A Survey of Arabic Text Representation and Classification Methods

Rami Ayadi[1], Mohsen Maraoui[2], Mounir Zrigui[3]

[1] LaTICE laboratory, University of Sfax, Sfax, Tunisia
ayadi.rami@planey.tn
[2] Computational Mathematics Laboratory, University of Monastir, Monastir, Tunisia
maraoui.mohsen@gmail.com
[3] LaTICE laboratory, Faculty of science of Monastir, Monastir, Tunisia
mounir.zrigui@fsm.rnu.tn

Abstract. In this paper we have presented a brief current state of the Art for Arabic text representation and classification methods. First we describe some algorithms applied to classification on Arabic text. Secondly, we cite all major works when comparing classification algorithms applied on Arabic text, after this, we mention some authors who proposing new classification methods and finally we investigate the impact of preprocessing on Arabic TC.

Keywords: Arabic, impact of preprocessing, Text classification.

1 Introduction

Many researchers have been worked on text classification in English and other European languages such as French, German, Spanish, and in Asian languages such as Chinese and Japanese. However, researches on text classification for Arabic language are fairly limited.

The text classification problem is composed of several sub problems, which have been studied intensively in the literature such as the document indexing, the weighting assignment, document clustering, dimensionality reduction, threshold determination and the type of classifiers. Several methods have been used for text classification such as: Support Vector Machines (SVMs), K Nearest Neighbor (KNN), Neural Networks (NN), Naïve Bayes (NB), Decision Trees (DT), Maximum Entropy (ME), N-Grams, and Association Rules.

Term indexing and weighting aim to represent high quality text. High quality in text mining usually refers to some combination of relevance, novelty, and interestingness.

Several approaches have been used to index and weight terms but all of them share the following characteristics [1]: The more the number of times a term occurs in documents that belong to some category, the more it is relative to that category.

The more the term appears in different documents representing different categories, the less the term is useful for discriminating between documents as belonging to different categories. The most commonly used weighting approach is the Term Frequency Inverse Document Frequency tf-idf.

2 Related work

Researches on the field of Arabic TC fall into two categories: applying and comparing classification algorithms on Arabic text, and investigates the impact of dimensionality reduction.

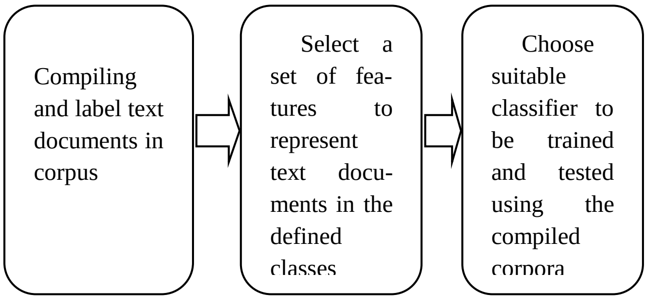
Compiling and label text documents in corpus
Select a set of features to represent text documents in the defined classes
Choose suitable classifier to be trained and tested using the compiled corpora

Fig. 1. Building Text Classification System Process

2.1 Classification and Comparing Algorithms on Arabic Text

El Koudri [2] classified Arabic web documents automatically by Naive Bayes (NB) which is a statistical machine learning algorithm, is used to classify non-vocalized Arabic web documents (after their words have been transformed to the corresponding roots) to one of five pre-defined categories. Cross validation experiments was used to evaluate the NB categorizer. Elkoudri used a corpus of 1500 text documents belonging to 5 categories each category contains 300 text documents. With 2,000 terms/roots, the categorization accuracy varies from one category to another with an average accuracy over all categories of 68.78 %. Furthermore, the best categorization performance by category during cross validation experiments goes up to 92.8%.

Maximum entropy (ME) used by El-Halees [3] and Sawaf [4] for classifying Arabic text documents to classify news articles. The first author preprocessed data using natural language processing techniques such as tokenizing, stemming and part of speech then he used maximum entropy method to classify Arabic documents. The best classification accuracy reported was 80.41% and 62.7% by Sawaf using statistical methods without any morphological analysis.

Al Zoghby [5] introduce a new system developed to discover soft-matching association rules using a similarity measurements based on the derivation feature of the Arabic language. In addition, He presents the features of using Frequent Closed Itemsets (FCI) concept in mining the association rules rather than Frequent Itemsets (FI).

Meshleh [6] implements a Support Vector Machines (SVMs) based text classification system for Arabic language articles using CHI square method as a feature selection method in the pre-processing step of the Text Classification system design procedure. Comparing to other classification methods, the system shows a high classification effectiveness for Arabic data set in term of F-measure (F=88.11). He used an in-house collected corpus from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-Hayat, Al-Ahram, and Al-Dostor as well as a few other specialized websites. The collected corpus contains 1445 documents that vary in length. These documents fall into nine classification categories. Also in another work [7], he investigates the effectiveness and performance of six (CHI, NGL, GSS, IG, OR and MI) commonly used feature selection methods with SVMs evaluated on an Arabic dataset, he concludes conclude that CHI, NGL and GSS performed most effective with SVMs for Arabic TC tasks.

Harrag in [8] presents the results of classifying Arabic text documents using a decision tree algorithm. The results show that the suggested hybrid approach of Document Frequency Thresholding using an embedded information gain criterion of the decision tree algorithm is the preferable feature selection criterion. The study concluded that the effectiveness of the improved classifier is very good and gives generalization accuracy about 0.93 for the scientific corpus and 0.91 for the literary corpus. Experiments are performed over two self-collected data corpus; the first one is from the scientific encyclopedia “Do You Know” (هل تعلم). It contains 373 documents belonging to 1 of 8 categories (innovations, geography, sport, famous men, religious, history, human body, and cosmology), each category has 35 documents. The second corpus is collected from Hadith encyclopedia (موسوعة الحد ثٌ الشر فٌ) from “the nine books” (الكتب التسعة). It contains 435 documents belonging to 14 categories.

The key Nearest Neighbor (kNN) algorithm, which is known to be one of top performing classifiers applied for the English text along with the Support Vector Machines (SVMs) algorithm, has been implemented by Al-Shalabi [9] to the problem of Arabic text categorization. He used Document Frequency threshold (DF) method to keyword extraction and reduction dimensionality. The results show that kNN is applicable to Arabic text; has been reached a 0.95 micro-average precision and recall scores, using a corpus from newspapers (Al-Jazeera, An-Nahar, Al-Hayat, Al-Ahram, and Ad-Dostor) and from Arabic Agriculture Organization website. The corpus consists of 621 documents belonging to 1of 6 categories (politics 111, economic 179, sport 96, health and medicine 114, health and cancer 27, agriculture 100). They preprocessed the corpus by applying stopwords removal and light stemming.

Kheirsat in [10] presented a machine learning approach for classifying Arabic text documents. For the problem of high dimensionality of text documents, embeddings are used to map each document (instance) into R (the set of real numbers) representing the tri-gram frequency statistics profiles for a document. Kheirsat classifies a test text document by computing Manhattan/Dice distance similarity measure to all training documents and assign the class of the training document with smallest/largest computed distance to the test text document. Kheirsat concluded that classification using the Dice measure outperformed classification using the Manhattan measure. Although the Manhattan measure has provided good classification results for English text documents, it does not seem to be suitable for Arabic text documents. Kheirsat collected her corpus from Jordanian Arabic newspapers (Al-Arab, Al-Ghad, Al-Ra’I, Ad-Dostor). The corpus consisted of text documents covering four categories: sports, economy, technology and weather. The technology and weather documents were very small in size ranging from1 to 4 KB. Sports and economy documents were much larger ranging from 2 to 15KB for sports documents and 2 to 18KB for economy documents. The smaller documents constituted about 2% of the total number of documents in the sports and economy category. She applied stop words removal and used 40% for training and 60% for testing.

Harrag in [11] proposes the application of Artificial Neural Network for the classification of Arabic language documents. An Arabic corpus is used to construct and test the ANN model and he discussed the methods of document representation, assigning weights that reflect the importance of each term. Each Arabic document is represented by the term weighting scheme. As the number of unique words in the collection set is big, the Singular Value Decomposition (SVD) has been used to select the most relevant features for the classification. The experimental results show that ANN model using SVD achieves 88.33% which is better than the performance of basic ANN which yields 85.75% on Arabic document classification.

Some studies are compared classification algorithms on Arabic text. Hmeidi reported in [12] a comparative study of two machine learning methods on Arabic text categorization. He evaluated K nearest neighbor (KNN) algorithm, and support vector machines (SVM) algorithm using the full word features and considered the tf.idf as the weighting method for feature selection, and CHI statistics as a ranking metric. Experiments showed that both methods were of superior performance on the test corpus while SVM showed a better micro average F1 and prediction time. The used training and testing data sets are subsets of the most common newspapers in Jordan which are called Alrai and Addustour newspapers. The number of training articles was 2206 articles, and the number of testing articles is 29 articles. The collected documents belong to one of two categories (sport and economic).

In [13] Abbes compared Triggers Classifier (TR-Classifier) and KNN to identify Arabic topic. Performances are acceptable, particularly for TR-classifier, using reduced sizes of vocabularies. For the TR-Classifier, each topic is represented by a vocabulary which has been built using the corresponding training corpus. Whereas, the kNN method uses a general vocabulary, obtained by the concatenation of those used by the TR-Classifier.), the average recall and precision for KNN and TR are 0.75, 0.70 and 0.89, 0.86 respectively. Abbas collected 9,000 articles from Omani newspaper (Al-Watan) of year 2004. The corpus belongs to 1 of 6 categories (culture, economic, religious, local news, international news). The corpus includes 10M word including stopwords. After removing stopwords and infrequent words the vocabulary size became 7M words. Tf.idf was used as weighting schemes.

Duwairi in [14] compared the performance of three classifiers for Arabic text classification as KNN, Naïve Bayes, and Distance-Based classifier. Each documents were preprocessed by removing punctuation marks and stop words, then all of them are represented as a vector of words (for the case of Naïve bayes, he used a vector of words and their frequencies). As a method to reduce the dimensionality of feature vector, the author use Al-shalabi stemmer. Experimental results show that NB outperforms the other two algorithms. Duwairi collected 1,000 documents fall into 10 predefined categories; each category contains 100 documents. The set of categories include: sport, economic, internet, art, animals, technology, religious, politics, medicine and plants.

In another work, three classification algorithms on Arabic text are compared by Kannan in [15]; the three algorithms were KNN, NB, and Rocchio. The research results reveal that Naïve Bayes was the best performer (F1=0.8209), followed by kNN (F1=0.7871) and Rocchio (F1=07882). The used corpus is collected from online newspapers (Al-Jazeera, An-Nahar, Al-Hayat, Al-Ahram, and Ad-Dostor). The corpus consists of 1445 documents belonging to 9 categories (medicine, sport, religious, economic, politics, engineering, low, computer, and education). They applied light stemming approach for feature reduction and 4-folds cross-validation was performed for evaluation.

The performance of two popular text classification algorithms (SVMs and C5.0) is evaluated in [16] by Al-Harbi to classify Arabic text using seven Arabic corpora. The average accuracy achieved by SVMs is 68.65%, while the average accuracy achieved by C5.0 is 78.42%. One of the goals of their paper is to compile Arabic corpora to be benchmark corpora. The authors compiled 7 corpora consisting of 17,658 documents and 11,500,000 words including stopwords but the corpora are not available publically.

Bawaneh applied KNN and NB on Arabic text and conclude that KNN has better performance than NB [17], they also conclude that feature selection and the size of training set and the value of K affect the performance of classification. The Researchers also posed the problem of unavailability of freely accessible Arabic corpus. The in-house collected corpus consists of 242 documents belonging to 6 categories. Authors applied light stemming as a feature reduction technique and tf-idf as weighting scheme, they also performed cross-validation test.

Alsaleem in [28] investigate Naïve Bayesian method (NB) and Support Vector Machine algorithm (SVM) on different Arabic data sets for TC. The data used are The Saudi Newspapers (SNP) [1], the data set consist of 5121 Arabic documents of different lengths that belongs to 7 categories, the categories are (Culture " الثقافية ", Economics " الإقتصادية " , General " العامة " , Information Technology " تكنولوجيا المعلومات " , Politics "السياسية", Social " الأجتماعية ", Sport " الرياضة "). The average of three measures obtained against SNP Arabic data sets indicated that the SVM algorithm (F1=0.778) outperformed NB algorithm regards to F1=0.74, Recall and Precision measures.

El-Halees in [18] compared six well known classifiers, which are: Maximum entropy, Naïve Bayes, Decision Tree, Artificial Neural Networks, Support Vector Machine, and k-Nearest Neighbor using the same data sets and the same experimental settings. The recall, precision and f-measure for the classifiers are computed and compared. The author compared the methods after preprocessing and all stop words are removed and he found that the performance of Naïve Bayes is the best (F1= 91.81), the performance of Maximum Entropy, Support Vector Machine and Decision Tree are acceptable, but the performance of k-Nearest Neighbor and Artificial Neural Networks was bad.

However, after using Information Gain as feature selection, the data was reduced significantly and the performance of k-Nearest Neighbor and Artificial Neural Networks improved significantly. The performance of Naïve Bayes did not change but still the best classifier to Arabic corpus.

In these experiments, the author used an Arabic documents collected from Aljazeera Arabic news channel. The documents categorized into six domains: politics, sports, culture and arts, science and technology, economy and health. The author applied stop words removal and normalization and used 10-folds cross-validation for testing.

Ismail in [29] implemented the Sequential Minimal Optimization (SMO), Naïve Bayesian (NB) and J48 (C4.5) Algorithms using weka[1] program and compared between the algorithms in accuracy and time to get the result. The data set used consists of 2356 documents of different lengths. Each document was manually labeled based on its contents and the domain that it was found within, these documents categorized to six categories (Sport, Economic, Medicine, Politic, Religion and Science) where 60% of the data used as training and the remaining 40% used as testing. Token reduction approach for documents is used to minimize storage requirements and all the types of stop words are removed. The results show that the Sequential Minimal Optimization (SMO) classifier achieves the highest accuracy (96.08%) and the lowest error rate (3.42%), followed by the J48 (C4.5) classifier (90.48% and 9.52%), then by the Naive Bayes (NB) classifier (85.60% and 14.4%). The second part of the results shows that the time needed to build the SMO model is the faster one (5.2 seconds), followed by NB model (12.0 seconds), then J48 classifier which takes a highest amount of time (291.72 seconds).

Table 1. F_mesure

	F-measure without reduction	F-measure with Information Gain
Maximum entropy	85.96	83.83
Naïve Bayes	91.81	83.9
Decision Tree	71.91	74.48
Artificial Neural Networks	10.81	74.33
Support Vector Machine	88.33	88.33
k-Nearest Neighbor	38.6	70.07

Duwairi in [19]- [20] propose a distance-based classifier for categorizing Arabic text. Each category is represented as a vector of words in an m-dimensional space, and documents are classified on the basis of their closeness to feature vectors of categories. The classifier, in its learning phase, scans the set of training documents to extract features of categories that capture inherent category specific properties; in its testing phase the classifier uses previously determined category-specific features to categorize unclassified documents. Stemming was used to reduce the dimensionality of feature vectors of documents. The accuracy of the classifier was tested by carrying out several categorization tasks on an in-house collected Arabic corpus. The average accuracy reported was 0.62 for the recall and 0.74 for the precision. He collected 1000

[1] http://www.cs.waikato.ac.nz/ml/weka/

text documents belonging to 10 categories (sport, economic, internet, art, animals, technology, plants, religious, politics, and medicine). Each category contains 100 documents. She used 50% for training and 50% for testing.

Alruily describes an initial prototype for identifying types of crime in a Arabic text within the crime domain. Two approaches are explored to perform recognition tasks. The first approach completely relies on direct recognition using gazetteers. The second approach is a rule-based system. Rules are built based on the predefined crime indicator list that contains some important keywords [21].

Abbas in [22] proposed Triggered (TR) classifier. Triggers of a word Wk are ensemble of words which highly correlated with it. The main idea of TR-Classifier is computing the average mutual information (AMI) for each couple of words from the training documents and testing document, and then assigns the topic that highest AMT to the test document. The best recall achieved is 0.9.

In [23], Ayadi applied inter-textual distance theory to classify any anonymous Arabic text according to criteria of lexical statistic; this requires integration of a metric for classification task using a database of lemmatized corpus.

Syiam presented an Arabic text categorization system based on Machine learning algorithms and many algorithms for stemming and feature selection [24]. The document is represented using several term weighting schemes and finally the k-nearest neighbor and Rocchio classifiers are used for classification process. Experiments show that the hybrid method of statistical and light stemmers is the most suitable stemming algorithm for Arabic language and the hybrid approach of document frequency and information gain is the preferable feature selection criterion and normalized-tfidf is the best weighting scheme. Finally, Rocchio classifier has the advantage over k-nearest neighbor classifier in the classification process and gives generalization accuracy of about 98%.

2.2 The Impact of Dimensionality Reduction in TC

Duwairi in [25] analyzed and compared three feature reduction techniques; stemming, light stemming, and word clusters using K-nearest-neighbor classifier applied to Arabic text. The purpose of employing the previous methods is to reduce the size of document vectors without affecting the accuracy of the classifiers. Comparison metrics are size of document vectors, classification time, and accuracy (in terms of precision and recall). The corpus consists of 15,000 documents belonging 3 categories: sports, economics, and politics. In terms of vector sizes and classification time, the stemmed vectors consumed the smallest size and the least time necessary to classify a testing dataset that consists of 6,000 documents. The light stemmed vectors superseded the other three representations in terms of classification accuracy.

Thabtah tested and compared three variations of vector space models (VSMs) (these variations are Cosine coefficient, Dice coefficient and Jacaard coefficient) and term weighting approaches (IDF, WIDF, ITF and log (1+tf)) using KNN algorithm [26]. The Experimental results on different Arabic text categorization data sets collected from online Arabic newspapers including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, and Al-Dostor. With regards to F1 results, the author concluded that Dice based TF.IDF (94.91) and Jaccard based TF.IDF (94.91) outperformed Cosine based TF.IDF (90.93), Cosine based WIDF (75.94), Cosine based ITF (91.02), Cosine based log(1+tf) (92.65), Dice based WIDF (81.01), Dice based ITF (89.63), Dice based log(1+tf) (85.21), Jaccard based WIDF (81.01), Jaccard based ITF (89.63), and Jaccard based log(1+tf) (85.21).

In [33][32], the author made an experimental study for compare two approaches of reduction dimensionality and verifies their effectiveness in Arabic document classification. Firstly, he apply latent Dirichlet allocation (LDA) and latent semantic indexing (LSI) for modeling the corpus contained 20.000 documents. He generates two matrices LDA (documents/topics) and LSI (documents/topics). Then the SVM algorithm is used for document classification, which is known as an efficient method for text mining. Classification results are evaluated by precision, recall and F-measure. The experiment shows that the results of dimensionality reduction via LDA outperform LSI in Arabic topic classification.

Said in [27] presented an evaluation study of the benefits of using morphological tools in Text Classification. The study includes using the raw text, the stemmed text, and the root text. The stemmed and root text are obtained using two different preprocessing tools. The results show that using light stemmer combined with a good performing feature selection method such as mutual information enhances the performance of Arabic Text Categorization especially for small sized data sets and small threshold values. Additionally, using the raw text leads to the worst performance in small datasets while its performance was among the best tools in large datasets. This may explain the contradiction in the results obtained previously in the literature of the Arabic text categorization since the performance of the preprocessing tools is affected by the characterizes of the dataset used.

3 Discussion

From previous discussion, most of related work in the literature used small in-house collected corpus, and applied one or two classifiers to classify one corpus which is not enough to evaluate Arabic TC. Thus, there are contradictions between results of researches in the literature because of using different corpora and different preprocessing techniques.

In addition, the impact of text preprocessing and different term weighting schemes combinations on Arabic text classification using popular text classification algorithms has not been studied in the literature. Also, there is a debate among researchers about the benefits of using morphological tools in TC.

We summarized the research problems in the following points:

- Debate among researchers about the benefits of using English morphological tools in TC. To the best of our knowledge, the benefits of using Arabic morphological tools (stemming and light stemming) is not address for Arabic Language; only [14][19] applied on single corpus belong to only 3 categories.
- The impact of text preprocessing and different term weighting schemes combinations on Arabic text classification using popular text classification algorithms has

not been studied in the literature. Only [25][27] have addressed the impact of morphological analysis tools on Arabic text classification. Their work is not comprehensive regarding Arabic corpora, classifiers, and term weighting schemes.

- The lack of availability of publically free accessible Arabic Corpora.
- The lack of standard Arabic morphological analysis tools.
- Most of related works in the literature used small in-house collected corpus.
- Most of related works in the literature applied one or two classifiers to classify one corpus. This is not enough to evaluate Arabic TC.
- There are contradictions between results of researches in the literature because of using different corpora and different preprocessing techniques.

For this, we have focused as objectives to build the largest publically free accessible Arabic Corpora, implement and integrate Arabic morphological analysis tools, conduct a comprehensive study about the impact of text preprocessing on Arabic text classification, and develop a method of representation and indexing text reflecting more semantics.

We start by building our data set. The corpus contains 20.000 documents that vary in writing styles. These documents fall into 10 categories that equal in the number of documents. In this Arabic dataset, each document was saved in a separate file within the directory for the corresponding category, i.e., the documents in this dataset are single-labeled. Tables 2 and 3 show more specified details about the collection.

Table 2. Dataset

NB of text in the corpus	20.000
NB of words in the corpus	2 .523 .022
Size of corpus (Mb)	34.0 Mb
NB of category	10

Table 3. Number of documents in each category

OATC	**NB of text**	**Average number of words per text**	**Number of words per category**	**Category Size (Mo)**
Sport	2 000	141.261	282 522	2.99
regional	2 000	125.723	251 447	2.71
Culture	2 000	168.485	336 971	3.62
world	2 000	105.701	211 402	2.26
National	2 000	136.739	273 479	2.97
political	2 000	164.356	328 712	3.53
Economic	2 000	148.922	297 845	3.27
Student	2 000	203.485	406 971	4.50
Investigation	2 000	253.602	507 205	5.43
Judicial	2 000	126.93	253 860	2.70

The corpus are collected from online Arabic newspapers including http://www.attounissia.com.tn, www.alchourouk.com/, www.assabahnews.tn/, http://jomhouria.com/, Table 4 summarize the percentage of split between different sources. As we can show for example, the "sport" category is composed of 25% from Attounissia, 25% from Alchourouk, 25% from Assabahnews, 25% from Jomhouria.

Table 4. Percentage of split between different sources

Source	Attounissia	Alchourouk	Assabahnews	Jomhouria
Sport	25%	25%	25%	25%
Regional	-	50%	50%	-
Culture	25%	25%	25%	25%
word	25%	25%	25%	25%
National	25%	25%	25%	25%
Political	-	100%	-	-
Economic	50%	-	-	50%
Student	100%	-	-	-
Investigation	100%	-	-	-
Judicial Incidents	25%	25%	25%	25%

4 Conclusion

In this paper we have presented a brief current state of the Art for Arabic text representation and classification methods. All major problems in the most of related work in the literature used small in-house collected corpus, and applied one or two classifiers to classify one corpus which is not enough to evaluate Arabic TC. Thus, there are contradictions between results of researches in the literature because of using different corpora and different preprocessing techniques. In addition, the impact of text preprocessing and different term weighting schemes combinations on Arabic text classification using popular text classification algorithms has not been studied in the literature. Also, there is a debate among researchers about the benefits of using morphological tools in TC. For this, we have built the largest publically free accessible Arabic Corpora and in future work we will implement and integrate Arabic morphological analysis tools, conduct a comprehensive study about the impact of text preprocessing on Arabic text classification, and develop a method of representation and indexing text reflecting more semantics.

References

1. Saad, M. K. (2010). The impact of text preprocessing and term weighting on arabic text classification (Doctoral dissertation, The Islamic University-Gaza).
2. El-Kourdi M., Bensaid A. and Rachidi T. (2004). Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics. August, Geneva. Pp51—58

3. El-Halees A. (2007), “Arabic Text Classification Using Maximum Entropy”, The Islamic University Journal (Series of Natural Studies and Engineering), 15(1), pp. 157-167.
4. Sawaf H., Zaplo J., Ney H. (2001), “Statistical Classification Methods for Arabic News Articles”, In the Workshop on Arabic Natural Language Processing, ACL'01, Toulouse, France.
5. Al-Zoghby A., Eldin AS., Ismail NA., Hamza T. (2007), “Mining Arabic Text Using Soft Matching association rules”, In the Int. Conf. on Computer Engineering & Systems, ICCES'07, pp 421 - 426.
6. Mesleh A. (2007), “Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System”, Journal of Computer Science, 3(6), pp. 430-435.
7. Mesleh A. (2007), “Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study”, In the 12th WSEAS Int. Conf. on APPLIED MATHEMATICS, Cairo, Egypt.
8. Harrag F., El-Qawasmeh E., Pichappan P. (2009), “Improving Arabic text categorization using decision trees”, In the 1st Int. Conf. of NDT '09, pp. 110 – 115.
9. Al-Shalabi R., Kannan G., Gharaibeh H. (2006), “Arabic text categorization using KNN algorithm”, In the Proc. of Int. multi conf. on computer science and information technology CSIT06.
10. Khreisat L. (2009), “A machine learning approach for Arabic text classification using N-gram frequency statistics”, Journal of Informetrics, Elsevier, 3(1), pp. 72-77.
11. Harrag F., El-Qawasmeh E. (2009), “Neural Network for Arabic text classification”, In the 2nd Int. Conf. of Applications of Digital Information and Web Technologies, ICADIWT '09, pp. 778 – 783.
12. Hmeidi I., Hawashin B., El-Qawasmeh E. (2008), “Performance of KNN and SVM classifiers on full word Arabic articles”, Journal of Advanced Engineering Informatics 22, pp. 106–111.
13. Abbas M., Smaili K., Berkani D. (2009), “Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies”, In The 3rd Int. Conf. on Arabic Language Processing, CITALA2009, Mohammadia School of Engineers, Rabat, Morocco.
14. Duwairi R. (2007), “Arabic text Categorization”, In the Int. Arab journal of information technology, 4(2).
15. Kanaan G., Al-Shalabi R., Ghwanmeh S. (2009), “A comparison of text-classification techniques applied to Arabic text”, Journal of the American Society for Information Science and Technology, 60(9), pp. 1836 – 1844,.
16. Al-Harbi S., Almuhareb A., Al-Thubaity A., Khorsheed M., Al-Rajeh A. (2008), “Automatic Arabic Text Classification”, In JADT’08, France, pp. 77-83.
17. Bawaneh M., Alkoffash M., Al-Rabea A. (2008), “Arabic Text Classification using K-NN and Naïve Bayes”, In Journal of Computer Science, 4 (7), pp. 600-605.
18. El-Halees A (2008)., “A Comparative Study on Arabic Text Classification”, Egyptian Computer Science Journal 20(2).
19. Duwairi R (2005)., “A Distance-based Classifier for Arabic Text Categorization”, In the Proc. of the Int. Conf. on Data Mining, Las Vegas, USA.
20. Duwairi R. (2006), “Machine Learning for Arabic text Categorization”, Journal of the American Society for Information Science and Technology, 57(8), pp. 1005-1010.
21. Alruily M., Ayesh A, Zedan H. (2009), “Crime Type Document Classification from Arabic Corpus”, In the 2nd Int. Conf. on Developments in eSystems Engineering, pp.153-159.
22. Abbas M., Smaili K., Berkani D. (2009), “A Trigger-based Classifier”, In The 2nd Int. Conf. on Arabic Language Resources and Tools (MEDAR 2009), 22-23, Cairo, Egypt.

23. Ayadi, R., Maraoui, M. and Zrigui, M. (2011) ‘SCAT: a system of classification for Arabic texts’, Int. J. Internet Technology and Secured Transactions, Vol. 3, No. 1, pp.63–80.
24. Syiam M., Fayed Z., Habib M. (2006), “An Intelligent System for Arabic Text Categorization”, In IJICIS, 6(1), pp. 1-19.
25. Duwairi R., Al-Refai M., Khasawneh N. (2009), “Feature reduction techniques for Arabic text categorization”, Journal of the American Society for Information Science, 60(11), pp. 2347-2352.
26. Thabtah F., Hadi W. Musa, Al-shammare G. (2008), “VSMs with K-Nearest Neighbour to Categorize Arabic Text Data”, in the Proc. of the World Congress on Engineering and Computer Science, WCECS’2008, San Francisco, USA.
27. Said D., Wanas N., Darwish N., Hegazy N. (2009), “A Study of Arabic Text preprocessing methods for Text Categorization”, In the 2nd Int. conf. on Arabic Language Resources and Tools, Cairo, Egypt.
28. Alsaleem S. (2011), “Automated Arabic Text Categorization Using SVM and NB”,in the International Arab Journal of e-Technology, Vol. 2, No. 2, pp. 124-128.
29. Majed Ismail Hussien, Fekry Olayah, Minwer AL-dwan, Ahlam Shamsan (2011), “ARABIC TEXT CLASSIFICATION USING SMO, NAÏVE BAYESIAN, J48 ALGORITHMS”, in the IJRRAS 9 (2), pp. 306-316.
30. Maraoui, M., Antoniadis, G., & Zrigui, M. (2009, July). SALA: Call System for Arabic Based on NLP Tools. In IC-AI (pp. 168-172).
31. Zouaghi, A., Zrigui, M., & Antoniadis, G. (2008). Automatic Understanding of Spontaneous Arabic Speech-A Numerical Model. TAL, 49(1), 141-166.
32. Ayadi, R., Maraoui, M., & Zrigui, M. (2014). Latent Topic Model for Indexing Arabic Documents. *International Journal of Information Retrieval Research (IJIRR)*, *4*(1), 29-45.
33. Ayadi, R., Maraoui, M., & Zrigui, M. (2015, October). LDA and LSI as a Dimensionality Reduction Method in Arabic Document Classification. In *International Conference on Information and Software Technologies* (pp. 491-502). Springer International Publishing.