Arabic Text Classification review

Rasha Elhassan[1] and Mahmoud Ahmed[2]

[1] Information Technology Department, Garden city collage, Khartoum, 11111, Sudan
[2] Dean of Mathematical Science collage, University of Khartoum, Khartoum, 11111, Sudan

[1]rashamamoun@yahoo.com, [2]mali@uofk.edu.sd

ABSTRACT

A millions of the documents are available free and online. These documents must be first organized systematically for its proper utilization to make a decision from it. There are a lot of applications that help in organizing the documents. Text classification is deal with how the document belongs to its suitable class or category. Arabic language is richness and a very complex inflectional language which makes ordinary analysis a very complex task. This paper focuses on the published research in the field of Arabic text classification and presents a scientific view about the process of it and camper the evaluation of text classification techniques that were used.

Keywords: *Text classification, Arabic Text Classification, Data Preprocess, Classifier.*

1. INTRODUCTION

Text classification is a common technique in the data mining and it utilized to give useful information from the large amount of data. Recently there is a dynamically growing in the information, in various fields on the Internet, which made the classified of the text one of the most important research issues in the data mining. Text classification is to labeling documents with thematic categories from a predefined set known in advance [1]. Text classification can define as the process of classifying or structuring documents into a predefined set of categories according to a group structure that is known in advance [2]. The terms text categorization is sometimes 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, like sorting by author, by publisher, or by language [3].

The main goal of text mining or classification is to extract the information with value from unstructured textual resources. Also to deals with the operations like, retrieval, classification and summarization [4]. There are two approaches to the text classification: rule base 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].

Text classification have been used in many applications such as e-mail filtering and routing; news monitoring; spam filtering and searching for interesting information on web sites [7].

Arabic language has rich morphology and a complex orthography [8]. This may decreased the research in the area of Arabic text classification [9]. There are 28 letters in Arabic language and the language is written from right to left. It has two genders: feminine and masculine. Numerical there are singular, dual, and plural numbers. Grammatical there are three cases: nominative, accusative, and genitive. A noun has three grammatical cases: nominative case when it is subject; accusative case when it is the object of a verb; and the genitive case when it is the object of a preposition [10].

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]. The size of feature vectors may increases by richness of the language and that make task of feature selection more important to avoid the irrelevant feature [11].

There are three main phases in the text classification task: (a) Data pre-processing, (b) text classification and (c) evaluation [12, 13].

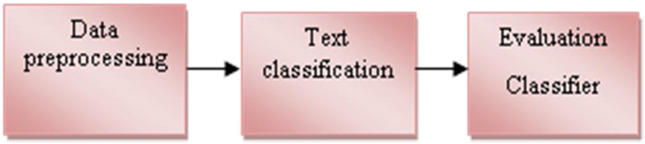

Fig. 1. Text Classification Phases

There are many standard data sets for English text classification that available free but for Arabic text classification unfortunately there is no free standard data set [14]. Almost of the researchers in the field of Arabic text classification collected their own corpus from the online web sites [15]. The range of the selected documents that presented in the data set was from 242 texts divided into six classes [16] up to 17,658 texts divided into one to seven classes in each one [13].

The paper is organized as follows. Section 2 presents a brief summary of the previous work in the field of Arabic text classification follow by section 3 which describe the steps of text classification phases with references to the Arabic text classification literature. Section 4 presented the results of the literature and at the last section 5 and 6 presented the conclusion and future work was presented.

2. PREVIOUS WORK

Previous work on Arabic text classification has used has used Decision Trees [17], Naïve Bayes [18], Support vector machine [19]. [20] compared performance of three classifiers for Arabic text categorization: Naïve Bayes, K-Nearest Neighbor and Distance–Based. The recall, precision, error rate and fallout are using to compare the accuracy of the classifiers. In-house Arabic corpus that consists of 1000 documents, which divided in 10 categories, as collected to experiment. The data was preprocessed by removed the stop word and extracted the root of the word. Each document is then represented as a vector of words. Stemming word was used to reduce the dimensionality of feature vectors of documents. Accuracy of classifier expressed in term of Precision, Recall, Fallout and error rate. The results of the experimentation show that the Naïve Bayes classifier outperforms the other two (more than 95%).

In [21] the authors evaluated the use of K-NN and Naive Bayes to classify the Arabic text. The k-fold cross-validation method is used to test the accuracy. One of difficulties that noted by authors was to found Arabic corpus for evaluating text categorization. They used in house corpus consists of 242 documents which belong to six categories. Another difficult was the huge number of features or keyword found in Arabic language. They extracted feature set of keyword to improve the performance. The effectiveness of KNN started to decline at k>15. The result showed that K-NN classifier had better performance. The average accuracy reported was about 84% to K-NN, while Naive Bayes classifier achieved 73%.

[22] 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.

By using a recognized statistics technique [13] classified Arabic text documents on seven different Arabic corpora. A tool was implemented for feature extraction and selection and the performance of two popular classification algorithms, (Support vector machine and C5.0, has been evaluated on classifying Arabic corpora. C5.0 classifier gives better accuracy.

An up-to-date review of the work done in the field of Arabic text classification was presented by [23]. 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 TEXT CLASSIFICATION PROCESS

The process of Arabic text classification contains in general three phases. Each phase can divided into sequence steps:

3.1 Data Pre-processing

This phase is necessary to clean the texts by removing worthless information which is of lower value to the text [16]. In this phase all non Arabic word, digits, numbers, and punctuation marks are remove [5, 10, 14], also removing the stop word, which have not any meaning in the content of the document, like (so لذلك, and for بالنسبة). 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 “15 ,5] ”ى].

After removing unnecessary word from the documents the extraction of the word root [8] can done by using stemming techniques. The root can be extracted by three ways: the root-based stemmer; the light stemmer; and the statistical stemmer [19]. Next the documents are indexed and represented as a vector of weighted terms [5, 19]. Some authors included the feature selection in the step of data preprocessing. [8] investigated the performance of five feature selection and found that CHI performed most effective.

The following table (1) summarized the step of data preprocessing in the Arabic text classification literature and give information about the corpus witch used by the authors:

Table 1: corpus and data preprocessing

Refer.	Corpus Type.	Documents no.	Categories no.	Data preprocessing
[5]	In House	1545	9	Remove digits, numbers, and punctuation marks, non Arabic letters, normalization some Arabic letters
[10]	In House	1132	6	Root-based, light and statistical stemmer
[12]	Saudi Newspapers	5121	7	Remove digits, numbers, and punctuation marks, non Arabic letters, normalization some Arabic letters.
[19]	In House	1132	6	Stop word remove and stemming (extract the root)
[13]	seven genres	17,658	1 to 7	Remove stop words.
[15]	In House	2363	6	Remove stop words, digits, numbers, and punctuation marks.
[20]	In House	1000	10	Remove stop words, punctuation marks and stemming word and store in feature vector.
[21]	In House	242	6	Remove stop words.
[22]	SNP	5121	7	Remove digits, punctuation marks, non Arabic letters, normalization some Arabic letters and Arabic function words were removed.
[24]	In House	2235	2	tf.idf measure for feature selection

3.2 Text Classification

This step the training matrix is used to train the classification algorithm [23]. After comparing the result of their performance with other algorithm the authors decided to used the best result in build their classifier. Support Vector Machine (SVM) [5], Naïve Bayes (NB) [20], Rocchio [10], Cosine [12] and k-nearest neighbor (KNN) [21] were used to build the final classifier

3.3 Evaluation

Text classification performance evaluated in terms of computational of efficiency and effectiveness of categorization [8]. There are some techniques that are used to measure the evolution of the classifier. One of those techniques: F1, precision and recall that are used in the field of information retrieve and machine learning [12].
The authors had used different types of measures to test their classifiers and this may not justified the result. [5] had shown the results in term of Precision, recall, F1¬ accuracy measure and [20] used Precision, Recall, Fallout and error rate as accuracy measure. Also the K-fold cross validation method used to test the accuracy [21].

The following table summarizes the classifier and the type of accuracy measures that were used to measured the evolution of the algorithm:

Table 2: The evolution

Refer.	Classifier Used	accuracy measure	Classifier	Accuracy
[5]	SVM/NB/ KNN	Precision, recall, F1.	SVM	90%
[10]	KNN/ Rocchio	Precision, recall	Rocchio	98%
[12]	Cosine, Jaccard, and Dice	Micro recall.	Cosine	95%
[13]	SVM/ C5.0	Accuracy	C5.0	78%
[15]	(SMO), NB, J48.	Percentage split method.	SMO	96%
[19]	SVM/NB/ KNN/ Rocchio	Leave One method.	SVM	90%
[20]	NB/KNN/ Distance-Based	Precision, Recall, Fallout and error rate.	NB	-
[21]	KNN/NB	K-fold cross validation.	KNN	84%
[22]	SVM/NB	Recall, Precision, and F1.	SVM	3.8%
[24]	KNN/SVM	Recall, Precision, and F1.	SVM	95%

4. Methods

An analytic, comparative approach was used in order to:

- Brief view in available Arabic text classification literature.
- Explain the text classification phases and show it in the literature.
- Camper between the results in the literature and analyzes the comparison.

5. Results and Analysis

Analyzing the results in the literature was found:

- There are no standard Arabic corpuses that can use easily. Almost all of the authors had justified this with the lack of Arabic corpus in general [5, 10, 20].
- The techniques of removing stop words, digits, numbers, punctuation mark and non Arabic words were use to prepare the text for classification. Some of them extracted the root [10, 19, 20] and the other preferred not to used stemming of the word [5, 12, 13, 15, 21, 22, 25]. Their justified this with the problem of conflated many terms to the same root form [5].
- Almost all of authors used Recall, Precision, and F1. [5, 12, 19, 21].
- There are differences between the classifiers in the accuracy, error rate, and time taken to build the classification [16].
- Support vector machine classifier outperformed K-Nearest Neighbor, Rocchio and NB [2. 19, 22, 8].

6. Conclusion and future work

In despite of the growth of Arabic documents, there is a view corpuses that available free and contain a bid data sit. There is a need to build a benchmarking data set for Arabic text classification. Stemming Arabic word till now is a big problem in the field of Arabic text classification
The future work is to design and develop a model for Arabic text categorization based on rough set theory. This will led to build a standard Arabic corpus which will be available for free.

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., "Text classification," Idea Group Inc., 2005.

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

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

[11]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.

[12]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.

[13]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.

[14]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.

[15]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.

[16]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.

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

[18]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.

[19]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.

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

[21]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.

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

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

[24]Hmeidi I., Hawashin B., El-Qawasmeh, E. "Performance of KNN and SVM classifiers on full word Arabic articles". Advanced Engineering Informatics,vol:106–111, 2008.