A Comparative Survey of Authorship Attribution on Short Arabic Texts

Siham Ouamour and Halim Sayoud(✉)

University of Science and Technology Houari Boumediene, Algiers, Algeria
siham.ouamour@uni.de, halim.sayoud@gmail.com

Abstract. In this paper, we deal with the problem of authorship attribution (AA) on short Arabic texts. So, we make a survey on a set of several features and classifiers that are employed for the task of AA. This investigation uses characters, character bigrams, character trigrams, character tetragrams, words, word bigrams and rare words. The AA is ensured by 4 different measures, 3 classifiers (Multi-Layer Perceptron (MLP), Support Vector Machines (SVM) and Linear Regression (LR)) and a new proposed fusion called VBF (i.e. Vote Based Fusion). The evaluation is done on short Arabic texts extracted from the AAAT dataset (AA of Ancient Arabic Texts). Although the task of AA is known to be difficult on short texts, the different results have revealed interesting information on the performances of the features and classification techniques on Arabic text data. For instance, character-based features appear to be better than word-based features for short texts. Furthermore, the proposed VBF fusion provided high performances with an accuracy of 90% of good AA, which is higher than the score of the original classifier using only one feature. Globally, the results of this investigation shed light on the efficiency and pertinency of several features and classifiers in AA of short Arabic texts.

Keywords: Natural language processing · Artificial intelligence
Authorship attribution · Arabic language · Short texts · Text-mining

1 Introduction

As per definition, the task of author recognition can be divided into several fields:

- authorship attribution (AA) or identification: consists in identifying the author(*s*) of a set of different texts;
- authorship verification: consists in checking whether a piece of text is written or not by an author who claimed to be the writer;
- authorship discrimination: consists in checking if two different texts are written by a same author or not [1];
- plagiarism detection: in this research field we look for the sentences or paragraphs that are taken from another author [2];
- text indexing and segmentation: which consists in segmenting the global text into homogeneous segments (*each segment contains the contribution of only one author*) by giving the name of the appropriate author for each text segment [3].

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 479–489, 2018.
https://doi.org/10.1007/978-3-319-99579-3_50

Although several works are reported for the English and Greek [4] languages, the authors have not found a lot of serious research works made with Arabic texts. That is why; they propose an overall research work of AA that handles several texts written by 10 ancient Arabic travelers who wrote several books describing their travels. A special Arabic corpus has been built by the authors of this paper in order to assess several features and classifiers. The paper is organized as follows: In Sect. 2, we quote some previous works related to AA. In Sect. 3, we describe our textual corpus. Section 4 defines the different classifiers and distances used during the experiments. Results are presented in the Sect. 5 and an overall conclusion is given.

2 Related Works

Authorship attribution consists in identifying the author of a given text. Several works have tested different features during the last three decades. For instance, Holmes in 1994 [5], Stamatatos in 2000 [6] and Zheng in 2006 [7] proposed taxonomies of features to quantify the writing style. Mendenhall in 1887 [8] proposed sentence length counts and word length counts. A significant advantage of such features is that they can be applied to any language. Several researchers used lexical features to represent the author style. However other works used common words instead [9, 10]. Hence, various sets of words have been used for English, we can quote the works of Abbasi and Chen in 2005 [11]; the works of Argamon in 2003 [12]; the works of Zhao and Zobel in 2005 [13]; and the works of Koppel and Schler in 2003 [14]. Similarly, in the works of Argamon in 2007 [15], A new interesting feature was proposed by [16] and [17], namely: the *word n-grams*, which provided very good performances. Concerning the character *n*-grams, the application of this approach to AA has shown an interesting success. Character bigrams and trigrams have been used in the works of Kjell [18]. In the works of Forsyth and Holmes [19], one found that bigrams and character *n*-grams of variable-length performed better than lexical features. They have been successfully used in the works of Peng [20], Keselj [21] and Stamatatos [22]. On the other hand, it is not only the feature which is important; in fact, the choice of a suitable classifier is important too. That is, in 2010, Jockers and Witten [23] tested five different classifiers. Concerning the Arabic language, there are not a lot of works that are reported. However, we can cite some recent works such as those reported by Sayoud 2012 [1] and Shaker [24]. Sayoud conducted an investigation on authorship discrimination between two old Arabic religious books: the Quran (*The holy words of God*) and Hadith (*statements of the prophet Muhammad*) [1]. Shaker investigated the AA problem in Arabic, using Function Words [24]. In this investigation, we are interested in using several features and classifiers for an evaluation in Arabic stylometry. The AAAT dataset is built by the authors of this paper for a purpose of AA.

3 Description of the Text Dataset

Our textual corpus is composed of 10 groups of old Arabic texts extracted from 10 different Arabic books. The books are written by ten different authors and each group

contains different texts belonging to a unique author. This set of texts has been collected in 2011 from "Alwaraq library" (www.alwaraq.net); we called it AAAT. Furthermore, this corpus represents a reference dataset for AA in Arabic, which has been used by several researchers working in this field.

The texts of the corpus are quite short: the average text length is about 550 words and some texts have less than 300 words.

4 Classification Methods

For the evaluation task, we have evaluated 4 distances (Manhattan, Cosine, Stamataos, and Canberra distances) and 3 classifiers (SVM, MLP and LR).

Several features are also used, namely: characters, character n-grams, words, word n-grams and rare words in order to find the most reliable characteristic for the Arabic language. Furthermore, a Vote Based Fusion (*VBF*) has been proposed to enhance the overall classification performances.

4.1 Manhattan Distance (Man)

The Manhattan distance between two vectors X and Y of length n is defined as follows:

$$\mathrm{Man}(\mathrm{X},\mathrm{Y}) = \sum_{\mathrm{i}=1}^{n} \left|\mathrm{X_i} - \mathrm{Y_i}\right| \tag{1}$$

4.2 Cosine Distance

Cosine similarity is a measure of similarity between two vectors X and Y (of length n) that measures the cosine of the angle between them (*denoted by* θ).

The cosine distance, $cos(\theta)$, is represented using a dot product and magnitude as:

$$\cos\theta = \frac{X.Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i * Y_i}{\sqrt{\sum_{i=1}^{n} \left(X_i\right)^2} * \sqrt{\sum_{i=1}^{n} \left(Y_i\right)^2}} \tag{2}$$

4.3 Stamatatos Distance (Sta)

This distance was introduced by Stamatatos [25] to measure texts similarity. It was successfully employed in AA. It is given by the following formula:

$$Sta(\mathrm{X},\mathrm{Y}) = \sum_{i=1}^{n} \left[2\left(\mathrm{X_i} - \mathrm{Y_i}\right)/\left(\mathrm{X_i} + \mathrm{Y_i}\right)\right]^2 \tag{3}$$

4.4 Canberra Distance (Can)

The Canberra distance between vectors X and Y is given by the following equation:

$$\mathrm{Can}(X,Y) = \sum_{i=1}^{n} \left| \frac{(X_i - Y_i)}{X_i + Y_i} \right| \tag{4}$$

4.5 Sequential Minimal Optimization-Based Support Vector Machines (SVM)

In machine learning, SVM are supervised learning models with associated learning algorithms that analyze data and recognize patterns. They are used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier.

Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other.

A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVM can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

The SVM is a very accurate classifier that uses bad examples to form the boundaries of the different classes. Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming problem that arises during the training of the SVM. The SMO algorithm is used to speed up the training of the SVM. In our application, we solved the multi-class problems by using pairwise classification technique.

4.6 Multi-layer Perceptron (MLP)

The MLP is a feed-forward neural network classifier that uses the errors of the output to train the neural network: it is the "training step". The MLP is organized in layers: one input layer of distribution points, one or more hidden layers of artificial neurons (nodes) and one output layer of artificial neurons. Each node, in a layer, is connected to all other nodes in the next layer and each connection has a weight (which can be zero). The MLP is considered as universal approximator and is widely used in supervised machine learning classification. The MLP can use different back-propagation schemes to ensure the classifier training.

4.7 Linear Regression

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norms

(*as with least absolute deviations regression*), or by minimizing a penalized version of the least squares loss function as in ridge regression. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models.

Usually, the predictor variable is denoted by the variable X and the criterion variable is denoted by the variable y. Most commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X. Less commonly, linear regression could refer to a model in which the median of the conditional distribution of y given X is expressed as a linear function of X.

Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

4.8 Classification Process

The general classification process is divided into two methods: Training Model based Classification and Nearest Neighbor based Classification. In the first type, a training step is required to build the model or the centroid (*in case of similarity measures*); afterward, the testing step could be performed by using the resulting model. In the second type, the training is not required, since a simple similarity distance is computed between the unknown document and each referential text: the smallest distance gives an indication on the most probable class. Furthermore two types of measures are employed: a simple distance and a centroid based distance. The first type is known to be inaccurate, while the second one (*i.e. centroid*) is more accurate and robust against noises. The first classification type includes the following classifiers: Centroid based Similarity measures, Multi-Layer Perceptron, SMO-based Support Vector Machines and Linear Regression; whiles the second classification type includes only the nearest neighbor similarity measures. After every identification test, a score of good AA is computed in order to get an estimation on the overall classification performances.

5 Experiments of Authorship Attribution

In this section, we present the different experiments of AA, which are conducted on the historical Arabic texts. Several features are tested such as: characters, character bigrams, character trigrams, character tetragrams, words, word bigrams, word trigrams, word tetragrams and rare words. On the other hand, different types of classifiers (MLP, SVM and LR) and distances are employed to ensure the AA classification.

The AA Score (AAS) is calculated by using the *RandAccuracy* formula, as follows:

$$AAS\ score\ =\ Rand\ Accuracy = \frac{number\ of\ texts\ that\ are\ well\ attributed}{total\ number\ of\ texts} \tag{5}$$

5.1 Comparative Performances

For a purpose of comparison, several figures are represented and commented on to make a comparative study of the different features and classifiers.

That is, Fig. 1 summarizes the overall best results given by each classifier. In this figure, we remark that the Manhattan centroid distance seems to be very accurate, with a score of 90%, followed by the classifiers MLP and SVM, with a score of 80%, after that, we retrieve the Manhattan nearest neighbor distance and the LR classifier, which provide a score of 70%. Finally, the remaining distances: Canberra, Cosine and Stamatatos distances, give the worst performances, score of 60%.

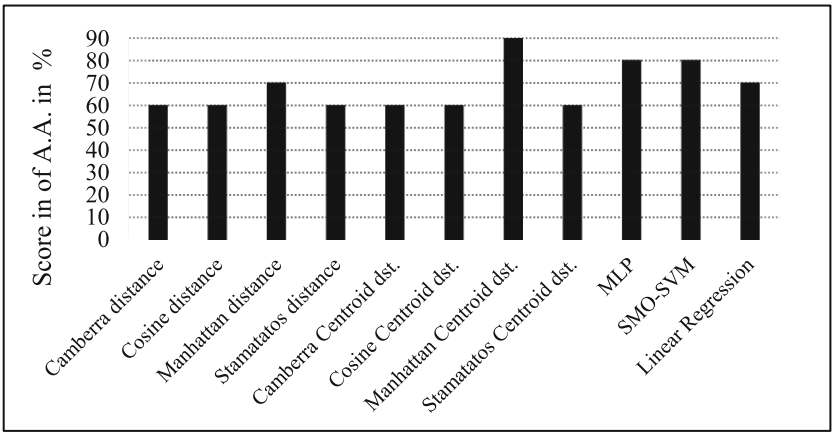

Fig. 1. Best scores of authorship attribution (*AAS*) given by the different classifiers.

In Fig. 2, we have presented the average AA performances for every feature. Those performances are obtained by calculating the mean of all the feature scores.

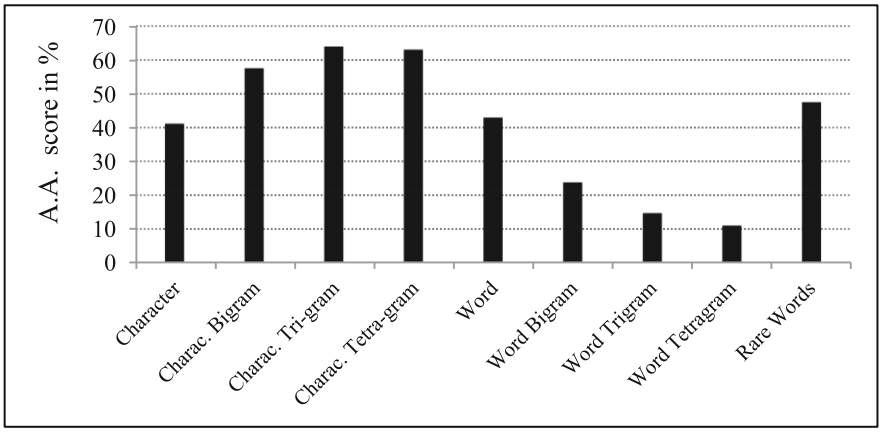

Fig. 2. Overall authorship attribution score for the different features used.

From Fig. 2, we can deduce that the best feature in these experiments is character trigrams, followed by character tetragrams, character bigrams and rare words. The performances of AA continue to decrease respectively by using words, characters, word bigrams, word trigrams and finally, word tetragrams, which represents the worst features in our experiments. In overall, we notice two important points: On one hand, the AAS increases with the character n-gram size (i.e. the size n) and decreases with the word n-gram size. On the other hand, character n-grams seems to be more accurate than word n-grams and rare words.

Similarly and in a dual form, Fig. 3 displays the average scores that are obtained by the different classifiers. These scores of performance are obtained by calculating the mean of all the scores of a specific classifier. So, we notice that the machine learning classifiers are the most accurate, especially the SMO-SVM (*average score exceeding 70%*), which provides high performances of AA. The MLP is strongly accurate with a score of about 70% of good attribution and the linear regression is quite interesting (*score over 60%*). On the other hand, we notice that the distances are less accurate in the overall, since the average attribution scores do not exceed 58.33%.

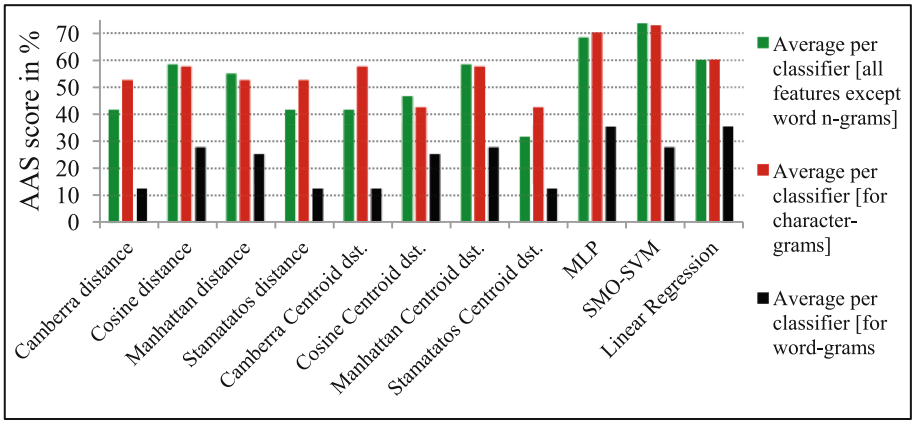

Fig. 3. Average AA score per classifier.

Once again, we can observe that character n-grams are better than word n-grams according to this same figure (Fig. 3) and we can also notice that the system presents a failure when using word n-grams. These last ones seem to be not suitable for short texts: this result is logical because short texts do not contain enough words or enough word n-grams either to make a fair statistical representation of the features.

Figure 4 presents the best score given by each feature. We see that a score of 90% is given by character tetragrams, followed by a score of 80% for character bigrams, character trigrams and rare words, thereafter, a score of 70% for words, 60% for characters, 50% for word bigrams, and a score of 20% for word trigrams and tetragrams.

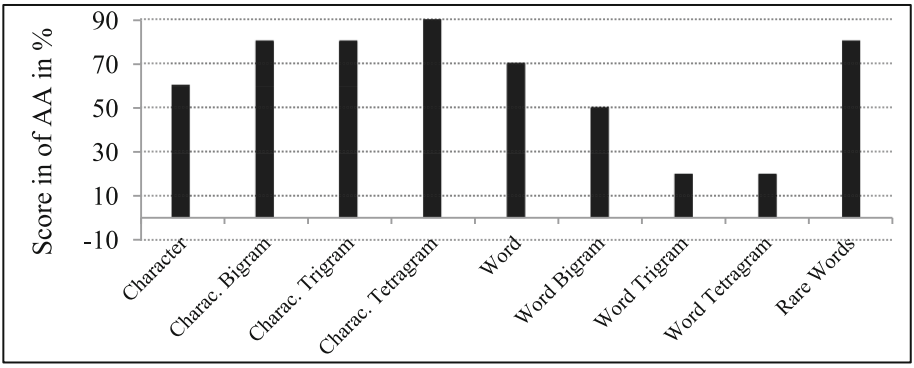

Fig. 4. Best score obtained with the different features.

5.2 Vote Based Fusion

In order to enhance the attribution performance, we thought to use several classifiers, which are combined in order to get a lower discrimination error: this combination is called Fusion. The fusion in the broad sense can be performed at different hierarchical levels or processing stages [26], as follows:

- Feature level, where the feature sets of different modalities are combined;
- Score (*matching*) level is the most common level where the fusion takes place. The scores of the classifiers are normalized and then combined in a consistent manner;
- Decision level where, the outputs of the classifiers establish the decision via techniques such as majority voting.

In this investigation, we have chosen to use the SMO-SVM classifier, which seems to be the best classifier in our experiments. The proposed fusion method is done at the decision level and is called "Vote-Based Fusion technique" or VBF. It consists in fusing the output decisions of the different systems (*i.e. each system uses the SVM classifier with one specific feature*) as it is described in Eq. 6. For the choice of the features, we have decided to keep only the most pertinent ones, namely those presenting a "best-score" of at least 80%. So according to Fig. 5, those pertinent features are: Character bigram; Character trigram and Rare words.

$$VBF_{Fusion} = Round\{(\alpha_1.Char2gram_{CLASS} + \alpha_2.Char3gram_{CLASS} + \alpha_3.RareWords_{CLASS})\frac{1}{\alpha_1 + \alpha_2 + \alpha_3}\}, \quad (6)$$

where CLASS represents the classifier output and α_i is a constant smaller than one.

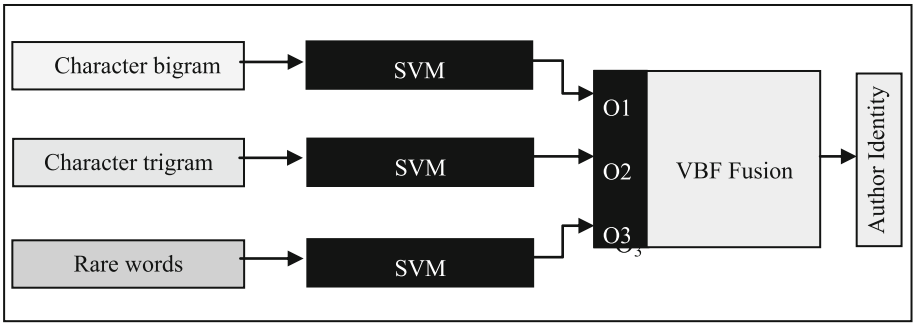

Fig. 5. Vote fusion technique. The outputs O_j are fused to produce the author identity.

The same previous experiments of AA have been conducted by using the proposed fusion technique. Results show that the fusion provides an accuracy of 90%, which is higher than all the scores provided by the SVM. This result is interesting since it shows that it is possible to enhance the identification accuracy only by combining several features and/or classifiers together. Furthermore, it is important to mention that an accuracy of 90% with short texts is motivating, since previous works showed that the minimum amount of required text for a fair AA is at least 2500 tokens [27].

6 Conclusion

An investigation of AA has been conducted on an old Arabic set of text documents that were written by ten ancient Arabic travelers. In this investigation, eleven different classifiers and distances have been used for the attribution task, by using nine different features. Moreover a fusion technique, called VBF, has been proposed to enhance the AA performances. The main conclusions of the different experiments can be summarized by the following points:

- Character bigram, trigram and tetragram appear to be interesting: Character tetragrams appear to be suitable for distances *(Manhattan, Canberra, Cosine and Stamatatos*), while for the machine learning, character bigram is the most accurate one.
- Manhattan centroid distance has shown excellent performances with an accuracy of 90% when using character tetra-grams. The performances of this distance are more or less comparable to those of the SVM, which is considered very reliable.
- As expected theoretically, the SVM has shown excellent average performances in most experiments, which recommends the use of this type of classifier in AA.
- Character-based features are better than word-based ones for short documents.
- The proposed VBF fusion provided high performances with an accuracy of 90% of good AA, which highly recommends the use of the fusion in AA.
- Although the word-based features did not give good results, rare words have presented good scores for almost all the classifiers. This result shows that some linguistic information of the author style are embedded in the rare words.

Finally, we think that the results of this investigation are interesting since they shed light on the efficiency of several features and classifiers in AA of short Arabic texts. As perspectives, one proposes to evaluate our system on dialectical Arabic language.

References

1. Sayoud, H.: Author discrimination between the Holy Quran and Prophet's statements. Lit. Linguist. Comput. **27**(4), 427–444 (2012)
2. Chowdhury, H.A., Bhattacharyya, D.K.: Plagiarism: taxonomy, tools and detection techniques. In: Paper of the 19th National Convention on Knowledge, Library and Information Networking (NACLIN 2016) held at Tezpur University, Assam, India (2016)
3. Sayoud, H.: Segmental analysis based authorship discrimination between the Holy Quran and Prophet's statements. Can. Soc. Digit. Hum., Digital Studies Journal (2015)
4. Tambouratzis, G., Hairetakis, G., Markantonatou, S., Carayannis, G.: Applying the SOM model to text classification according to register and stylistic content. Int. J. Neural Syst. **13**(1), 1–11 (2003)
5. Holmes, D.I.: Authorship attribution. Comput. Humanit. **28**, 87–106 (1994)
6. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Comput. Linguist. **26**(4), 471–495 (2000)
7. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: writing style features and classification techniques. J. Am. Soc. Inf. Sci. Technol. **57**(3), 378–393 (2006)
8. Mendenhall, T.C.: The characteristic curves of composition. Science **9**, 237–249 (1887)
9. Argamon S., Levitan, S.: Measuring the usefulness of function words for authorship attribution. In: Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (2005)
10. Burrows, J.F.: Word patterns and story shapes: the statistical analysis of narrative style. Lit. Linguist. Comput. **2**, 61–70 (1987)
11. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. Intell. Syst. **20**(5), 67–75 (2005)
12. Argamon, S., Saric, M., Stein, S.: Style mining of electronic messages for multiple authorship discrimination: first results. In: Proceedings of 9th ACM SIGKDD, pp. 475–480 (2003)
13. Zhao, Y., and Zobel, J.: Effective and scalable authorship attribution using function words. 2nd Asia Information Retrieval Symposium (2005)
14. Koppel, M., Schler J.: Exploiting stylistic idiosyncrasies for authorship attribution. In: IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis, pp. 69–72 (2003)
15. Argamon, S., et al.: Stylistic text classification using functional lexical features. J. Am. Soc. Inform. Sci. Technol. **58**(6), 802–822 (2007)
16. Peng, F., Shuurmans, D., Wang, S.: Augmenting naive Bayes classifiers with statistical language models. Inf. Retrieval J. **7**(1), 317–345 (2004)
17. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, Markov chains and author unmasking: an investigation. In: Proceedings of the International Conference on Empirical Methods in Natural Language Engineering, pp. 482–491 (2006)
18. Kjell, B.: Discrimination of authorship using visualization. Inf. Process. Manag. **30**(1), 141–150 (1994)
19. Forsyth, R., Holmes, D.: Feature-finding for text classification. Lit. Linguist. Comput. **11**(4), 163–174 (1996)

20. Peng, F., Shuurmans, D., Keselj, V., Wang, S.: Language independent authorship attribution using character level language models. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 267–274 (2003)
21. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. Pacific Association for Computational Linguistics, pp. 255–264 (2003)
22. Ouamour, S., Sayoud, H.: Authorship attribution of ancient texts written by ten arabic travelers using character N-Grams. CITS-2013, Athens, Greece, CITS (2013)
23. Jockers, M.L., Witten, D.M.: A comparative study of machine learning methods for authorship attribution. Lit. Linguist. Comput. **25**(2), 215–223 (2010)
24. Shaker, K.: Investigating features and techniques for Arabic authorship attribution, PhD thesis Heriot-Watt University (2012)
25. Stamatatos, E.: Author identification using imbalanced and limited training texts, text-based Information Retrieval, pp. 237–241 (2007)
26. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. Trans. Circ. Syst. Video Technol. **14**(1), 4–20 (2004)
27. Ouamour, S., Khennouf, S., Bourib, S., Hadjadj, H., Sayoud H.: Effect of the text size on stylometry-application on arabic religious texts. In: International Conference on Computer Science Applied Mathematics and Applications, pp 215–228, Vienna, Austria (2016)