A Comparative Study of Effective Supervised Learning Methods on Arabic Text Classification

Rachid Sammouda[1,*]

[1] Department of Computer Science, King Saud University, Riyadh, Saudi Arabia

Summary
Nowadays, Arabic Text Classification (ATC) is attracting researchers' attention in many fields, including text mining, web search, social media, security, and other fields. Text Classification or Categorization (TC) is the process of classifying text documents to proper categories based on their contents. Few studies have been developed for the comparison of supervised learning (SL) methods on ATC. Consequently, this paper is concerned with ATC of Arabic documents. The proposed approach adopted for this comparative study consists of three steps: (i) document pre-processing step where Arabic stop words, punctuations, diacritics, common prefix and suffix (Arabic words light stemmer) are removed from the Arabic documents, (ii) document filtering step where the words strings are converted into number of individual words vectors using term frequency transform (TFT) technique, inverse document frequency transform (IDFT) technique and both, (iii) classification step where a comparison of eight effective known SL methods is adopted for ATC. The impact of using TFT, IDFT and both on the effectiveness of these SL methods is also studied. The results show that the accuracy of 10-fold cross validation test mode obtained by LSVM classifier with IDFT technique is the highest compared to other SL methods used in this study. This outcome can be used in the future as a guidance for developers of ATC applications.

Key words:
Text Classification of Arabic documents, Supervised Learning Methods, Arabic Light Stemmer, and Weka Tool.

1. Introduction

Text classification is becoming a significant task tool of many applications which deal with a huge text information. Large scale of electronic text documents of many languages are growing developing every day [1]. Arabic text documents are part of this growth and there is a need to store, retrieve and mine in many tasks of different applications. In the literature [2], a text classification system was proposed for Arabic language. This system used the N-grams (unigrams and bigrams) and single terms (bag of words) were also used as a representation of Arabic documents in the pre-processing step. Later, the k-nearest neighbours (kNN) classifier was employed for Arabic text classification. The results illustrated that the documents representation by using unigrams and bigrams outperformed the representation based on the bag of words in terms of accuracy.

Srivastava et al. [3] proposed a statistical method named Maximum Entropy method to classify Arabic news documents. A multi-word term extraction method for Arabic text was proposed in [4]. In this method, multi-word terminology from Arabic corpus was extracted. From the respective of linguistic, some linguistic needs to filter and extract the candidates of multi-word terminology. Abainia et al. [5] proposed to use the support vector machine (SVM) classifier to classify the Arabic text documents. The results of SVM were the highest compared to the results of kNN classifier.

Moreover, Hmeidi et al. [6] studied the influence of khoja root-based stemmer and light stemming on the results of naïve Bayes (NB), SVM, kNN, decision tree (J48), and Decision Table (DT) classifiers for ATC. The conclusion showed that the results of SVM and NB classifiers with light stemming were better than other classifiers. The same deduction was drawn up by the works of Al-Badarneh [13] and Ayedh et al. [7] based on several pre-processing methods. In addition, Al-Molegi et al. [8] and Khreisat [9] have developed some approaches for ATC using the combination of N-grams with Manhattan, Euclidean distances and Dice measures. The results showed that the combination of tri-gram with Dice measure attained a high performance.

Al-Anzi et al. [10] presented a method for Arabic text classification based on a Latent Semantic Indexing (LSI) with clustering techniques where the similar unlabelled documents are grouped into a pre-defined number of classes. The results exposed that the method labelled the documents without any training dataset. Another research of Al-Anzi et al. was in [11]. This work developed a new method for ATC using a LSI and cosine similarity. The results showed that the features of LSI were significantly achieved with better accuracy than the features of TF-IDF. Also, these results proved that the kNN and SVM with the cosine measure attained the highest performance. Although the most works in the literature have achieved an acceptable performance, Arabic language is a rich and ancient language that needs robust text classification algorithms to deal with different aspects of the Arabic language, such as morphology, vocabulary and syntax. The authors in [12] addressed some of these challenges. Furthermore, Al-Anzi

Manuscript received December 5, 2017
Manuscript revised December 20, 2017

et al. [13] proposed to use the conventional TF-IDF with different machine learning classifiers for ATC. The remaining part of this paper is organized as follows: Section 2 explains the proposed approach in more details. Experiment and discussion are presented in Section 3. Finally, Section 4 summarizes the conclusion of this work.

2. Proposed Approach

The proposed approach aims to find out the best SL model which achieves high accuracy for ATC system. It consists of three steps as shown in Fig. 1.

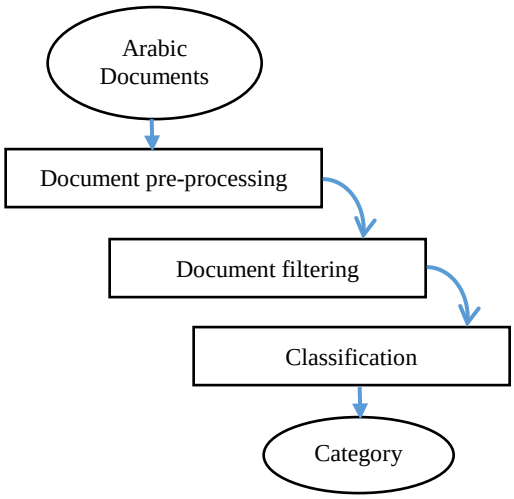

Fig. 1 Proposed Approach for ATC system.

For every document in the dataset, the pre-processing step is responsible to remove stop words, punctuations, diacritics, common prefix and suffix (words light stemmer) from it.

In document filtering step, we apply three transforms to convert Arabic text words attributes into a set of attributes representing words occurrences. They are TFT, IDFT and TFT-IDFT. The reason for using these three representations is to show the effect of each representation on the next step.

In TFT, the Arabic word frequencies are transformed into $log(1 + fij)$, where fij is the frequency of word i in document j. IDFT transformed the word frequencies into $fij * log\left(\frac{num\ of\ Docs}{num\ of\ Docs\ with\ word\ i}\right)$, where fij is the frequency of word i in document j.

In the final step, we use a set of eight effective well known classifiers to show which one is efficient for ATC system. The eight classifiers adopted in this study are: (1) k-nearest neighbors (kNN), (2) naïve Bayes (NB), (3) linear support vector machine (LSVM), (4) decision Tree (J48), (5) Bayes Network (BN), (6) Random Forest (RF), (7) Random Tree (RT) and (8) Random Committee (RC).

3. Experiment and Discussion

3.1 Data Set Description

We select a dataset of five categories from the publicly corpus dataset, namely "Arabic news articles", collected by Diab Abu Aiadh [14]. The selected dataset contains 1500 documents belonging to five different categories (Art, Economy, Health, Law and Politics), each one contains 300 documents. Table 1 shows the statistics of the selected corpus dataset.

Table 1: Selected Dataset Statistics.

Name of Category	***Number of Documents***
Art	300
Economy	300
Health	300
Law	300
Politics	300
Total	1500

3.2 Tool Description

We implement our experiment using the Waikato Environment for Knowledge Analysis (WEKA), developed at the University of Waikato in New Zealand [15]. We use it because of its popularity for machine learning and data mining fields.

3.3 Evaluation Metrics

10-fold cross-validation test mode is employed in the evaluation process. The dataset is divided into 10 folds, each fold is used for testing and other folds for training. The average result is used as a final result. Three popular metrics are computed for evaluating the output results. These metrics are the *accuracy (ACC)*, *precision* (P) and *recall (R)*.

$$Acc = (TP + TN)\ (TP + FP + TN\ /\ + FN) \quad (1)$$

$$P = TP\ /\ (TP + FP) \quad (2)$$

$$R = TP\ /\ (TP + FN) \quad (3)$$

Where TP is the true positive rate, FP is the false positive rate, TN is the true negative rate and FN is the false negative rate.

3.4 Results

Here, we computed the accuracy, precision and recall of all examples in the testing dataset folds. From Table 2, we see

that BN and RF with TFT achieved the highest results with 95.2667% and 96.0667%, respectively, compared to all other classifiers. In contrast, the results in Table 3 and Figure 2 show that the LSVM classifier with IDFT achieved the highest accuracy with 97.3333%, compared to all classifiers for all transformation methods.

Table 2: Results of TFT representation with the eight classifiers.

Classifier Model	*Accuracy (%)*	*Weighted Avg. of Precision*	*Weighted Avg. of Recall*
kNN	72.5333	0.847	0.725
NB	93.5333	0.937	0.935
LSVM	89.2667	0.921	0.893
J48	83.9333	0.840	0.839
BN	95.2667	0.955	0.953
RF	96.0667	0.961	0.961
RT	67.5333	0.679	0.675
RC	90.5333	0.905	0.905

Table 3 shows the results of IDFT for the eight classifiers used in the experiment.

Table 3: Results of IDFT representation with the eight classifiers.

Classifier Model	*Accuracy (%)*	*Weighted Avg. of Precision*	*Weighted Avg. of Recall*
kNN	73	0.851	0.730
NB	93.8	0.939	0.938
LSVM	97.3333	0.974	0.973
J48	84.8	0.848	0.848
BN	95.4667	0.957	0.955
RF	95.4	0.955	0.954
RT	66.8	0.667	0.668
RC	89.8	0.897	0.898

Table 4 shows the results of TF-IDFT for the eight classifiers used in the experiment.

Table 4: Results of TF-IDFT representation with the eight classifiers.

Classifier Model	*Accuracy (%)*	*Weighted Avg. of Precision*	*Weighted Avg. of Recall*
kNN	73	0.851	0.730
NB	93.8	0.939	0.938
LSVM	96.8	0.969	0.968
J48	84.8	0.848	0.848
BN	95.4667	0.957	0.955
RF	95.9333	0.960	0.959
RT	67.2667	0.676	0.673
RC	90.4667	0.904	0.905

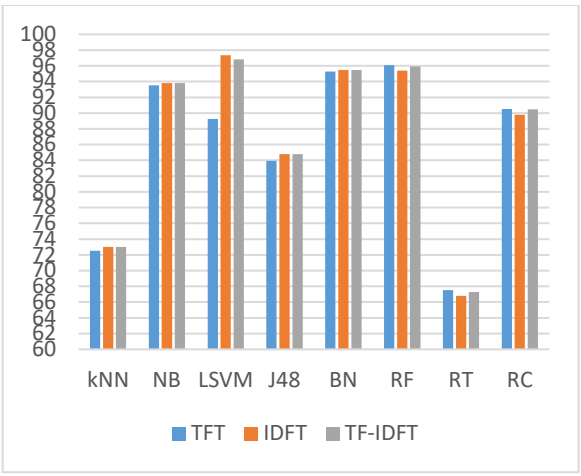

Fig. 2: The effect of TFT, IDFT and TF-IDFT on the accuracy of the eight classifiers used in the experiment.

From Figure 3 and 4, we see that LSVM, BN and RF achieved the highest average precision and recall values compared to others classifiers.

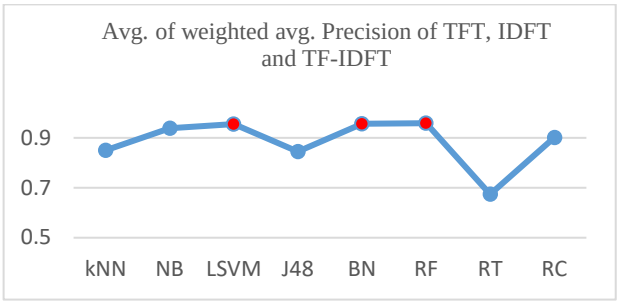

Fig. 3: The Average of weighted avg. precision of TFT, IDFT and TF-IDFT for the eight classifiers used in the experiment.

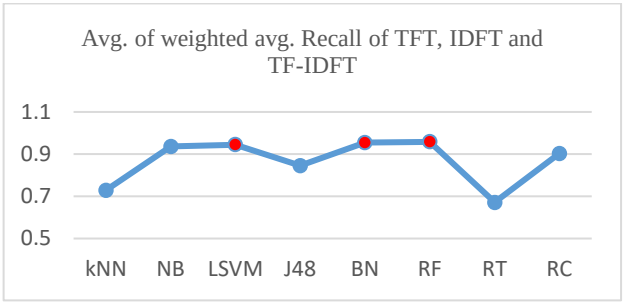

Fig. 4: The Average of weighted avg. recall of TFT, IDFT and TF-IDFT for the eight classifiers used in the experiment.

4. Conclusion and future work

In this paper, a text classification task for Arabic documents is addressed. It proposed a new approach for a comparative study of popular SL methods on ATC problem. This

approach consists of three steps which are: (i) document pre-processing step where Arabic stop words, punctuations, diacritics, common prefix and suffix (Arabic words light stemmer) are removed from the Arabic documents, (ii) document filtering step where the words strings are converted into number of individual words vectors using term frequency transform (TFT) technique, inverse document frequency transform (IDFT) technique and both, and (iii) classification step where a comparison of eight effective well known SL methods is used to classify the Arabic text.

Finally, we conducted several experiments on the publicly "Arabic news articles" corpus dataset via Weka tool. The results show that the accuracy of 10-fold cross validation test mode obtained by LSVM classifier with IDFT technique is the highest compared to other SL methods. In future work, we will propose a new SL method to classify Arabic text retrieved from a large database by the method proposed in [16].

Acknowledgment

This work was supported by a special fund in the Research Center of the College of Computer and Information Sciences (CCIS) at King Saud University.

References

[1] Al-Shargabi, B., Olayah, F. and Romimah, W.A., (2011) "An experimental study for the effect of stop words elimination for arabic text classification algorithms", International Journal of Information Technology and Web Engineering (IJITWE), Vol. 6, No. 2, pp.68-75.

[2] Al-Shalabi, R. and Obeidat, R., (2008) "Improving KNN Arabic text classification with n-grams based document indexing", In: Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, pp. 108-112.

[3] Srivastava, A.N. and Sahami, M. eds., (2009) Text mining: Classification, clustering, and applications, CRC Press.

[4] Park, H.H., Park, J. and Kwon, Y.B., (2015) "Topic clustering from selected area papers", Indian Journal of Science and Technology, Vol. 8, No. 26.

[5] Abainia, K., Ouamour, S. and Sayoud, H., (2015) "Neural Text Categorizer for topic identification of noisy Arabic Texts", In: Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of, IEEE, pp. 1-8.

[6] Hmeidi, I., Al-Ayyoub, M., Abdulla, N.A., Almodawar, A.A., Abooraig, R. and Mahyoub, N.A., (2015) "Automatic Arabic text categorization: A comprehensive comparative study", Journal of Information Science, Vol. 41, No. 1, pp.114-124.

[7] Al-Badarneh, A., Al-Shawakfa, E., Bani-Ismail, B., Al-Rababah, K. and Shatnawi, S., (2017) "The impact of indexing approaches on Arabic text classification", Journal of Information Science, Vol. 43, No. 2, pp.159-173.

[8] Ayedh, A., Tan, G., Alwesabi, K. and Rajeh, H., (2016), "The effect of preprocessing on arabic document categorization", Algorithms, Vol. 9, No. 2, p.27.

[9] Al-Molegi, A., Izzat Alsmadi, H.N. and Albashiri, H., (2015), "Automatic learning of arabic text categorization", Int. J. Digit. Contents Appl, Vol. 2, No. 1, pp.1-16.

[10] Khreisat, L., (2009), "A machine learning approach for Arabic text classification using N-gram frequency statistics", Journal of Informetrics, Vol. 3, No. 1, pp.72-77.

[11] Al-Anzi, F.S. and AbuZeina, D., (2016) "Big data categorization for arabic text using latent semantic indexing and clustering", In: International Conference on Engineering Technologies and Big Data Analytics (ETBDA 2016), pp. 1-4.

[12] Al-Anzi, F.S. and AbuZeina, D., (2017), "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing", Journal of King Saud University-Computer and Information Sciences, Vol. 29, No. 2, pp.189-195.

[13] Al-Anzi, F.S. and AbuZeina, D., (2015), "Stemming impact on Arabic text categorization performance a survey", In: Information & Communication Technology and Accessibility (ICTA), 2015 5th International Conference on, IEEE, pp. 1-7.

[14] Diab Abu Aiadh, "Dataset for Arabic document classification", 2017. [Online]. Available: http://diab.edublogs.org/dataset-for-arabic-document-classification/. [Accessed: 21- Dec- 2017].

[15] WEKA, "Data Mining Software in Java", 2017. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka. [Accessed: 21- Dec- 2017].

[16] Gumaei, A., Sammouda, R., and Al-Salman, A. S. (2017), "An Efficient Algorithm for K-Rank Queries on Large Uncertain Databases", International Journal of Computer Science and Network Security (IJCSNS), Vol. 17, No.4, pp. 129.

Rachid Sammouda, Associate professor, computer science educator & researcher. Accomplished career demonstrating consistent success as an Educator and relentless Researcher at higher education levels. Outstanding track record in assuring student success with high teaching skills. Seasoned in conceiving and building programs from the ground up through proven competencies in projects and program management, and staff support and empowerment. Effective communicator with excellent planning, organizational, and negotiation strengths as well as the ability to lead, reach consensus, establish goals, and attain results. Specialist whose qualifications include a PhD degree in Artificial Intelligence; with detailed knowledge in Digital Image Processing, Computer Sciences, Programming Languages, Artificial Intelligence and Computing with Artificial Neural Networks. Several years of experience in the creation and deployment of solutions for Medical Image Processing and Analysis for diverse medical purposes. Leading groups in successful conferences organizations across a wide platform of topics.