Arabic Text Categorization using Machine Learning Approaches

Riyad Alshammari
College of Public Health and Health Informatics King Saud Bin Abdulaziz University for Health Sciences
P.O. Box 22490, Riyadh 11426 Kingdom of Saudi Arabia

***Abstract*—Arabic Text categorization is considered one of the severe problems in classification using machine learning algorithms. Achieving high accuracy in Arabic text categorization depends on the preprocessing techniques used to prepare the data set. Thus, in this paper, an investigation of the impact of the preprocessing methods concerning the performance of three machine learning algorithms, namely, Naïve Bayesian, DMNBtext and C4.5 is conducted. Results show that the DMNBtext learning algorithm achieved higher performance compared to other machine learning algorithms in categorizing Arabic text.**

***Keywords*—Arabic text; categorization; machine learning**

I. Introduction

Constructing an automated text categorization system for Arabic articles/documents is a difficult work as a result of the unique nature of the Arabic language. Arabic language consists of 28 letters and is written from right to left. It has a distinctive morphology and orthography principles.

The number of text information accessible on the Internet has increased rapidly on the last few years since many private and public organizations are publishing their text information such as documents, news, books, etc. on the World Wide Web (WWW). This creates a vast amount of text information that makes the manual categorization of text information a very impractical task. Thus, the development of automated text categorization/classification system is important work.

Text categorization is the automatic mapping of texts to predefine labels (classes). Text categorization methods are deployed in many systems such as searching e-mails, spam filtering, and classification of news. Most of the text categorization methods are developed to deal with texts written in the English language. Hence, there are not suitable to documents written in the Arabic language.

Several works had been conducted on this topic using machine learning algorithms such as decision trees [1]–[3], k-nearest neighbour [4]–[8], Support Vector Machines (SVMs) [5], [9] regression model [10], [11], and other techniques and algorithms [12]–[14]. The previous study compared the performance of the classifiers based on the English articles utilizing the Reuters Newswire Corpus.

Several papers had focused on the use of machine learning algorithms to categorize Arabic Text. El-Kourdi et al. [15] used Naïve Bayes machine learning Algorithm to classify Arabic website data. They achieved an accuracy of ≈ 69%. Mesleh [16] utilized three machine learning algorithms (SVM, KNN and Naïve Bayes) to classify Arabic data that were collected from Arabic newspaper. They divided the texts into nine labels that were Economics, Education, Computer, Engineering, Medicine, Law, Politics Sports and Religion. He used Chi-square statistics as the feature sets. He achieved high performance with F-measure equal to 88.11%. Furthermore, Syiam et al. [17] applied many stemming techniques and different features selection for Arabic text categorization. They discovered that combination of light stemmer and statistical method achieved the best performance.

On the other hand, few studies focus their study on the performance of automatically classifying Arabic language using Arabic corpora [18], [19]. This is due to the fact that the Arabic language is highly abundant in grammars and needs special handlings for morphological analysis. Furthermore, constructing an automated text categorization system for Arabic articles/documents is a challenging work as a result of the unique nature of the Arabic language. The Arabic language has complicated morphological principles comparing to English. The root of the words can have either tri-letter (most of the words), Quad-letter, Penta-letter or even Hexa-letter [20]. As a result, the Arabic Language demands enormous processing to construct an accurate categorization system [21]. Thus, Arabic text categorization is considered as a very challenging task for researchers because of the language complexity.

In this paper, an automated categorization system has been introduced based on machine learning algorithms for Arabic text documents. The impact of the preprocessing techniques related to the term weighing schemes for Arabic Text has not been studied in the literature. In this research paper, we focus on exploring this impact on Arabic corpora to improve the categorization accuracy by investigating different machine learning approaches, mainly Naïve Bayesian, DMNBtext and C4.5 algorithms.

The rest of this paper is organized as follows. Section II presents the methods that includes the machine learning algorithms employed, details description of the data sets, features and evaluation criteria. The experimental results and analysis are discussed in Section III. Finally, conclusions are drawn and future work is discussed in Section IV.

II. Methods

This section describes the methods employed in this

research paper.

A. Corpora

This section contains a brief description of the corpora used to evaluate the ML algorithms. Two corpora have been used to perform the experiments that were collected and formed from online text documents by [22]. These were BBC Arabic and CNN Arabic corpora. The BBC Arabic corpus collected from the BBC Arabic website (bbcarabic.com) that contains 4763 text documents,1.8 M words and 106K distinct keywords (when removing stop words). The corpus consisting of seven categories that are Middle East News, World News, Business & Economy, Sports, International Press, Science & Technology, and Art & Culture where each category contains the following document numbers: 2356, 1489, 296, 219, 49, 232, and 122, respectively.

On the other hand, the CNN Arabic corpus collected from the CNN Arabic website (cnnarabic.com) contains 5070 text documents, 2.2 M words and 144K distinct keywords (when removing stop words). The corpus consisting of six categories that are Business, Entertainments, Middle East News, Science & Technology, Sports, and World where each category contains the following documents numbers: 836, 474, 1462, 526, 762, and 1010, respectively.

B. Text Preprocessing Techniques

The text documents must go through a preprocessing phase. The preprocessing phase usually consists of the tasks of document conversion, tokenization, stop-word removal and stemming. The task of stemming is to remove all the affixes and suffixes from a word to extract its root. Since the Arabic language has different variations in representing text, three stemming techniques have been applied which are: Khoja stemming [23], light stemming [24] techniques and compared with Raw text (no stemming).

The next task is the feature extracting/selection. This phase, the influence of text preprocessing functions on text categorization is measured, especially the impact of using stemming from Arabic text categorization. In this task, a term weigh is representing each document as a weight vector; this regularly mentioned as the bag of words method. The goal of this work is to measure the influence of text preprocessing tasks on text categorization, especially the impact of using stemming with term weighing on Arabic text. Therefore, three stemming techniques with twelve-term weighing schemes have been applied (Table I). For instance, Term frequency (*tf)* measures the occurrence of term *t* in document *d* while document frequency (*df)* counts the number of documents that the term *t* presented at least once. On the other hand, the inverse document frequency (*idf*) measures how common the term in all the documents. The *idf* is going to be low if the term appears in many documents and high if the term appears in few documents.

TABLE. I. TERMS WEIGHING SCHEMES

Schema	Description
bool	boolean model (0 means absent while 1 means
idf	Inverse Document Frequency
tf	Term Frequency
tfidf	Term Frequency-Inverse Document
tfidf-norm-minFreq3	Apply normalization with minimum frequency
tfidf-norm-minFreq5	Apply normalization with minimum frequency
wc	output word counts
wc-minFreq3	Minimum frequency <3 for wc
wc-minFreq5	Minimum frequency <5 for wc
wc-norm	Apply normalization for wc
wc-norm-minFreq3	Apply normalization with minimum frequency
wc-norm-minFreq5	Apply normalization with minimum frequency

C. Machine Learning Algorithms

For this work, three Data Mining algorithms have been used that are C4.5, Naïve Bayesian and DMNBtext.

1) C4.5: C4.5 is an improvement of the ID3 decision tree based algorithm where the tree is built in a top-down approach. It applies the divide-and-conquer strategy to construct a decision tree by dividing the input space into local regions based on a distance metric. It uses information theory of choosing attributes to be selected in the root and internal nodes. The process of constructing a decision tree by C4.5 algorithm starts at the root node and is repeated until a leaf node is encountered. Additional detailed information on the C4.5 algorithm can be obtained in [25].

D. Naïve Bayesian

Naïve Bayesian is a statistical classification system based on the application of Bayeś theorem. This classification technique examines the relationship between an instance of each class and each feature assuming that all feature values are conditionally independent in a data set. It separately considers each feature and obtains a conditional probability for the associations linking the feature values and the class. The class with the highest probability value is selected as the predicted class. Additional detailed information on the Naïve Baysien algorithm can be obtained in [26].

E. DMNBtext

The learning process for the Bayesian network from data consists of two fundamental elements that are structure learning and parameter learning. When the Bayesian network has fixed structure, parameters learning can have two kinds of approaches that are discriminative and generative learning. The generative learning parameter appears to be more efficient while the discriminative parameter learning considers being more efficient. Hence, Discriminative parameter learning for Bayesian networks for text (DMNBtext)

consists of the benefits of discriminative learning and Bayesian network. Accordingly, it provides efficient, practical and straightforward discriminative parameter learning approach that discriminatively calculates frequencies from a dataset and after that uses the appropriate frequencies to estimate parameters. Additional detailed information on the DMNBtext algorithm can be obtained in [27].

F. Metrics for Evaluation

Accuracy is used as an evaluation criterion for text classification to measure the performance of the learning algorithms. The Accuracy, (1), reflects the total number of flows that are correctly classified from All- classes (the ones which the algorithm aims to classify):

$$Accuracy = \frac{TP + TN}{All} \quad (1)$$

The desired outcomes are to obtain a high percentage value for the Accuracy. In this paper, stratified 10 fold-cross validation is used to evaluate each of the learning algorithms on each datasets. To this end, Weka [28] is used with the default parameters for running the learning algorithms.

III. Experimental Results and Analysis

In this section to evaluate the different preprocessing methods and to show the effectiveness of data mining algorithms in finding sound signatures, we use two different datasets (BBC and CNN) to provide some measure of choosing proper preprocessing methods with classifier generalization (robustness).

Results presented in Tables II and III illustrate that the DMNBtext-based classification approach is observed to provide outstanding performances on both datasets. C4.5 delivers better performance on the BBC datasets while achieved lower performance on the CNN dataset compared with the other two data mining algorithms. Naïve Bayesian algorithms performance is the second best classifier comparing with the different two learning algorithms on both datasets.

On term of the both the stemming and preprocessing methods, the DMNBtext learning algorithms are achieving higher performance when Khoja and Light stems with boolean, idf, tfidf, wc and wc-norm preprocessing methods on the BBC data set. The DMNBtext attained an accuracy of 99% on the BBC data set. On the other hand, evaluating the DMNBtext on the CNN data sets, it produced higher performance using Light stem and Raw Text with boolean, idf, tf, tfidf and wc-norm preprocessing methods. The DMNBtext achieved an accuracy higher than 93% on the CNN data set. Moreover, C4.5 produced higher performance using the Light stem and Raw Text.

One of the significant concerns with Arabic Text categorization is the lack of standardized public Arabic corpora. Furthermore, most of the text data is obtained from online websites or newspapers. Hence, the performance of the machine learning models is biased to such corpora and would be difficult to generalize the models to all Arabic text.

TABLE. II. Accuracy Performance for each Classifier on the BBC Data using Stratified 10-Fold Cross-Validation

Preprocessing Methods	DMNBtext			C4.5			Naïve Bayes		
	Khoja stem	Light stem	Raw Text	Khoja stem	Light stem	Raw Text	Khoja stem	Light stem	Raw Text
bool	99.0	98.9	98.7	99.3	99.5	99.5	92.3	91.0	91.0
idf	99.0	99.0	98.8	99.4	99.5	99.5	75.0	77.2	78.6
tf	99.0	98.9	98.7	99.4	99.5	99.5	78.0	80.3	81.5
tfidf	99.0	99.0	98.8	99.4	99.5	99.5	78.0	80.2	81.5
tfidf-norm-minFreq3	98.5	98.4	98.3	99.4	99.5	99.3	83.5	77.6	71.8
tfidf-norm-minFreq5	98.5	98.2	98.3	99.3	99.5	99.4	85.8	80.3	76.0
wc	99.0	98.9	98.7	99.4	99.5	99.5	75.0	77.1	78.4
wc-minFreq3	98.4	98.1	98.2	99.5	99.4	99.5	74.8	77.4	78.9
wc-minFreq5	98.5	98.2	98.3	99.3	99.4	99.5	75.7	78.8	79.9
wc-norm	99.0	98.9	98.7	99.4	99.5	99.5	83.1	78.0	72.3
wc-norm-minFreq3	98.4	98.1	98.2	99.4	99.5	99.5	84.9	79.9	74.6
wc-norm-minFreq5	98.5	98.2	98.3	99.3	99.5	99.5	86.8	82.1	79.1

TABLE. III. ACCURACY PERFORMANCE FOR EACH CLASSIFIER ON THE CNN DATA USING STRATIFIED 10-FOLD CROSS-VALIDATION

Preprocessing Methods	DMNBtext			C4.5			Naïve Bayes		
	Khoja stem	Light stem	Raw Text	Khoja stem	Light stem	Raw Text	Khoja stem	Light stem	Raw Text
bool	93.0	93.5	93.6	75.6	77.2	78.5	86.3	87.5	89.1
idf	93.0	93.5	93.6	76.3	78.5	78.5	88.4	89.4	89.9
tf	93.0	93.5	93.6	76.3	78.3	78.5	88.1	87.8	88.1
tfidf	93.0	93.5	93.6	76.3	78.3	78.5	88.2	88.0	88.3
tfidf-norm-minFreq3	92.9	93.4	93.5	76.5	78.5	78.3	89.3	88.7	88.4
tfidf-norm-minFreq5	92.4	93.1	93.5	77.2	79.4	79.0	89.3	89.7	89.6
wc	93.0	93.5	93.5	76.4	78.5	78.5	88.5	89.5	90.0
wc-minFreq3	92.9	93.4	93.5	76.0	79.1	78.3	88.2	89.5	90.1
wc-minFreq5	92.4	93.1	93.5	75.8	80.7	79.0	87.4	89.5	89.5
wc-norm	93.0	93.5	93.5	76.7	79.6	79.2	88.6	86.5	85.7
wc-norm-minFreq3	92.9	93.4	93.5	76.5	79.1	78.9	89.0	89.0	89.0
wc-norm-minFreq5	92.4	93.1	93.5	77.1	78.9	79.3	89.0	89.5	90.0

IV. CONCLUSION

Arabic Text categorization is considered one of the difficult problems in classification using machine learning algorithms. Achieving high accuracy in Arabic text categorization depends on the preprocessing techniques used to prepare the data set. Thus, in this paper, an investigation of the impact of the preprocessing techniques in relation to the performance of three machine learning algorithms, namely Naïve Bayesian, DMNBtext and C4.5 is conducted.

The preprocessing step is an essential element in text categorization. Many preprocessing techniques that can be applied but it is very complicated to find the most suitable one. In this research paper, several preprocessing techniques were evaluated on Arabic text corpora using Machine learning algorithms. The DMNBtext learning algorithm is showing superior performance compared to other machine learning algorithms in categorizing Arabic text with different preprocessing techniques.

In the future, the plan is to apply genetic programming and deep learning techniques to improve the performance of the models and to use more extensive datasets.

REFERENCES

[1] D. D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," in *Third annual symposium on document analysis and information retrieval*, vol. 33, 1994, pp. 81–93.

[2] C. Apte, F. Damerau, S. Weiss *et al.*, *Text mining with decision rules and decision trees*. IBM Thomas J. Watson Research Division, 1998.

[3] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving arabic text categorization using decision trees," in *Networked Digital Technologies, 2009. NDT'09. First International Con- ference on*. IEEE, 2009, pp. 110–115.

[4] R. Al-Shalabi and R. Obeidat, "Improving knn arabic text classification with n-grams based document indexing," in *Pro- ceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt*, 2008, pp. 108–112.

[5] I. Hmeidi, B. Hawashin, and E. El-Qawasmeh, "Performance of knn and svm classifiers on full word arabic articles," *Advanced Engineering Informatics*, vol. 22, no. 1, pp. 106–111, 2008.

[6] B. Masand, G. Linoff, and D. Waltz, "Classifying news stories using memory based reasoning," in *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1992, pp. 59–65.

[7] W. Lam and C. Y. Ho, "Using a generalized instance set for automatic text categorization," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 81–89.

[8] Y. Yang, "Feature selection in statistical learning for text categorization," in *14th International Conference on Machine Learning, 1997*, 1997.

[9] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[10] N. Fuhr, S. Hartmann, G. Lustig, M. Schwantner, K. Tzeras, and G. Knorz, "Air/x-a rule based multistage indexing system for large subject fields," in *RIAO*, vol. 91, 1991, pp. 606–623.

[11] Y. Yang and C. G. Chute, "An example-based mapping method for text categorization and retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 12, no. 3, pp. 252–277, 1994.

[12] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. Khorsheed, and A. Al-Rajeh, "Automatic arabic text classification," 2008.

[13] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*. ACM, 1998, pp. 148–155.

[14] R. M. Duwairi, "Arabic text categorization." *Int. Arab J. Inf. Technol.*, vol. 4, no. 2, pp. 125–132, 2007.

[15] M. El Kourdi, A. Bensaid, and T.-e. Rachidi, "Automatic arabic document categorization based on the naïve bayes algorithm," in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*. Association for Computa- tional Linguistics, 2004, pp. 51–58.

[16] A. Moh'd A Mesleh, "Chi square feature extraction based svms arabic language text categorization system," *Journal of Computer Science*, vol. 3, no. 6, pp. 430–435, 2007.

[17] M. M. Syiam, Z. T. Fayed, and M. B. Habib, "An intelligent system for arabic text categorization," *International Journal of Intelligent Computing and Information Sciences*, vol. 6, no. 1, pp. 1–19, 2006.

[18] A. M. El-Halees, "Arabic text classification using maximum entropy," *IUG Journal of Natural Studies*, vol. 15, no. 1, 2015.

[19] L. Khreisat, "Arabic text classification using n-gram frequency statistics a comparative study." *DMIN*, vol. 2006, pp. 78–82, 2006.

[20] S. Al-Fedaghi and F. Al-Anzi, "A new algorithm to generate arabic root-pattern forms," in *proceedings of the 11th national Computer*

Conference and Exhibition, 1989, pp. 391–400.

[21] F. Harrag and E. El-Qawasmah, "Neural network for arabic text classification," in *Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the*. IEEE, 2009, pp. 778–783.

[22] M. K. Saad and W. Ashour, "Arabic text classification using decision trees," in *Proceedings of the 12th international work- shop on computer science and information technologies CSIT*, vol. 2, 2010, pp. 75–79.

[23] S. Khoja and R. Garside, "Stemming arabic text," *Lancaster, UK, Computing Department, Lancaster University*, 1999.

[24] L. Larkey, L. Ballesteros, and M. Connell, "Light stemming for arabic information retrieval," *Arabic computational mor- phology*, pp. 221–243, 2007.

[25] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[26] G. H. John and P. Langley, "Estimating continuous distribu- tions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[27] J. Su, H. Zhang, C. X. Ling, and S. Matwin, "Discriminative parameter learning for bayesian networks," in Proceedings of the 25th international conference on Machine learning. ACM, 2008, pp. 1016–1023.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10– 18, 2009.