International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 1, February 2019, pp. 652~659
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i1.pp652-659

The effect of training set size in authorship attribution: application on short Arabic texts

Mohammed AL-Sarem[1] , Abdel-Hamid Emara[2]
[1]Department of Information System, Taibah University, Madinah, KSA
[1]Department of Computer Science, Saba'a Region University, Mareb, Yemen
[2]Department of Computer Science, Taibah University, Madinah, KSA
[2]Computers and Systems Engineering Department, Al-Azhar University, Egypt

Article Info

Article history:

Received Apr 14, 2018
Revised Sep 3, 2018
Accepted Sep 26, 2018

Keywords:

Arabic language
Authorship attribution training set size
Linear regression
Mahalonobis distance
MLP classifier

ABSTRACT

Authorship attribution (AA) is a subfield of linguistics analysis, aiming to identify the original author among a set of candidate authors. Several research papers were published and several methods and models were developed for many languages. However, the number of related works for Arabic is limited. Moreover, investigating the impact of short words length and training set size is not well addressed. To the best of our knowledge, no published works or researches, in this direction or even in other languages, are available. Therefore, we propose to investigate this effect, taking into account different stylomatric combination. The Mahalanobis distance (MD), Linear Regression (LR), and Multilayer Perceptron (MP) are selected as AA classifiers. During the experiment, the training dataset size is increased and the accuracy of the classifiers is recorded. The results are quite interesting and show different classifiers behaviors. Combining word-based stylomatric features with n-grams provides the best accuracy reached in average 93%.

Corresponding Author:

Mohammed Al-Sarem,
Department of Information System, Taibah University,
PO box 344, Medina, Kingdom of Saudi Arabia.
Email: mohsarem@gmail.com

1. INTRODUCTION

Unlike text categorization, authorship attribution (AA) is a subfield of linguistics analysis which aims to identify the original author among a set of candidate authors [1]. The idea can be created basically as follows: for given text sets of known authorship, the author of the unseen text is estimated by matching the anonymous text to one author of the candidate set. The estimated function depends on the extracted human "stylometric fingerprint" features. Despite author's writing style that can change from topic to topic, some

olds and the widely use by hundreds of millions of people, only a very small number of authorship attribution studies that employ machine-learning methods has been published for Arabic texts so [2]. Abbasi and Chen, e.g., applied support vector machine (SVM) and C4.5 decision trees on political and social Arabic web forum messages [3]. Despite the notable results they have, the dataset is quite large to extract enough features. Stamatatos [4] tested also the use of SVM on Arabic newspaper. Although the experiment was conducted to investigate the effect of class imbalance problem, the findings encouraged us to investigate which size and length effect the performance of AA classifiers has. In [2], the linear discriminant analysis (LDA) is used. For attributing text, function words were used as feature. At training and testing phase, the used texts were divided into two chunks: the first chunk with 1000-words, while the second chunk with 2000-words. The findings indicate that the longer the text is, the higher the obtained performance. The best performance they obtained with the

second chunk was around 87% accuracy. The same findings were reported in [5] in which the authors examined the performance of several classifiers, namely, SVM, MLP, Linear Regression, Stamatatos distance and Manhattan distance. The results confirmed the findings by Eder in 2013 for the English language [6] and the minimum size of textual data is 2500 words per documents.

Existing works on authorship attribution are conducted with long text size. Besides that, there are many problems facing researchers in this domain, especially in Arabic:

a. Researches in AA are quite few and are not tackled as much as in other languages;
b. Absence a benchmark data sets, full adaptable support tools lead to increase the exerted effort in extracting the necessary features;
c. Nature of Arabic language also adds extra steps to text preprocessing, and words in Arabic tend to be short, which might reduce the effectiveness of some stylistic features, such as word length distribution [3].

Like any supervised learning problem, solving the AA problems go through three key phases: (i) data preprocessing phase where the typical dataset collection, feature extraction, feature selection, and dimension reduction are necessary steps; (ii) training/testing phase: at this step, the classifiers are selected and the model is built; and (iii) pattern evaluation: to ensure the performance of the selected classifier, k-fold cross validation with different feature combination is conducted and accuracy of the models are computed.

Current paper tries to draw the researchers' attention to increase their efforts to enhance research in Arabic language domain. Since the AA problems in Arabic domain are many, current study focuses, on one hand, on investigating the effect of training set size with short text length. The reason behind such selection is to provide insights on the way the training set size impacts the accuracy performance. On the other hand, to help guide others in selecting suitable classifiers based on the size of the training set so that it will give the optimum result.

In the preparation for this research, there are, to the best of our knowledge, no published works or researches in this direction or even in other languages. Since there are numerous methods which have been developed to tackle the AA problem, the researcher limits his investigation to selected methods and writing style features; however, the idea is still valid to assess other classifiers with different writing style combination.

This paper is organized as follows: Section II gives a general overview of the authorship attribution problem, the selected writing styles, and the selected classifiers. Section III presents the conducted experiment and finding results and Section IV provides some comparison with other existing works in term of accuracy. Finally, Section V summarizes the whole paper.

2. AUTHORSHIP ATTRIBUTION

2.1 The selected writing style features

To identify the author of an unseen text, set of popular features are extracted in order to quantify the writing style. In the literature, several features can be found. The interested reader is referred to [1], [5], [7] and [8]. These features can be categorized into seven main groups: lexical, character, syntactic, semantic, content-specific, structural and language-specific. In this paper, the lexical features, namely word-based features have been selected. The features were combined with n-gram method and Part-of-Speech and organized different sets for testing as follows:

Set1: Word-based Lexical Features (WLF).
Set2: N-Gram.
Set3: Part-of-Speech (POS)
Set4: N-gram + POS
Set5: WLF+N-gram
Set6: WLF+POS
Set7: WLF+N-gram+ POS

2.2 The selected classifiers

Since there are various machine learning methods that can be used for AA and due to many classifiers can be selected, our selection was randomly and the following methods were used: Mahalanobis distance (MD), Linear Regression (LR), and Multilayer Perceptron (MP).

2.2.1. Mahalanobis distance

The *Mahalanobis* distance is a measure of distance between an observation $\vec{x} = (x_1, x_2, x_3, \dots, x_N)^T$, a set of its means $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \dots, \mu_N)^T$ and a distribution D (covariance matrix S) as follows:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})} \quad (1)$$

It can also be defined as a dissimilarity measure between two random vectors $\vec{x}$ and $\vec{y}$ of the same distribution with the covariance matrix S:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})} \quad (2)$$

In the cases the where the covariance matrix is the identity matrix, the Mahalanobis distance is reduced to the Euclidean distance. In addition to this, if the covariance matrix is diagonal, then the resulting distance measure is called a normalized Euclidean distance:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}} \quad (3)$$

where s_i is the standard deviation of the x_i and y_i over the sample set. The *Mahanalobis* distance is widely used in LDA, Clustering analysis, and classification techniques.

2.2.2. Linear regression

LR is one of the oldest methods that is widely used in predictive analysis. The main idea behind it is to minimize the sum of the squared errors to fit a straight line to a set of data points. The LR model assumes that the relationship between the dependent variable y_i and the p-vector of *regressors* x_i is linear. Formally, the model takes the form:

$$y = X\beta + \varepsilon \quad (4)$$

where, y is called the *regressand* or *response variable*, X are *regressors* or *predictor variables*, β are a "*estimated effects*" or *regression coefficients*, and ε is an error term.

2.2.3. Multilayer perceptron

The MLP is a classical neural network classifier in which the errors of the output are used to train the network [9]. When applying MLP classifier, we should take in consideration the high computational effort for training [10] and problem falling in a local minima which leads to some errors of classification. In nature, the MLP is a class of feed forward network. It consists of three layers of nodes with at least one or more hidden layer(s). To ensure training of the networks, several back- propagation techniques have been utilized. Figure 1 represents a MLP with a single hidden layer (interested reader can refer [11].

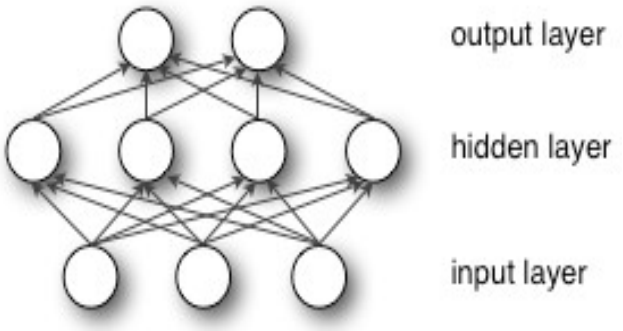

Figure 1. A MLP with single hidden layer

3. EXPERIMENT AND FINDING RESULTS

3.1. Dataset

Our dataset corpus consists of 4631 texts documents which were extracted from *Dar Al-ifta AL Misriyyah* website. The texts include Islamic fatwas from 1896 to 1996. The original texts were unstructured and required special tool to extract the fatwa structure from the whole documents. For this purpose, a C# tool was implemented. Since we seek to investigate the size and word length effects on AA classifiers, we grouped fatwas by the word length and size into different groups. Then, a set with the smallest test size is used later during the experiment. Table 3 summarizes the statistical features of the used dataset. The word size ranges widely. It varies from 11 words to 1500 words per texts.

3.2. Design of experiment

As mentioned previously, AA problem can be considered a supervised learning task which means there is a need to train the model, and then test it on a set of unseen texts. Despite the whole text mining process, we focus here only on the training/test phases and evaluate the performance of the classifiers by the number of right predicted texts. As said earlier, a combination of WLF, N-gram, and POS features are used and three types of classifiers are employed (i.e., *Mahalanobis* distance, MLP, and Linear Regression).

Figure 2 presents in addition to the typical solution of an AA problem, the way that we are followed during the experiment (the colored lines in red). Below, the general procedure of the experiment:

a. After executing the whole training and attributing phase, the finding results are recorded. Then, at the next iteration, we increase the training set size and the whole procedure is executed again and again. We limit the experiment with 4-fold iterations. However, there is a possibility to increase number of iterations based on the partitioned training sets.
b. The classifier type is still the same until we investigate all the partitioned training sets.
c. The all writing feature combinations are examined, and then the *accuracy* of the model is recorded.

To avoid a bias that may occur, the partitioned training sets were collected carefully and *t-tests* were conducted to ensure that the training sets are not significant differ. The training sets were organized as follows:

a. TS1: the training set with 330 documents
b. TS2: the training set with 378 documents
c. TS3: the training set with 414 documents
d. TS4: the training set with 438 documents

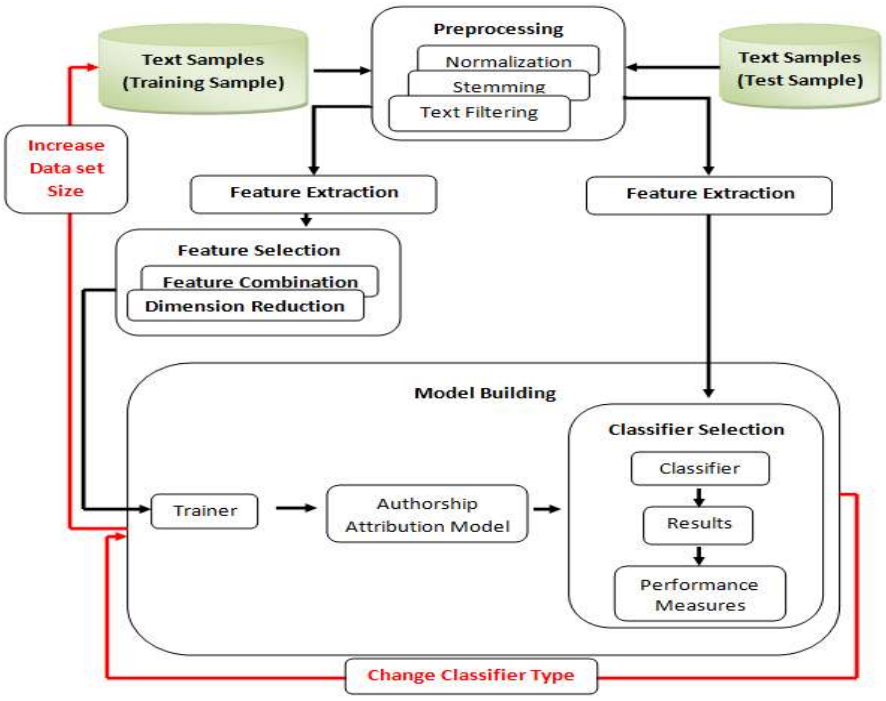

Figure 2. Design of experiment

3.3. Experimental findings

3.3.1. Authorship attribution using WLS features

We recorded the accuracy performance after employing the classifiers with different training sets as shown in Table 1. As we can see, the classifiers performed differently within the training sets. In almost data sets, the MLP overcomes others classifiers when the training sets have been increased.

Table 1. Performance of classifiers with WLF features

Classifier Type	TS4	TS3	TS2	TS1
MLP	77.08%	91.67%	79.167%	75%
LR	72.916%	91.67%	70.83%	83.33%
MD	64.58%	66.67%	37.5%	58.33%

Interesting finding is notated that accuracy of the MLP classifier increased with increasing the training set size, then decreased vastly. The same thing can be said about LR and MD methods with some nuance change. The LR performed better than MD. In addition, it gives also better results where the training set is quite few.

3.3.2. Authorship attribution using N-gram features

To investigate impact of the training sets size on AA using word *N-gram*. Indeed, the value of the gram can be selected randomly. However, to capture more words, we adjusted the *n-gram* to be five. Table 2 shows the performance of classifiers regarding each training set. Table 3 summarizes the statistical features of the used dataset. The word size ranges widely. It varies from 11 words to 1500 words per texts.

Table 2. Performance of classifiers with N-gram

Classifier Type	TS4	TS3	TS2	TS1
MLP	89.58%	91.67%	91.67%	100%
LR	89.58%	91.67%	87.5%	100%
MD	95.83%	94.44%	95.83%	100%

Table 3. Statistical features of dataset

Statistical Features	Textual Data Feature				
	# words	# sentences	Avg. words/sent.	# paragraphs	Avg. words length
Mean	125.26	4.51	28.04	12.4975	4.514755
St. Dev.	147.73	3.2850	22.115	12.5138	0.2897
Variance	21824.9	10.79	489.079	156.596	0.08395
Max.	1844	31	195.5	160	5.585
Min.	11	2	4.75	160	3.761

Comparing the classifiers' performance with n-gram with their performance with WLF, we note an improvement in AA accuracy. However, with increase the training set size the performance deteriorated. Furthermore, the MD method overcomes other classifiers even with increasing the training set size as shown in Figure 3.

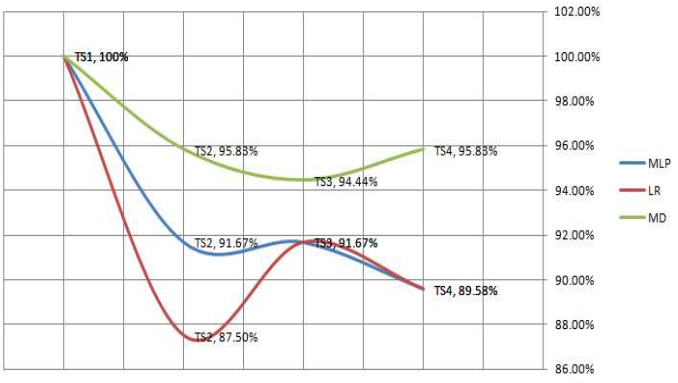

Figure 3. Classifier Performance with N-gram feature

3.3.3. Authorship attribution using POS features

With POS method, the performance of the classifiers are different. Both the accuracy of MLP and LR methods increases with increasing the training set size until a certain point, then decreases again with increasing the size. The MLP classifier decreases significantly comparing with LR. On the opposite, the MD method shows a different behavior. The accuracy of MD classifier is still better than other classifiers as shown in Figure 4.

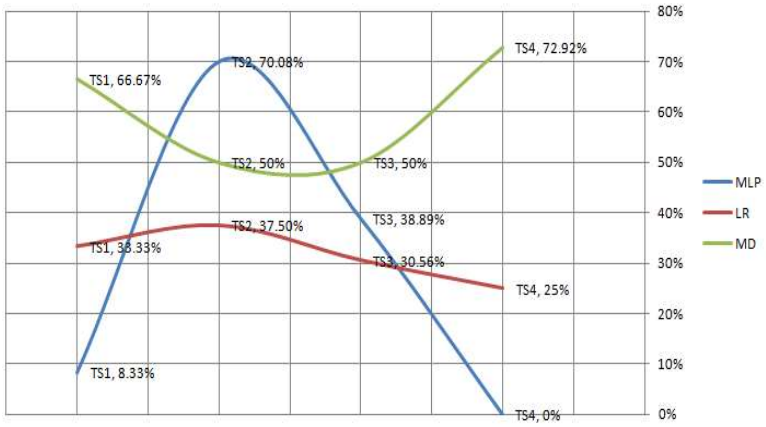

Figure 4. Classifier performance with POS

3.3.4. Authorship attribution with all features

In this part, we investigate performance of the classifiers with different combination. Table 4 summarizes all the finding results with different combinations. In most cases, the LR overcomes the other classifiers. The MD provides low performance accuracy, especially when the WLF features combined with POS. In general, combining POS with N-grams enhances the performance of attribution task. However, the accuracy is still lower than the WLF and n-gram combination. It is also notable that the classifier's performance affected negatively with increasing the training set size. In contrary, applying the full combination of WLF, N-grams, and POS lead to decreasing the accuracy of AA.

Table 4. Performance of classifiers with different combination

Methods		Training Sets (TS)			
		TS4	TS3	TS2	TS1
WLF+N-gram	MLP	83.33%	91.67%	83.33%	69.44%
	LR	66.67%	91.67%	83.33%	62.5%
	MD	50%	61.11%	60.42%	41.67%
WLF+ POS	MLP	79.167%	83.33%	66.67%	58.33%
	LR	79.167%	86.11%	70.83%	48.61%
	MD	37.5%	47.22%	41.67%	34.22%
All	MLP	87.5%	69.44%	83.33%	66.67%
	LR	62.5%	94.44%	87.5%	62.5%
	MD	41.67%	61.11%	56.25%	30.56%
n-gram + POS	MLP	91.67%	91.67%	81.25%	79.17%
	LR	87.5%	91.67%	79.17%	86.11%
	MD	75%	77.78%	62.5%	62.5%

4. COMPARISON WITH OTHER WORKS

As conclusion of the previous sections, we found that combining word-based stylomatric features with n-grams provides the best accuracy reached in average 93%. Thus, current section compares performances of those classifiers that are used in the reviewed works with our findings in term of accuracy. Even if they obtained come from different datasets, the comparison gives an indication of the performance of the different methods. Table 5 presents the comparison regardless the text size. It presents accuracy of the used classifiers of those works mentioned throughout this paper regardless the text size, whilst Table 6 present the accuracy taking in consideration the same text size.

Table 5. Comparison of the best reached accuracy in this work with accuracy of other methods regardless the text size

Reference	Accuracy	Data
Our Work	100%	Islamic Fatwas collected from *Dar Al-ifta AL Misriyyah* website
Shaker and Corne [2]	87.63%	Arabic books obtained from the website of the Arab Writers Union
Abbasi and Chen [3]	85.4%	Arabic web forum messages from Yahoo groups
Stamatatos [4]	93.4%	Arabic newspaper report of *Alhayat*
Ouamour et al. [5]	100%	Ancient historical books in Arabic
Altheneyan and El Bachir Menai [12]	97,84%	Arabic books collected from *Alwaraq* website

Table 6. Comparison of the best reached accuracy in this workwith accuracy of other methods regarding the text size

Reference	Accuracy	Data Size
Our Work	93%	[11-1500] words per document
Ouamour et al. [5]	80%	[100-1500] words per document
Al-Ayyoub et al. [8]	59.8%	Up to 140 words per tweets

5. CONCLUSIONS AND FUTURE WORK

This study has investigated the effect of increasing training set size on performance of authorship attribution classifiers. The experiment was designed to measure accuracy of classifiers with short Arabic texts. It was also designed to investigate the performance of AA classifier using different combination of stylomatric features. We limited our experiment to assess three well-known classifiers, namely linear regression, multilayer perceptron, and Mahanolobis distance. However, the research methodology is still valid with other classifiers.

The overall results show that the classifiers have different behaviors regarding the used AA features and training set size. The MLP classifier exceeds the other classifiers when the WLF features are used. The MLP, to a certain point, is positively affected by the increase in the training set size, then decreased. The n-gram methods lead to decreasing the classifiers' performance with increasing the training set size. However, generally speaking, the n-gram features provide the best results among the all stylomatric features. With POS features, the classifiers show different behaviors. The MLP classifier decreases significantly compared with LR. On the contrary, the MD shows a different behaviors and provides better accuracy than MLP and LR.

We also applied the classifiers with different features combination (WLF+ n-Gram, WLF+ POS, n-Gram+ POS, and a set of all features). The MD provides low accuracy especially when the WLF features combined with POS. Combining POS with N-grams improves the performance of attribution task. Contrary to the expected results, applying the full combination of WLF, N-grams, and POS leads to decreasing the accuracy of AA.

For future work, we intend to extend the experiments to investigate accuracy of other classifiers and to a larger training set. We also plan to investigate the impact of other feature selection methods on the performance of the AA problem.

REFERENCES

[1] A. Al-Falahi, M. Ramdani, M. Bellafkih, M. Al-Sarem, "Authorship Attribution in Arabic Poetry' Context Using Markov Chain classifier," *IEEE*, 2015.

[2] Shaker, K., Corne, D., "Authorship Attributionin Arabic Using A Hybrid Of Evolutionary Search And Linear Discriminant Analysis," In: *2010UK Workshopon Computational Intelligence (UKCI)*, pp.1–6. 2010, Doi:10.1109/UKCI.2010.5625580.

[3] Abbasi, A., Chen, H., "Applying Authorship Analysis To Arabic Webcontent," In: Kantor, P., Muresan, G., Roberts, F.,Zeng, D.D., Wang, F.-Y., Chen, H., Merkle, R.C. (Eds.), *Intelligence and Security Informatics*, vol. 3495. Springer-Verlag, Berlin, Heidelberg, pp. 183–197, 2005.

[4] Stamatatos, E., "Author Identification: Using Text Sampling To Handle The Class Imbalance Problem," *Inf. Process. Manage*. 44 (2), 790–799, 2008, http://dx.doi.org/10.1016/j.ipm.2007.05.012.

[5] Ouamour, S., Khennouf, S., Bourib, S., Hadjadj, H., & Sayoud, H. "Effect of the Text Size on Stylometry-Application on Arabic Religious Texts". In *Advanced Computational Methods for Knowledge Engineering* pp. 215-228, Springer, Cham., 2016.

[6] Eder, M., "Does Size Matter? Authorship Attribution, Small Samples, Big Problem," *Lit. Ling. Comput*. 2013. doi:10.1093/llc/fqt066.

[7] Sara El Manar El Bouanani, Ismail Kassou, "Authorship Analysis Studies: A survey," *International Journal of Computer Applications* (0975-8887) Volume 86-No 12, January 2014.

[8] M. Al-Ayyoub, Y. Jararweh, A., Rababa'ah, and M., Aldwairi. "Feature Extraction and Selection for Arabic Tweets Authorship Authentication," *J. Ambient Intell Human Comput*. 8:383-393, 2017, DOI 10.1007/s12652-017-0452-1.

[9] Sayoud, H., "Automatic Speaker Recognition-Connexionnist Approach," PhD thesis, USTHB University, Algiers, 2003.
[10] F.J Tweedie, S. Singh, and D.I. Holmes, "Neural Network Application in Stylometry: The Federalist Papers," *Computers and the Humanities*, vol. 30, pp. 1-10, 1996.
[11] Pal, S. K., & Mitra, S. "Multilayer Perceptron, Fuzzy Sets, and Classification," *IEEE Transactions on neural networks*, 3(5), 683-697, 1992.
[12] Altheneyan AS, "Menai MEB Naïve Bayes Classifiers for Authorship Attribution of Arabic Texts," *J King Saud Univ Comput Inf Sci*, 26 (4):473–484, 2014.

BIOGRAPHIES OF AUTHORS

Mohammed Al-Sarem is an assistant professor of information system at the Taibah University, Al Madinah Al Monawarah, KSA. He received the PhD in Informatics from Hassan II University, Mohammadia, Morocco in 2014. His research interests center on E-learning, educational data mining, Arabic text mining, and intelligent and adaptive systems. He published several research papers and participated in several local/international conferences

Abdel Hamid Emara received his BSc, MSc, and PhD in Systems and computers engineering from Al-Azhar university in 1992, 2000, 2006, respectively. He works as a lecturer in Systems and Computers Engineering Department at Al-Azhar University in Cairo, Egypt. He has experience of 12 years which includes both academic and research. He is currently an Assistant Professor in Computer Science Department at University of Taibah at Al- Madinah Al Monawarah, KSA. He is research interest includes Knowledge Discovery and Data Mining, Data Security, Data analysis and Artificial intelligence. He has published quality International journal papers.