associated class labels. In other word, A TID-List is simply a data structure that stores the appearances of an attribute value and the class attribute in the input data set. This data structure is very useful when it comes to computing the support and confidence of the ruleitem and thus saves resources associated with time and memory usage [22].

Consider for example two ruleitems as follows (A1), Class1 and (K2), Class 1 and assume that these ruleitems have the following TID-Lists (1,3,4,7,11,15,22) and (2,4,11,15,16,18,21,25) respectively. Furthermore, assume that the minimum support is 3 meaning these two items are frequent ruleitems of size 1 since each of them contains a single attribute value in its antecedent (right-hand-side). Now, to validate whether the new candidate ruleitem (A1,K2), Class1 is frequent, MCAR algorithm simply intersects the TID-Lists of frequent ruleitems of size 1, e.g. (A1), Class1 and (K2), Class 1 in order to determine whether the candidate ruleitem of size 2 is frequent. So, (1,3,4,7,11,15,22) gets intersected with (2,4,11,15,16,18,21,25) and the resulting TID-List (4,11,15) is actually the locations of ruleitem (A1,K2), Class1 in the training data set. Then taking the cardinality of this set we can determine that this ruleitem has support (3) which is greater than or equal to the minimum support (3) and thus this ruleitem is frequent.

The rule discovery method described earlier is very simple if compared with other AC mining method such as CBA that necessitates multiple training data set scans and consumes more time and memory. In fact, MCAR rule discovery method requires only one single data scan and then performs simple intersection between the TID-Lists of ruleitems of size N-1 to generate candidate ruleitems of size N. Once all frequent ruleitems are discovered, MCAR algorithm generates the subset of those which hold larger confidence than the minimum confidence threshold as rules. When all rules are generated then the algorithm applies a ranking procedure to favor rules over each other. The basis of this rule favoring procedure is mainly the confidence value, and then support value and lastly the size of the rules (number of attributes values in the rule body). If two or more rules having similar confidence, support and rule size then the rank will be random.

Once all rules are sorted, then MCAR uses the database coverage pruning to remove redundant rules from taking any role in the prediction step. More details on the database coverage pruning can be found in [22]. The output of the pruning is the subset of rules that are high predictive and those represent the classifier. Once the classifier is produced its predictive power is tested using cross validation or on test data set. The prediction procedure of MCAR works as follows: Given a test data case, the algorithm goes over the set of rules starting from the highest ranked rule and selects the rule that its body matches test data case and assigns its class to the test data case. The outcome of the prediction phase of MCAR is the error rate which is simply calculated by dividing the number of correct classification in the test data set by the size of the test data set.

B. Expert Multi Class based on Association Rule

An Expert Multi Class based on Association Rule (EMCAR) goes through three main phases excluding the preprocessing phase which is optional: training, construction of the knowledge base, and forecasting of new cases as shown in Fig. 1. It should be noted that there is an optional step called data preprocessing in cases the input data is unstructured text collection that requires processing. During the first phase, it scans the input data set to find frequent items in the form <AttributeValue, class> of size 1. These items are called

frequent one-items. Then the algorithm repeatedly joins them to produce frequent two-items, and so forth. It should be noted that any item that appears in the input data set with a frequency less than the MinSupp threshold gets discarded.

Once all frequent items of all sizes are discovered, then the EMCAR algorithm checks their confidence values in which those hold a confidence value larger than the MinConf threshold become a class association rule (CAR). Otherwise, the item gets deleted. Therefore the complete set of CARs represents items in the training data set which are statistically representative and hold high confidence values. The next step is to choose a subset of the complete set of CARs to form the knowledge base. The proposed algorithm treats categorical attributes.

Data used by the proposed algorithm contain a header that indicates file name, attribute names, and a number of training cases. Values for each training data case are comma-separated, and the class attribute must be the last column in the header file. Details on the EMCAR which involves knowledge base construction, and forecasting of test cases, are given in the next subsections.

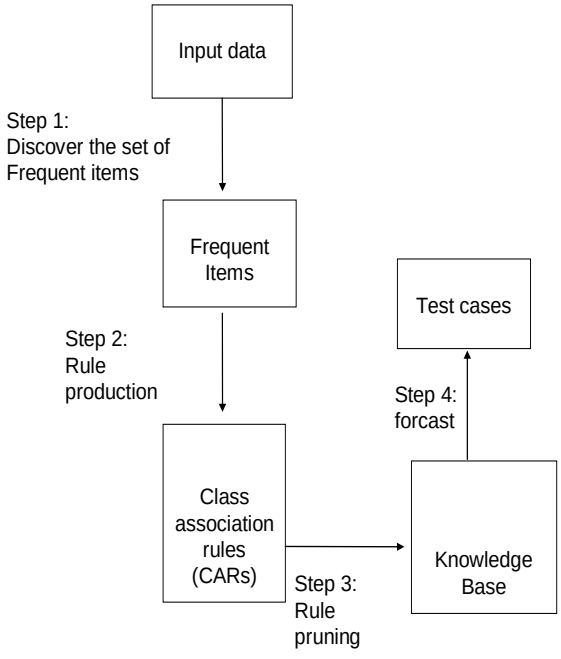

Figure 1. EMCAR Main Steps Adopted From (Thabtah, 2006) and Modified

C. Building the Knowledge Base

As mentioned earlier, one primary limitations of the AC approach in data mining is the exponential growth of rules [23] [22], and thus a primary motivation of this research is to generate a knowledge base with concise set of rules that end-user can control and understand easily. Prior to prune redundant rules and to build the knowledge base, rules must be sorted in order to give higher quality rule better priority to be chosen as part of the knowledge base. The proposed algorithm sorts the rules according to the following guidelines:

1) The rule with higher confidence is placed in a higher rank
2) If the confidence values of two or more rules are the same, then the rule with higher support gets a higher rank
3) If the confidence and the support values of two or more rules are the same, the rule with less number of attribute values in the antecedent gets a higher rank
4) If all above criteria are similar for two or more rules, then the rule which was produced first gets a higher rank

Building the knowledge base procedure in EMCAR is described as follows:

For each sorted rule (CAR) in a sequential manner, the EMCAR starts with the first one and applies it on the training data set; the rule gets inserted into the knowledge base if it covers at least one case regardless of the rule class similarity to that of the training case. In other words, the highest confidence rules is tested on the training cases, all training cases that are similar to the rule body are marked for removal and the rule gets inputted into the knowledge base. Once a rule gets inserted into the knowledge base, all training cases associated with it are discarded. In situations where a rule fails to cover any training case then it will be removed. The same process is repeated on the remaining rules in order until no more cases remains in the training data set or all rules are tested. The evaluation procedure in our algorithm guaranteed that only high confidence and quality rules are remains for prediction.

This may result in more accurate prediction on the training data set but not necessarily on new unseen test cases.

D. Prediction Method

In predicting test data case, the prediction method of EMCAR divides all rules which match the test case into groups one for each class label, and then counts the numbers of rules for each group. Lastly, it assigns the test case the class of the group with largest count. In cases where there are two or more groups with similar count, the assignment of the class to the test case is random. Unlike other current AC methods like MCAR which employ only the highest confidence rule for predicting the test case, our algorithm makes the prediction decision based on multiple rules, which is considered by previous research studies on prediction methods, i.e. [24] [22] an advantage since multiple high confidence and support rules contributed to the assignment decision. Finally, in cases when no rules in the classifier are applicable to the test case, the default class (majority class in the training dataset) will be assigned to that case.

IV. Experimental Results

The data used in our experiments is The Islamic data set, the data set consist of 2244 Arabic documents of different lengths that belongs to 5 categories, the categories are (Hadeeth "الحديث" , Aqeedah "العقيدة" , Lughah "اللغة" , Tafseer "التفسير" , Feqah "الفقة"), Table 1 represent the number of documents for each category.

Generally, TC task goes through three mainly steps: Data preprocessing, text classification and evaluation. Data preprocessing phase is to make the text documents suitable to train the classifier. Then, the text classifier is constructed and tuned using a text learning approach against from the training data set. Finally, the text classifier gets evaluated by some evaluation measures i.e recall, precision, etc. The next two sub-sections are devoted to discuss the main phases of the TC problem related to the data we utilized in this paper.

TABLE I. NUMBER OF DOCUMENTS PER CATEGORY

Category Name	Number of Documents
Hadeeth	462
Aqeedah	536
Lughah	434
Tafseer	388
Feqah	424
Total	2244

A. Data preprocessing

For the Islamic Arabic text collection, the data is organized with reference to document categories in which each document is stored as a separate text file in its related category folder. Each document is represented in a numerical vector where terms in the document correspond to numerical values according to their frequency in that document by utilizing String-To-Vector method in the WEKA tool. To conduct preprocessing operations such as stemming and stopwords elimination, approaches from [9] and [12] have been adopted. This includes the following stages:

1) Each document in the Islamic data set is processed to discard the numerical data as well as punctuation marks.
2) All the non-Arabic texts and function words are deleted.
3) The documents in the Islamic data set are stemmed in which all Arabic word derivatives are transformed into their single common root.
4) The non useful Arabic frequent list of words (stopwords) was eliminated from the Islamic data.

B. Results Discussion

Table 2 represents the average precision and recall results for the selected classification algorithms. The figures in that table give a clear indication that EMCAR algorithm outperformed the remaining classification algorithm on the Islamic data set. In particular, EMCAR achieved more precision on average 11.4%, 21.2%, 10.2% and 0.9% respectively than C4.5, KNN, NB, and SVM algorithms. Additionally, this algorithm gained respectively more recall on average 11.5%, 26.2 %, 10% and 1% than C4.5, KNN, NB and SVM algorithms. Table 2 also demonstrates that KNN algorithm is the least applicable classification approach towards the Islamic Arabic data set due to the low results of precision and recall. On the other hand, C4.5 algorithm produced 24 rules that represent most of the classes in the training data set. This method correctly covered 1897 out of 2244 documents. This means C4.5 learning approach is somehow not impacted with the unbalanced categories. In general, all classification algorithms except KNN showed very competitive performance with regards to precision and recall on the Islamic data set, as their generated results are very close to each other.

TABLE II. THE AVERAGE PRECISION AND RECALL RESULTS ON THE ISLAMIC DATA SET

Classification Algorithm	Average Precision	Average Recall
C4.5	0.850	0.848
KNN	0.752	0.701
NB	0.862	0.863
SVM	0.955	0.953
EMCAR	0.964	0.963

The confusion matrices for the selected algorithms in the experimentation have been derived as shown in Fig. 2. The confusion matrix in the Fig. 3 represents the distribution of documents for each class in the training data set. For instance, for the C4.5 (J48) algorithm, class “Aqeedah” have correctly covered 441 documents and incorrectly covered 95 documents (32 to “Feqah", 43 to “Hadeeth”, 5 to "Lughah" and 15 to “Tafseer”). On the other hand, for EMCAR algorithm, class “Tafseer” represents 388 documents in the training data set in which 382 of them are classified correctly and 6 of them are classified incorrectly (4 by class “Aqeedah”, 1 by "Feqah" and 1 by “Hadeeth”). In general, the confusion matrix is a helpful evaluation measure that reveals the performance of class labels with regards to the available documents in the training data set. In other words, it displays the number of documents covered correctly by the class and the number of documents incorrectly classified to other classes.

Fig. 3 depicts the error rate produced by the selected classification algorithms for the Islamic data set. This Fig. reveals that EMCAR and SVM have almost a similar error rate. However, SVM algorithm achieved a slightly less error rate by (1%) than EMCAR. Further, the error rate of KNN algorithm on the Islamic data set is very high (30.7%) which definitely makes this algorithm the least applicable to such data set.

An intensive analysis has been conducted on the performance of each selected classification approach for the document categories. Table 3 represents precision

and recall evaluation results for KNN, C4.5 SVM, NB, and EMCAR algorithms respectively. As it is shown in table 6, there are two categories "Aqeedah" and "Tafseer" that achieved nearly 100% prediction accuracy. All categories' performance for EMCAR and SVM algorithms with reference to precision and recall measures are Excellent. Further, almost similar performance to NB algorithm has been achieved by the C4.5 algorithm document categories. Lastly, the "Hadeeth" category produced less precision and recall results than the rest of the categories in most of the algorithms. This is due to the fact that this category is the highly overlapped with other categories.

C4.5

a	b	c	d	e	<-- Classified as
441	32	43	5	15	| a = Aqeedah
41	324	37	16	6	| b = Feqah
41	14	376	29	2	| c = Hadeeth
14	7	12	397	4	| d = Lughah
16	4	2	7	359	| e = Tafseer

KNN

a	b	c	d	e	<-- classified as
277	16	199	34	10	| a = Aqeedah
3	320	67	33	1	| b = Feqah
0	63	354	45	0	| c = Hadeeth
0	94	0	340	0	| d = Lughah
3	45	28	47	265	| e = Tafseer

SVM

a	b	c	d	e	<-- classified as
526	0	8	1	1	| a = Aqeedah
11	384	22	4	3	| b= Feqah
11	21	425	5	0	| c = Hadeeth
2	1	3	426	2	| d = Lughah
7	0	2	1	378	| e = Tafseer

NB

a	b	c	d	e	<-- classified as
512	3	10	0	11	| a = Aqeedah
10	300	103	3	8	| b = Feqah
31	20	401	1	9	| c = Hadeeth
3	11	26	394	0	| d = Lughah
26	6	18	9	329	| e = Tafseer

EMCAR

a	b	c	d	e	<-- classified as
528	0	6	1	1	| a = Aqeedah
4	395	20	3	2	| b = Feqah
9	8	437	5	3	| c = Hadeeth
3	4	5	420	2	| d = Lughah
4	1	1	0	382	| e = Tafseer

Figure 2: The Confusion Matrices Produced By Selected Classifiers

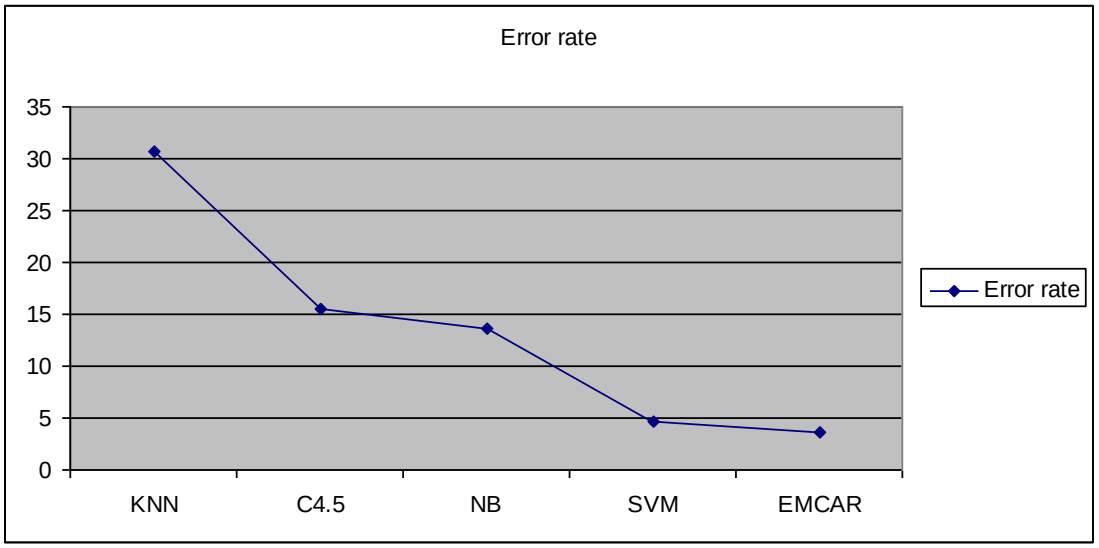

Figure 3: Error Rate Produced By Selected Classifiers

TABLE 3: PRECISION AND RECALL RESULTS PER CLASS FOR THE CLASSIFICATION ALGORITHMS

Category/Class	KNN		C4.5		SVM		NB		EMCAR	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Aqeedah	0.979	0.517	0.797	0.823	0.944	0.981	0.88	0.955	0.964	0.985
Feqah	0.595	0.755	0.85	0.764	0.946	0.906	0.882	0.708	0.968	0.932
Hadeeth	0.546	0.766	0.8	0.814	0.924	0.92	0.719	0.868	0.932	0.946
Lughah	0.681	0.783	0.874	0.915	0.975	0.982	0.908	0.937	0.979	0.968
Tafseer	0.96	0.683	0.93	0.925	0.984	0.974	0.922	0.848	0.979	0.985

V. CONCLUSIONS

Text mining is becoming increasingly important because of the huge number of documents available offline and online. Text classification is one of the significant tasks in the field of text mining. This task involves classifying text documents into a number of predefined classes based on their content. In this paper, the problem of Arabic text classification is investigated using different classification learning algorithms (C4.5, KNN, NB, SVM and EMCAR) from data mining and machine learning. WEKA, the open business intelligence tool, is employed in order to test the performance of these algorithms for the published Islamic Arabic data set and the EMCAR is implemented using VB.Net programming language. The basis of the comparison in the experimentation are different text evaluation metrics, including error-rate, precision, and recall. The results indicated that the least applicable learning algorithm towards the chosen Arabic data set is KNN. Moreover, the most applicable algorithm to the Arabic data set is EMCAR in which it derived higher results in all evaluation criteria than SVM, NB and C4.5, respectively. Further, the confusion matrix which represents the distribution of documents per category is derived for all the learning algorithms. The confusion matrices for the learning algorithms indicate that the "Hadeeth" category achieved the least results with respects to precision and recall.

REFERENCES

[1] Kroeze, J. Matthee, M. & Bothma, T., (2003) 'Differentiating between data-mining and text mining terminology', ACM: Proceeding of the 2003 annual research conference of the south African institute, Vol. 47, PP.93-101.

[2] Weiss, M., S., Indurkhya, N., Zhang, T., & Damerau, F., (2005) Text mining: predictive methods for analyzing unstructured information. Springer Science Inc.

[3] Feldman, R. & Sanger, J., (2007) the Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, NY: Cambridge University Press

[4] Song, M. (2009) Handbook of research on text and web mining technologies, information science reference, IGI global, pp. 1-22.

[5] Guo, Y., Shao, Z. & Hua, N. (2010) 'Automatic text categorization based on content analysis with cognitive situation models', Information Sciences 180, pp. 613–630

[6] Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons

[7] Shi, G. & Kong, Y. (2009) Advances in Theories and Applications of Text, Mining, IEEE: ICISE'09, pp. 4167 – 4170.

[8] Tan, A., H. (1999) 'Text mining: the state of the art and the challenges' ,Proceeding Of The Pakdd Workshop On Knowledge Discovery From Advanced Databases,. PP. 65-70

[9] Elkourdi M., Bensaid A. and Rachidi T. (2004) 'Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm', ACM Publication: Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, pp: 51-58

[10] Harrag, F. El-Qawasmeh, E. & Pichappan, P. (2009) 'Improving arabic text categorization using decision trees', IEEE, NDT '09, pp. 110 – 115

[11] Sebastiani, F. (2002) 'Machine learning in automated text categorization' ACM Publication: ACM Computing Surveys. Vol. 3(1) : pp.1-47.

[12] Syiam, M. M., Fayed, Z. T. & Habib, M. B. (2006) 'An Intelligent System For Arabic Text Categorization', IJICIS, Vol.6, No. 1

[13] Khreisat, L. (2006) 'Arabic Text Classification Using N-Gram Frequency Statistics: A Comparative Study', Proceedings of the 2006 International Conference on Data Mining, pp. 78-82.

[14] Harrag, F. & El-Qawasmeh, E. (2009) 'Neural Network for Arabic Text Classification', The Second International Conference on the Applications of Digital Information, London, UK, pp.805-810, Aug. 4-6, 2009.

[15] El-Halees, A. M.(2007) 'Arabic Text Classification Using Maximum Entropy', The Islamic University Journal, Vol. 15, No.1, pp 157-167.

[16] Duwairi, R. (2007) 'Arabic Text Categorization', International Arab Journal of Information Technology, Vol.4, No.2, pp 125 – 131.

[17] Al-Harbi, S. (2008) 'Automatic Arabic Text Classification', JADT'08: 9es Journées internationales d'Analyse statistique des Données Textuelles., pp. 77-83

[18] Joachims, T. (2001) 'A Statistical Learning Model of Text Classification for Support Vector Machines', SIGIR'01, pp. 1 – 9.

[19] Abdelhamid N., Ayesh A., Thabtah F., Ahmadi S. and Hadi W. (2012) MAC: A Multiclass Associative Classification Algorithm. Journal of Information and Knowledge Management 11(2): (2012)

[20] Yin X. and Han J. (2003) CPAR: Classification based on predictive association rule, Proceedings of the SDM (2003) pp. 369–376.

[21] Tang Z. and Liao Q. (2007). A New Class Based Associative Classification Algorithm. IMECS 2007: 685-689.

[22] Thabtah, F., Cowling, P., and Peng, Y. (2005) MCAR: Multi-class classification based on association rule approach. Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications (pp. 1-7).Cairo, Egypt.

[23] Baralis E., Chiusano S. and Garza P. (2008). A Lazy Approach to Associative Classification. IEEE Trans. Knowl. Data Eng. 20(2): 156-171.

[24] Li W., Han J. and Pei J. (2001). CMAR: Accurate and efficient classification based on multiple-class association rule. Proceedings of the ICDM'01, (pp. 369-376). San Jose, CA.

Dr. Wa'el Hadi is a lecturer in Management Information Systems at the Department of MIS, Petra University. In research, his current interests include data warehousing, data mining, and knowledge management. Dr. Wa'el received his PhD degree in Computer Information Systems from the Arab Academy for Banking and Financial Sciences, Amman, Jordan.

