I.J. Intelligent Systems and Applications, 2015, 06, 71-78
Published Online May 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2015.06.08

Arabic Text Categorization Using Logistic Regression

Mayy M. Al-Tahrawi
Computer Science Department, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan
Email: mayy.tahrawi@gmail.com

Abstract— Several Text Categorization (TC) techniques and algorithms have been investigated in the limited research literature of Arabic TC. In this research, Logistic Regression (LR) is investigated in Arabic TC. To the best of our knowledge, LR was never used for Arabic TC before. Experiments are conducted on Aljazeera Arabic News (Alj-News) dataset. Arabic text-preprocessing takes place on this dataset to handle the special nature of Arabic text. Experimental results of this research prove that the LR classifier is a competitive Arabic TC algorithm to the state of the art ones in this field; it has recorded a precision of 96.5% on one category and above 90% for 3 categories out of the five categories of Alj-News dataset. Regarding the overall performance, LR has recorded a macroaverage precision of 87%, recall of 86.33% and F-measure of 86.5%.

Index Terms— Logistic Regression, Arabic Text Categorization, Arabic Document Classification

I. INTRODUCTION

Text Categorization (TC) is the process of automatic classification of unseen texts into a specific category, like Religion, Computer, Economics, Law, Sport, … etc. This classification can be either supervised or unsupervised. Supervised classification, which is more accurate and is adopted in this research, depends on training the classification system using a set of labeled text documents.

The amount of online text documents, which needs automatic retrieval, and thus classification, increases rapidly with the rapid growth of the internet usage all over the world in different areas which depend on document categorization, like information retrieval, online news, digital libraries, topical crawling, spam filtering and automatic machine response to e-mails.

Interest in Arabic TC has grown recently with the emergent desperate need for Arabic automatic TC systems in the last few years, which can be attributed to many reasons: Firstly, the Arabic content on the internet exceeds 3% of the whole internet content and comes in the eighth rank [1]; this huge content needs to be searched, exchanged and retrieved as fast and accurate as possible. Secondly, the Arabic language is one of the seven official languages of the United Nations with more than 400 million Arabic native speakers. Thirdly, most of these Arabic native speakers cannot read English.

Many well-known English TC techniques or algorithms were investigated in Arabic TC and have proved to be efficient Arabic text classifiers; examples include the Na ïve Bayes algorithm (NB) [2-4], Support Vector Machines (SVM) [5-9], k-Nearest Neighbor (kNN) [4, 10], Decision Tree [7, 9, 11] besides others [4, 6, 12-16].

Logistic Regression (LR) is a well-known statistical algorithm which was used widely in information retrieval [17-22]. LR was also investigated algorithm in English TC by some researchers [23-36].

Regarding Arabic TC, LR was never investigated before in Arabic TC. In this research, we investigate using LR for Arabic TC. Experiments are conducted on Aljazeera Arabic News (Alj-News) [37] dataset, which consists of 1500 Arabic News articles distributed evenly among five categories: Art, Economic, Politics, Science and Sport. Some text-preprocessing steps are applied on the dataset to handle the special nature of Arabic text, and Chi Square (CHI) is used for FS. Detailed Results of these experiments are presented in Section five.

The rest of the paper is organized as follows: The problem of Arabic TC is presented in Section two, Logistic Regression algorithm is explained in brief in Section three, the Dataset is presented in Section four, Experiments, Results and Analysis of these results are presented in Section five and finally Conclusions take place in Section six.

II. ARABIC TEXT CATEGORIZATION (TC)

Building a TC system usually starts with a preprocessing stage to prepare texts for automatic categorization, then follows the classifier training stage and finally testing the classifier and evaluating its performance using some formal evaluation criteria.

Arabic texts share some pre-processing steps with texts written in other languages, like stop words removal, stemming, feature weighting and selection. However, due to the special nature of the Arabic Language, additional special types of preprocessing are needed. Details of the data pre-processing stage applied in our research is presented in the next subsection.

A. Data Pre-processing

The Arabic language differs from the Latin-based alphabets in many aspects. Firstly, it is written from right to left. Secondly, one letter can have different shapes depending on its position in the word; for example, (هـ, ـهـ , ه ، ـه) are four different shapes for one letter: ‘هـ’ at the beginning, ‘ـهـ ‘ in the middle of, and ‘ه’ or ‘ـه’ at the end of a word. Thirdly, the Arabic language exhibits two genders: masculine and feminine and three number classes: singular, dual, and plural. Moreover, the Arabic plurals are divided into two classes: regular and broken. Finally, Arabic nouns have three cases: nominative, accusative and genitive. As a result, Arabic language is very complex and rich, which justifies the difficulties in achieving precise automatic TC results when dealing with Arabic text documents.

The Arabic language consists of 28 letters (أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي) in addition to the Hamza (ء). All Arabic letters are considered consonants, except the letters (أ و ي) which are considered long vowels. Diacritics are used in the Arabic language, as short vowels, to show the correct pronunciation and thus the meanings of words. These are: Fatha, Damma, Kasra, Sukūn, Mad (ا), Shadda and Tanwin (◌ٍ,◌ٌ,◌ً). The only way to disambiguate diacritic-less Arabic words is to locate them within the context. Shapes and sounds of Arabic diacritics are listed in Table 1.

Table 1. Shapes and sounds of Arabic diacritics

Diacritic	**Example**	**Sound**
Fatha	بَ	Ba
Damma	بُ	Bu
Kasra	بِ	Bi
Madd	آ	Aa
Shadda	بّ	Bb
Sukun	بْ	B
Tanwin	بٍ باً بٌ	bun, ban, bin

Data pre-processing is the first stage in building TC systems for all languages. It aims mainly to reduce the number of features used in building automatic classifiers, thus reducing requirements of memory and processing resources. Data-preprocessing aims also to enhance classification accuracy by removing noisy features from the dataset. Typical Arabic TC pre-processing steps include, but not limited to, the following:

1) Tokenization: which converts a text document from a stream of characters into a sequence of tokens (features or terms) by recognizing delimiters such as white spaces, punctuations, special characters, … etc.
2) Removal of the non-Arabic letters.
3) Removal of numbers, diacritics, special characters and punctuations.
4) Removal of stop words: these include pronouns, conjunctions, and prepositions.
5) Stemming: reducing an inflected or derived word to its stem. The stem needs not to be a valid morphological root of the word as far as related words map to the same stem. The main advantage of this pre-processing step is to reduce the number of terms in the corpus so as to reduce the computational and storage requirements of TC algorithms. With the case of the highly derivative Arabic language, in which a large number of words can be formed using one stem, stemming is a valuable tool in reducing complexity of automatic TC.
6) Some optional pre-processing steps include removal of words with one-character length after stemming and words which occur infrequently in the corpus.

III. Logistic regression (LR) Algorithm

Logistic Regression (LR) is a well-known statistical algorithm which has the advantage of yielding a probability model that can be useful in many applications. LR was used widely in information retrieval. It has been studied in the field of machine learning [38-40], including English TC. Some studies have shown that the LR model is able to achieve similar English TC performance as SVMs [39, 41]. To the best of our knowledge, LR was never investigated before in Arabic TC.

Logistic Regression (LR) is a discriminative model that can be used for probabilistic categorization. It outputs the posterior probabilities for test examples that can be conveniently engaged in other systems. If our ultimate goal is classification, then given a test example x, LR can directly estimate the conditional probability of assigning a class label y to the example by [40]:

$$P(y|x)= \frac{1}{1 + \exp(-y\alpha^{T}\ x)} \qquad (1)$$

where α is the model parameter.

Logistic regression can be easily generalized to multiple classes by treating multi-class classification as several binary classification problems. The decision of whether to assign the class can be based on comparing the probability estimate with a threshold or, more generally, by computing which decision gives optimal expected effectiveness [42, 43].

For a logistic regression model to make accurate predictions for future inputs, we must avoid overfitting the training data. In this research, the Iteratively Re-weighted Least Squares (IRLS) nonlinear optimization algorithm is used as a fitting procedure [41, 44]. This technique uses the Newton-Raphson algorithm to solve the LR score equations [44].

IV. The DataSet

Different Arabic Datasets were used in the research field of Arabic TC, as no benchmark Arabic dataset exists. Aljazeera News Arabic Dataset (Alj-News), available at [37], is used in this research. Alj-News dataset is gathered from Al-Jazeera Arabic News Website. The dataset consists of 1500 Arabic news documents distributed evenly among five classes: Art, Economic, Politics, Science and Sport. Each class has 300 documents (240 for training and 60 for testing). This dataset was used in several researches in the literature of Arabic TC [9, 16, 45, 46].

The data pre-processing, applied on Alj-News dataset in this research, is explained in detail in Section II.A. In this research, the stemming algorithm of Khoja [47] is adopted. It is a well-known Arabic Stemmer which removes the longest suffix and the longest prefix. It then matches the remaining word with verbal and noun patterns to extract the root. The stemmer makes use of several linguistic data files such as a list of all diacritic characters, punctuation characters, definite articles, and stop words. The stop word list adopted by [47] is extended in this research to include 478 stop words rather than the list of just 168 stop words adopted by them.

Khoja stemmer has been developed in both C++ and Java and is available at [48]. The authors in [49] evaluated Arabic language morphological analyzers and stemmers and reported that Khoja stemmer achieved the highest accuracy in their experiments. The stemmer has also been used as part of an information retrieval system developed at the University of Massachusetts for the TREC-10 cross-lingual track in 2001. The authors in [50] reported that although the stemmer produced many mistakes, it improved the performance of their system immensely.

The steps of Khoja stemming algorithm [47] can be summarized as follows:

1) Format the word by removing any punctuation, diacritics and non-letter characters.
2) Ignore stop words.
3) Remove the definite article ال وال بال كال فال.
4) Remove the special prefix (و).
5) Remove and duplicate the last letter, if the last letter is a shadda.
6) Replace أ إ آ with ا
7) Remove Prefixes. للـ لـ سـ فـ
8) Remove Suffixes.كن هما كما
9) Match the result against a list of Patterns. فاعل افعل تفعيل فعال
10) Replace all occurrences of Hamza ؤ ئ ء with ا
11) Two letter roots are checked to see if they should contain a double character; if so, the character is added to the root.

Alj-News dataset ended up with the number of terms shown in Table 2 after applying all the text pre-processing steps.

Table 2. The final number of terms in Alj-News Dataset

CLASS	Number of Terms
Art	3745
Economic	2178
Politics	2984
Science	2806
Sport	3332
TOTAL	**15045**
FILTERED (after removing duplicates among classes)	**8218**

V. Experiments and Results

Details of Feature Selection and Reduction, Performance Evaluation Measures and Results of experiments conducted in this research are presented in subsections A through D.

A. *Feature Selection (FS)*

Feature Selection (FS) is widely used in TC, as most classifiers cannot afford to work with the huge number of features (terms) in the corpus. Add to this, the effect of using all terms in building a classifier on the classifier accuracy was always a great debate; many researchers believe that using all corpus terms adds both noise and processing requirements to the classifiers, while some researchers found FS to be harmful to categorization [51-54].

Using FS, the discriminating power of each term is computed, and only the top-scoring ones are used to build the classifier. Several FS methods are used in the literature of Arabic TC research, like Cross Validation [3], Chi Square (CHI) [5, 6, 16, 55-58], Information Gain(IG) [7, 45, 55], Document Frequency (DF) [45, 55], Mutual Information (MI) [45], Correlation Coefficient (CC) [45], Binary Particle Swarm Optimization- K-Nearest-Neighbor (BPSO-KNN) [9], Semi-Automatic Categorization Method (SACM) and Automatic Categorization Method (ACM) [59]. On the other hand, [60] selected features randomly and [15] didn't apply FS at all.

Chi Square (CHI) is used in the experiments of this research as a FS metric for selecting the most discriminating features in the dataset. CHI has proved to record high accuracy in classifying both English [7, 6, 16, 61-66] and Arabic [5, 6, 16, 55-58] texts. The CHI FS metric measures the lack of independence between a term and a class. It was originally used in the statistical analysis of independent events. Its application as a FS metric for TC purposes goes through the following steps:

1) For each term in each class in the training set, compute the CHI score to measure the correlation between the term and its containing class. CHI is computed for each term *t* in each class c_i as follows [67]:

$$\chi^2(t, c_i) = \frac{N \times (AD\text{-}CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \qquad (2)$$

where: *N* is the total number of training documents in the dataset, *A* is the number of documents belonging to class c_i and containing *t*, *B* is the number of documents belonging to class c_i but not containing *t*, *C* is the number of documents not belonging to class c_i but containing *t* and *D* is the number of documents neither belonging to class c_i nor containing *t*.

2) Combine the class-term CHI measures for terms that appear in more than one class in one score using the maximum or average score.

After deciding on the terms to be selected for building the classifier, the terms will be represented in the categorization system using one of the various presentations or weights used in the literature of TC.

Common examples include Term Frequency. Inverse Document Frequency (TF.IDF) [3, 5, 9, 14, 56, 59], Term Frequency (TF) [14, 15, 55, 57, 58], Document Frequency (DF) [55], Weighted IDF [14], Normalized Frequency [7, 16, 60-64], Boolean [6, 55, 61, 62, 64] and other FS methods like Cosine coefficient, Dice coefficient and Jacaard coefficient [68]. In this research, Normalized frequency is used to as a weighting scheme for term representation in the Vector Space Model.

B. Feature Reduction

A class-based local policy is applied, in this research, for selecting the best terms for building all the classifiers by selecting 1% of the topmost terms from each of the five classes. This policy has proved to achieve the best categorization performance compared to other reduction policies, like choosing the topmost corpus terms, or an equal number of terms from each class, as it gives each class a representative share in the final set of terms used to build the classifier [16, 61-64, 69]. The number of terms selected from each class and the total number of terms, after applying CHI and Feature Reduction, then removing duplicates is summarized in Table 3.

Table 3. The number of terms used to build the classifier

CLASS	**1% of Terms**
Art	37
Economic	22
Politics	30
Science	28
Sport	33
TOTAL	**150**
FILTERED	**135**

C. Performance Evaluation Measures

All classifiers are evaluated by computing their precision, recall and F1- measure. These three measures are known to be reliable evaluation measures of the classifier effectiveness and have been used widely in evaluating Arabic TC systems [4, 7- 9, 14-16, 60 , 68, 70, 71]. Precision is defined as the proportion of test files classified into a class that really belong to that class, while Recall is the proportion of test files belonging to a class and are claimed by the classifier as belonging to that class. Precision of a class ci, denoted by (Pi), is computed as [72]:

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

and Recall of a class c_i, denoted by *(*R_i*)*,is computed as [72]:

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

where $TP_{i,}$ FP_i and FN_i refer to Truly Positive, Falsely Positive and Falsely Negative claims of the classifier respectively.

The F1 measure, introduced by [73], is the harmonic average of both precision and recall. High F1 means high overall performance of the system. F1 is computed as follows [72]:

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \tag{5}$$

$$= \frac{2TP}{2TP + FP + FN} \tag{6}$$

After computing individual performance results for each class, results of all classes are microaveraged and macroaveraged to give an idea of the categorization performance on the dataset as a whole.

D. Results

Results of classifying Alj-News dataset using Logistic Regression are summarized in Table 4. The best recorded performance was on “Sports” class with precision of 96.49, recall of 91.67 and F1-measure of 94.0171. The next best performance was on the “Science” class with precision of 93.1034, recall of 90.00 and F1-measure of 91.525. “Art” was classified with a precision of 90.74 and an F1-masure of 85.96. Economic comes the fourth with an F1-measure of 81.967 and finally comes the “Politics” class with an F1-measure of 79.07. Nevertheless, “Politics” was the third performer using Recall as it recorded 85%. Comparisons of the performance of Logistic Regression on Alj-News dataset with other algorithms experimented on classifying the same dataset takes place in the following paragraph.

Ref. [9] proposed BPSO (Binary Particle Swarm Optimization)-KNN as a FS method for Arabic TC. They experimented three different classifiers on exactly the same dataset (Alj-News), with the same training and testing split, used in our research. These algorithms are: Support Vector Machines (SVM), Na ïve Bayes (NB) and Decision Trees (J48). They ended up with 5329 features after applying a set of pre-processing steps on the corpus. These preprocessing steps include removing hyphens, punctuation marks, numbers, digits, non-Arabic letters and diacritics. Then stop words and rare words (words that occur less than five times in the dataset) were removed. From these terms, they selected 2967 features to build the three classifiers. Results reached in their experiments on Alj-News are summarized in Tables 5 to 7 and comparisons of our results with the results of this research are summarized in Figs 1 to 3. As is clear from the Figures, LR classifier is a competitive algorithm to the best performers in their research.

Although [9] have worked on the same dataset, we used in this research, with exactly the same training and testing split, differences in the number of features used for building classifiers, FS and weighting methods adopted, as well as in the text pre-processing steps applied on the dataset documents make direct performance comparisons between our LR classifier and their classifiers unfair, since these differences are known to affect the classification performance to a great extent. Our intended near future work is to conduct direct comparisons between LR classifier and other well-known Arabic text classifiers using the same TC settings.

Other research works on Alj-News Arabic Dataset used different set of classes, different number of documents or different splits for training and testing subsets. We present here a comparison of the results on the common classes in our and their research experiments.

Ref. [74] used a version of Alj-News dataset with 16 categories, 7566 documents and 189815 features to test 3 algorithms on Arabic TC: SVM, kNN and GIS (Generalized Instance Set). Results of their experiments are summarized in Table 8.

Table 4. Results of Classifying Alj-News using LR

Class	Precision	Recall	F-Measure
Art	90.7407	81.6667	85.9649
Economic	80.6452	83.3333	81.9672
Politics	73.913	85	79.0698
Science	93.1034	90	91.5254
Sport	**96.4912**	**91.6667**	**94.0171**
MicroAverage	**86.3333**	**86.3333**	**86.3333**
MacroAverage	**86.9787**	**86.3333**	**86.5089**

Table 5. Accuracy by Class for SVM on Alj-News Dataset in [9]

Class	Precision	Recall	F-Measure
Art	93.4	95	94.2
Economic	96.2	85	90.3
Politics	78.9	93.3	85.5
Science	100	93.3	96.6
Sport	100	98.3	99.2
W. Avg.	93.7	93	93.1

Table 6. Accuracy by Class for Na ïve Bayes on Alj-News Dataset in [9]

Class	Precision	Recall	F-Measure
Art	86	71.7	78.2
Economic	85.2	86.7	86
Politics	66.2	85	74.5
Science	91.4	88.3	89.8
Sport	100	90	94.7
W. Avg.	**85.8**	**84.3**	**84.6**

Table 7. Accuracy by Class for J48 on Alj-News Dataset [9]

Class	Precision	Recall	F-Measure
Art	71.1	53.3	61
Economic	78.9	75	76.9
Politics	47.1	66.7	55.2
Science	84.9	75	79.6
Sport	91.7	91.7	91.7
W. Avg.	**74.7**	**72.3**	**72.9**

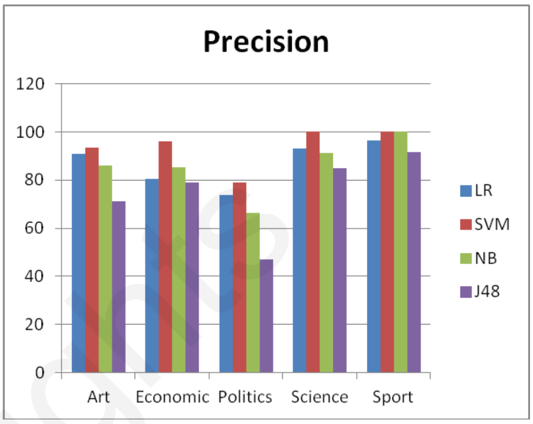

Fig. 1. Precision of LR versus others in [9] per class

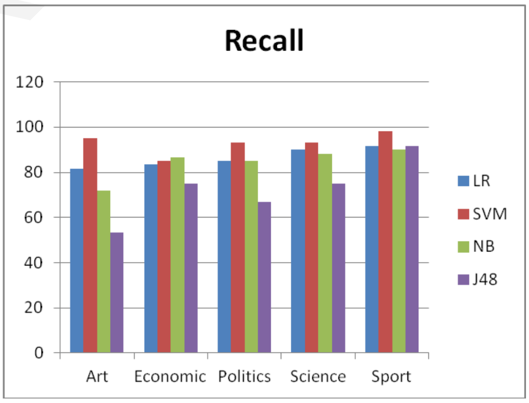

Fig. 2. Recall of LR versus others in [9] per class

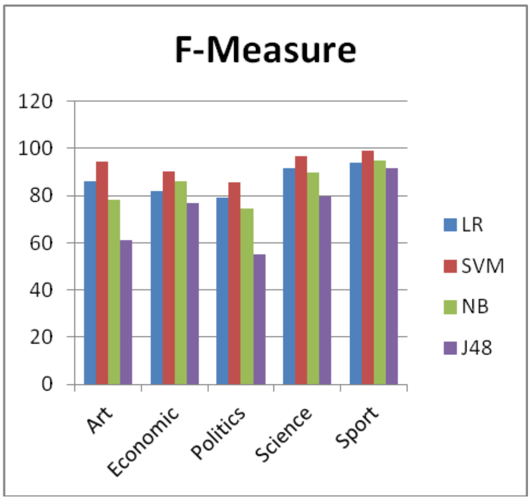

Fig. 3. F-Measure of LR versus others in [9] per class

Table 8. Results of research work of [74] on Alj-News Dataset

Algorithm	Precision	Recall	F1
SVM	78.1316	86.1111	81.9314
KNN	83.814	85.5740	84.6849
GIS	84.5085	85.3060	84.9054

Ref. [5] tested CHI FS in Arabic TC using 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 consists of 1445 documents. These documents fall into nine classification categories that vary in the number of documents. Data preprocessing was applied by removing digits, punctuation marks, non-Arabic letters, stop words and infrequent terms which occur less than 4 times in the training part of the corpus. In addition, Light Stemming was applied. His best results, which were achieved when extracting the top 162 terms for each classification class, are presented in Table 9 for the common classes between his and our research works. The overall performance of the three algorithms used in his research is summarized in Table 10.

Table 9. Results of research work [5] on Alj-News Dataset

Category	Precision	Recall	F-measure
Economics	93.02326	71.42857	80.80808
Politics	90	76.27119	82.56881
Sports	100	85.71429	92.30769

Table 10. Overall F-Results in research work [5] on Alj-News Dataset

Algorithm	F-measure
SVM	88.11
NB	84.54
kNN	72.72

It is apparent from these indirect comparisons that LR is a competitive Arabic TC algorithm using much less number of features (only 135 features compared to hundreds of thousands of features in other researches).

VI. Conclusion

In this research, Logistic Regression (LR) is investigated in Arabic Text Categorization (TC) for the first time in the literature of Arabic TC. Experiments are conducted on the widely-used Arabic Text Categorization Dataset (Alj-News). Chi Square is used for feature selection and a local policy is used to select a reduced feature set for building the LR classifier (only 1% of each class features). Using this very small feature set, LR has recorded very accurate classification performance (precision of 96.49, recall of 91.67 and F1-measure of 94.0171). In fact, these results conclude that LR is a promising competitive Arabic TC algorithm to the state-of-the-art ones in this field. It can be used for classifying larger datasets successfully as long as good Feature Selection and Reduction criteria are applied on the dataset. Our intended near future work is to compare LR directly to the state-of-the-art algorithms in this field using the same classification settings on larger datasets.

References

[1] http://www.InternetWorldStats.com (Accessed November, 2014).

[2] Yahyaoui M. “Toward an Arabic web page classifier”. Master project. *AUI.* 2001.

[3] El-Kourdi M, Bensaid A and Rachidi T. “Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm”. *In the 20th Int. Conf. on Computational Linguistics*, Geneva, August, 27, 2004.

[4] Duwairi R. “Arabic Text Categorization”. *International Arab Journal of Information Technology*, 2007; 4(2): 125 – 131. doi: 10.1002/asi.20360.

[5] Mesleh A. A. “Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System”. *Journal of Computer Science*, 2007; 3(6): 430 - 435.

[6] Al-Harbi S, Almuhareb A, Al-Thubaity A, Khorsheed M and Al-Rajeh A. “Automatic Arabic Text Classification”. *In: JADT’08*. France; 2008. pp 77 - 83.

[7] El-Halees A. M. “A Comparative Study on Arabic Text Classification”. *Egyptian Computer Science Journal*, 2008; 30(2).

[8] Al-Saleem S. “Associative Classification to Categorize Arabic Data Sets*”. International Journal of ACM Jordan,* 2010; 1(3): 118 - 127.

[9] Chantar H. K. and Corne D.W. “Feature subset selection for Arabic document categorization using BPSO-KNN”, *IEEE* , 2011. pp. 546 - 551. doi: 10.1109/NaBIC.2011.6089647.

[10] Al-Shalabi R, Kanaan G and Gharaibeh H. “Arabic text categorization using KNN algorithm”. *In the Proc. of Int. multi conf. on computer science and information technology CSIT06*; 2006.

[11] Harrag F, El-Qawasmeh E and Pichappan P. “Improving Arabic text categorization using decision trees”. *In IEEE, NDT '09*, 2009. pp 110 – 115. doi:10.1109/NDT.2009.5272214.

[12] Duwairi R. “A Distance-based Classifier for Arabic Text Categorization”. *In the Proc. of the Int. Conf. on Data Mining DMIN’05*, Las Vegas, USA; June, 2005. pp 20-23.

[13] Ghwanmeh S. “Applying Clustering of Hierarchical K-means-like Algorithm on Arabic Language”. *The Int. Journal of Information Technology* 2007; 3(3): 168-172.

[14] Kanaan G, Al-Shalabi R and Ghwanmeh S. “A comparison of text-classification techniques applied to Arabic text”. *Journal of the American Society for Information Science and Technology*, 2009; 60(9): 1836 – 1844. doi:10.1002/asi.v60:9.

[15] Khreisat L. “Arabic Text Classification Using N-Gram Frequency Statistics: A Comparative Study”. *In the Proceedings of the 2006 International Conference on Data Mining (DMIN 2006)*, Las Vegas, Nevada, USA; June 26-29, 2006. pp 78 - 82.

[16] AL-Tahrawi M. M. and Al-Khatib S. N. “Arabic Text Classification Using Polynomial Networks”. *Journal of King Saud University - Computer and Information Sciences*. In Press.

[17] Cooper W. S., Gey F. C. and Dabney D. P. “Probabilistic retrieval based on staged logistic regression”. *In: SGIR 92*, pp. 198–210, 1992.

[18] Fuhr N and Pfeifer U. "Combining model-oriented and description-oriented approaches for probabilistic indexing". *In: SIGIR 91*, pp. 46–56. 1991.

[19] Gey F. C. "Inferring probability of relevance using the method of logistic regression". *In: SIGIR 94*, pp.222–231, 1994.

[20] Ittner D. J., Lewis D. D. and Ahn D. D. "Text categorization of low quality images". *In: Symposium on Document Analysis and Information Retrieval,* pp. 301–315. 1995.

[21] Lewis D. D. and Gale W. A. "A sequential algorithm for training text classifiers". *In: SIGIR 94*, pp. 3–12. 1994.

[22] Sch ̈utze H., Hull D. A. and Pedersen J. O. "A comparison of classifiers and document representations for the routing problem". *In: SIGIR 95*, pp. 229–237. 1995.

[23] Alexander GENKIN DIMACS and David D. LEWIS. "Large-Scale Bayesian Logistic Regression for Text Categorization". *American Statistical Association and the American Society for Quality Technometrics*, August 2007, 49(3). DOI 10.1198/004017007000000245.

[24] Andrew Gelman and Jennifer Hill. "Data Analysis Using Regression and Multilevel/Hierarchical Models". *Cambridge University Press*. 2007.

[25] Amrita Paul. "Effect of imbalanced data on document classification algorithms". *Master Thesis. Auckland University of Technology*. 2014.

[26] Yiming Yang and Xin Liu. "A re-examination of text categorization methods*". In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), pp. 42-49, doi:10.1145/312624.312647

[27] P. Komarek and A. Moore. "Fast robust logistic regression for large sparse datasets with binary outputs". *In Proceedings of the International Workshop on Artificial Intelligence and Statistics*, New York,NY, 2003.

[28] Sujeevan Aseervathama,, Anestis Antoniadisb, Eric Gaussiera, Michel Burletc and Yves Denneulind. "A Sparse Version of the Ridge Logistic Regression for Large-Scale Text Categorization". *Pattern Recognition Letters (01 October 2010)*. doi:10.1016/j.patrec.2010.09.023

[29] Andrew Y. Ng, Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *In Neural Information Processing Systems*, 2002.

[30] Hanuman Thota , Raghava Naidu Miriyala , Siva Prasad Akula, .Mrithyunjaya Rao , Chandra Sekhar Vellanki ,Allam Appa Rao and Srinubabu Gedela. "Performance Comparative in Classification Algorithms Using Real Datasets". *JCSB*, Vol.2, February, 2009.

[31] Paul Komarek and , Andrew Moore . "Fast Logistic Regression for Data Mining, Text Classification and Link Detection". *Proceedings of NIPS2003* .

[32] Arild Brandrud Næss. "Bayesian Text Categorization". *MASTER'S THESIS, Norwegian University of Science and Technology*, 2007

[33] Tong Zhang and Frank J. Oles."Text Categorization Based on Regularized Linear Classification Methods". *Information Retrieval*, Vol. 4, pp. 5–31, 2001.

[34] Alexander Genkin, David D. Lewis AND David Madigan. "Sparse Logistic Regression for Text Categorization" . *Working Group on Monitoring Message Streams Project Report, April 2005*. 2005

[35] Georgiana Ifrim, G ökhan Bakir and Gerhard Weikum. "Fast Logistic Regression for Text Categorization with Variable-Length N-grams". *In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 354-362, 2008. doi:10.1145/1401890.1401936

[36] Weiwei Cheng and Eyke H ullermeier. "Combining Instance-Based Learning and Logistic Regression for Multilabel Classification". *Journal Machine Learning*, Volume 76 Issue 2-3, pp. 211 – 225, 2009. doi:10.1007/s10994-009-5127-5.

[37] http://filebox.vt.edu/users/dsaid/Alj-News.tar.gz. (January, 2014)

[38] Vapnik V. Statistical learning theory. New York: Wiley; 1998.

[39] Zhang J, Jin R, Yang Y and Hauptmann A. "Modified logistic regression: an approximation to SVM and its applications in large-scale text categorization". *In: Proc Twentieth Int Conf Machine Learning (ICML 2003)*, Washington, DC USA, August, 2003; pp. 21–24.

[40] Hoi SCH, Jin R, Lyu M.R. "Large-scale text categorization by batch mode learning". *In: Proc 15th Int World Wide Web conference (WWW2006)*, Edinburgh, England, UK, May, 2006.

[41] Komarek P and Moore A. "Making logistic regression a core data mining tool: a practical investigation of accuracy, speed, and simplicity". Technical Report TR-05—27, Robotics Institute, Carnegie Mellon University, May 2005.

[42] R. O. Duda and P. E. Hart. "Pattern Classiffication and Scene Analysis". *Wiley-Interscience*, New York, 1973.

[43] D. D. Lewis. "Evaluating and optimizing autonomous text classification systems". *In SIGIR '95*: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 246-254, 1995.

[44] P. J. Green. "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robustand Resistant Alternatives". *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2), pp. 149-192, 1984.

[45] Said D, Wanas N, Darwish N and Hegazy N. "A Study of Arabic Text preprocessing methods for Text Categorization". *In the 2nd Int. conf. on Arabic Language Resources and Tools*, April, 22-23, Cairo, Egypt, 2009; pp 230-236.

[46] Mohamed S, Ata W and Darwish N. "A new technique for automatic text categorization for Arabic documents". *In Proc. of the 5th IBIMA International Conference on Internet and Information Technology in Modern Organizations*, Cairo, Egypt; 2005. pp 13–15.

[47] Khoja S and Garside R. "Stemming Arabic text". Computing Department, Lancaster University, Lancaster; 1999.http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps (Accessed January 2014)

[48] http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip. (January, 2014)

[49] Sawalha M and Atwell E. "Comparative evaluation of Arabic language morphological analyzers and stemmers". *In: the Proc. of COLING'2008 22nd Int. Conf. on Computational Linguistics*, (poster volume); 2008. pp 107-110.

[50] Larkey L and Connell ME. "Arabic information retrieval at UMass in TREC-10". *In: Proceedings of TREC, Gaithersburg: NIST; 2001*. doi:10.1.1.14.9079.

[51] Joachims T. "Text categorization with support vector machines: learning with many relevant features". *Proc 10th Euro Conf Machine Learning (ECML) Springer-Verlag* London, UK 1998;1398:137–142.

[52] Brank J, Grobelnik M, Milic-Frayling N and Mladenic D. "Interaction of feature selection methods and linear

classification models", *Workshop on Text Learning held at ICML-2002*; 2002.

[53] Rogati M and Yang Y. "High- performing term selection for text classification". *CIKM'McLean, Virginia, USA, November, 2002*, pp. 4–9.

[54] Bekkerman R. "Distributional clustering of words for text categorization". Master's thesis, *CS Department, Technion-Israel* Inst. of Technology; 2003.

[55] Khorsheed M and Al-Thubaity A. "Comparative evaluation of text classification techniques using a large diverse Arabic dataset". *Lang Resources & Evaluation*, Springer, 2013; 47(2):513-538. doi: 10.1007/s10579-013-9221-8.

[56] Belkebir R and Guessoum A. "A Hybrid BSO-Chi2-SVM Approach to Arabic Text Categorization*". In IEEE Computer Systems and Applications (AICCSA), 2013 ACS International Conference, Ifrane* , 27-30 May 2013; pp 1-7. doi: 10.1109/AICCSA.2013.6616437 .

[57] Sharef B, Omar N and Sharef Z. "An Automated Arabic Text Categorization Based on the Frequency Ratio Accumulation". *The International Arab Journal of Information Technology*, 2014; 11(2): 213-221.

[58] Thabtah F, Eljinini M, Zamzeer M and Hadi W. "Na ïve Bayesian based on Chi Square to Categorize Arabic Data". *In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies*, Cairo, Egypt ; 2009. pp 930 - 935. doi:10.1.1.411.3605.

[59] Fodil L, Sayoud H and Ouamour S. "Theme Classification of Arabic Text: A Statistical Approach". *In Terminology and Knowledge Engineering 2014*, Berlin : Germany; 2014. pp 77-86.

[60] Sawaf H, Zaplo J and Ney H. "Statistical classification methods for Arabic news articles". *Arabic Natural Language Processing Workshop, ACL'2001,* Toulouse, France. 2001; pp 127–132.

[61] AL-Tahrawi M. M. and Abu Zitar R. "Polynomial networks versus other techniques in text categorization". *Int J Patt Recog Artif Intell (IJPRAI) 2008*; 22(2):295–322. doi: 10.1142/S0218001408006247.

[62] AL-Tahrawi M. M. "The Significance of Low Frequent Terms In Text Classification". *International Journal of Intelligent Systems*, 2014; 29(5): 389 – 406. doi: 10.1002/int.21643.

[63] AL-Tahrawi M. M. "Class-Based Aggressive Feature Selection For Polynomial Networks Text Classifiers – An Empirical Study". *UPB Scientific Bulletin, Series C*, In Press.

[64] Al-Tahrawi M. M. "The role of rare terms in enhancing the performance of polynomial networks based text categorization". *J Intell Learn Syst Appl* 2013;5:84–89. doi: 10.4236/jilsa.2013.52009.

[65] Eldos M. "Arabic Text Data Mining: A Root Extractor for Dimensionality Reduction". *ACTA Press, A scientific and Technical Publishing Company*; 2002.

[66] Eldin S. "Development of a computer-based Arabic Lexicon". *In the Int. Symposium on Computers & Arabic Language, ISCAL, Riyadh, KSA*; 2007.

[67] Zheng Z, Wu X and Srihari R. "Feature selection for text categorization on imbalanced data". *SIGKDD Explorations, ACM*, New York, NY, USA, 2004; 6(1):80–89. . doi:10.1.1.103.5069.

[68] Ababneh J, Almomani O, Hadi W, Kamel N, El-Omari T and Al-Ibrahim A. "Vector Space Models to Classify Arabic Text". *International Journal of Computer Trends and Technology (IJCTT)*, 2014; 7(4): 219-223.

[69] Lewis D. D. and Ringuette M. "A comparison of two learning algorithms for text categorization". *In: Proc Third Ann Symp Document Analysis and Information Retrieval (SDAIR'94)*, Las Vegas, USA, 1994; pp. 81–93. doi:10.1.1.49.860.

[70] El-Halees A. M. "Arabic Text Classification Using Maximum Entropy". *The Islamic University Journal*, 2007; 15(1): 157 - 167. doi: 10.1.1.124.361.

[71] Al-Saleem S. "Automated Arabic Text Categorization Using SVM and NB". *International Arab Journal of e-Technology*, 2011; 2(2): 124-128.

[72] Debole F and Sebastiani F. "An analysis of the relative hardness of Reuters-21578 subsets". *JASIS*; 2005. 56(6): 584–596.

[73] Van Rijsbergen C. J. "Information retrieval". 2nd edn. London: *Butterworths*; 1979.

[74] Awad, W. A. "Machine Learning Algorithms in Web Page Classification". *International Journal of Computer Science & Information Technology (IJCSIT),* 2012, 4(5), 93-101. doi: 10.5121/ijcsit.2012.4508.

Author's Profiles

Mayy M. Al-Tahrawi obtained her B.Sc. degree in Computer Science from Kuwait University in 1986, her M.Sc. also in Computer Science (Compiler Generators from Kuwait University in 1988, and her Ph.D. in Computer Information Systems from The Arab Academy for Banking and Financial Sciences, Jordan in 2006. She has worked as a lecturer in several universities in Jordan and Kuwait since 1988. Currently, she is a Senior Assistant Professor of Computer Science at Al-Ahliyya Amman University, Amman, Jordan. Her research interests are in Machine Learning, Pattern Recognition, Feature Selection, Text Categorization and Information Retrieval.