Modeling Of Arabic Language For Authorship Identification

Heba M . Khalil ,Ahmed Taha,Tarek . El-shishtawy

Abstract: With the vast volume of data processed in digital form today, the need for and capability of analysing and processing this data for forensic authorship authentication has increased. The focus of study has concentrated on English, Spanish, and German. Arabic language has received less attention from the academic community due to the difficulty and length of Arabic sentences. This article provides a set of stylometric features derived from the study of many articles' parts of expression, including adjectives ratio, sentence size, conjunctions, and others. This details is classified into two categories: statistical features and linguistic features. The AdaBoost and Bagging ensemble approaches have been proposed in this research to maximise predictive efficiency in Arabic articles by using multiple learning. The results indicate that the Bagging model achieves average accuracy of 91.5 %, while the AdaBoost model achieves the highest accuracy of 93.6 %.

Keywords: Forensic authorship authentication, Stylometric features, Ensemble methods.

1-Introduction

From the viewpoint of linguistic description, forensic authorship authentication aims to detect an anonymous article's primary author. The main point is that each writer has a distinct set of characteristics than those in his writing style. While the writing style may differ from subject to subject, certain uncontrolled continuous habits and writing styles are still successful over time. The chosen writing style fits one of the detected writers set to detect the author of unidentified text. To solve this problem, many techniques have been employed for different types of features. These methods were focused on extracting certain characteristics from language, such as length and frequency of words, to detect the writing style of the author. While statistical methods are useful for solving the problem of authorship authentication, when the length of the document is small, they fail. Many researches in forensic authorship authentication area depended on machine learning models [1] as, the authorship forensic authentication can be considered as problem of classification [2]. The main idea of the machine learning models to solve the forensic authorship authentication is classifying data samples in to detected classes. Many models of machine learning such as decision trees [3], naive Bayes (NB) [4], [5], [6], k-nearest neighbour (k-NN) [7], [8] and Support Vector Machine (SVM) [9], [10] are widely used in solving forensic authorship authentication problem. In forensic authorship authentication, a few of studies used ensemble models such as [11], [12], although it presented a high performance in machine learning to enhance the results. Ensemble techniques are learning models that create a series of classifiers and then identify new data points by obtaining their predictions with a (weighted) vote. A few studies of forensic authorship authentication concerned of Arabic language so, this paper intents to enhance the forensic au-thorship authentication in Arabic language. This paper contains five parts arranged as follows: section two presents numerous studies on the forensic authorship authentication issue that have been carried out in the last several years. The proposed method of forensic authorship authentication based on ensemble method is discussed in details in section three. The experimental results which carry out to evaluate the proposed method in section four. Finally, in section five the conclusion and the future work are discussed.

2- Background and Related Works

Although forensic authorship authentication may be regarded as a specific form of authorship study, ensemble methods are a common methodology in Machine learning where a group of classifiers with their results are aimed to best choices in some way [13]. In this section, we discussed a review of Arabic forensic authorship authentication, machine learning based forensic authorship authentication and different types of features.

Forensic Authorship Authentication

Forensic Authorship Identification is about identifying the author of an anonymous text document based on the writing style. Each author has different writing style detected by extracting different types of features extracted from his documents [7]. One of the following methods is used to extract features from the documents of authors, instance and profile based methods [1]. As shown in figure 1, instance method extract the features from each article and it helps to detect any variations of the style. Profile method as shown in figure 2, collects the documents of each author in one file and extracts the writing style. This approach helps recognize the most uncontrolled habits and characteristics represented in the author's writing style.

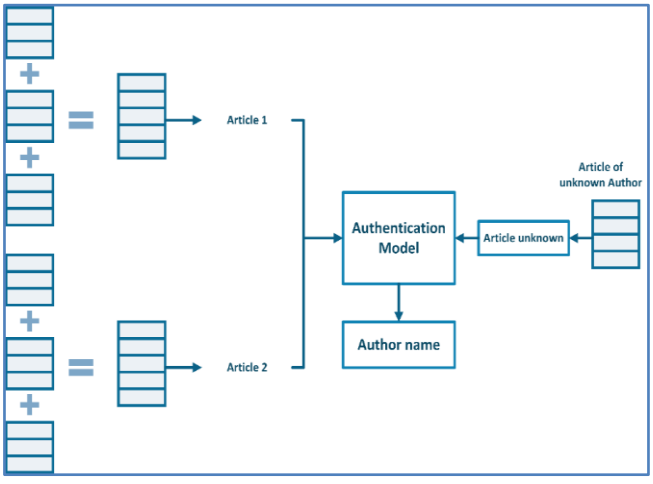


Figure.*1 the framework of instance based method.*

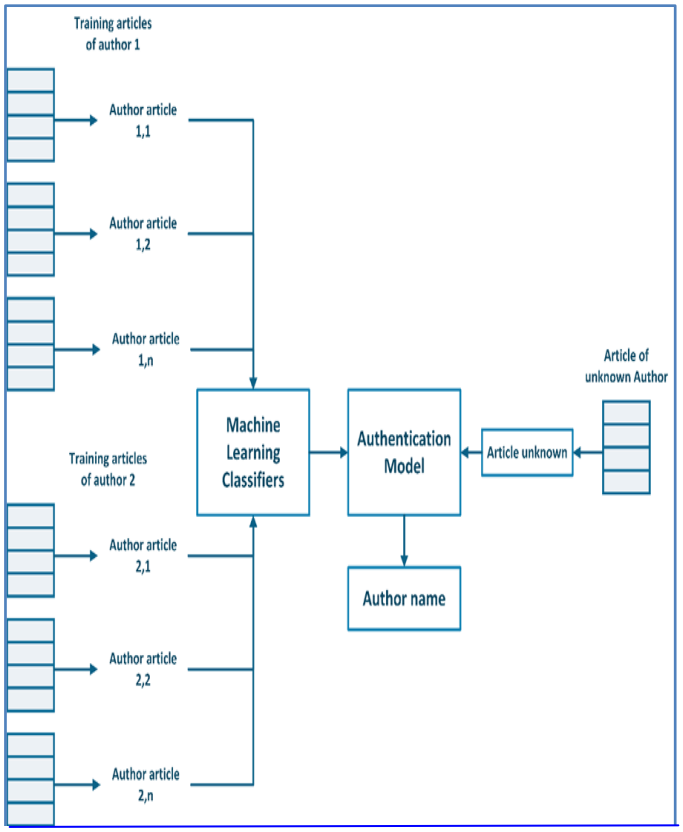


Figure.*2 the framework of the profile based method.*

Ensemble learning

Ensemble techniques are a kinds of research method that mixes a set of classifiers and then labels unique data points using their (weighted) vote prediction. By integrating the contribution of a series of classifiers, the main purpose of the ensemble approach is to enhance the efficiency of a classifier. It is obvious that the efficiency of the classifier varies and that there are more classifiers than others at any point. So, it's easier to work with a group of classifiers than to use each classifier alone. Boosting, bagging and random forests are the most common ensemble approaches. In order to prioritize the class name of each classifier and rate the class as optimum, the ensemble method relies on majority voting. Although a single classifier can make errors, if more than half the classifiers are in failure, the ensemble can only misidentify. Therefore, the performance of an ensemble is more powerful than a single classifier. [14].

Machine learning methods in forensic authorship authentication

There are many methods that solving the forensic authorship authentication problem based on machine learning. In [15] presented a method to solving authorship authentication problem for Arabic language. They used Arabic function words to extract the sets of features that used to authenticate the text of unknown author. They used set of machine learning classifiers such as a Linear Discriminant Analysis model and the Evolutionary Algorithm. The method achieved the accuracy in rang of 93%. The disadvantage of this method was that they depended on function words only to extract features. There were many types of features which obtained a better results.

Abooraig, R., et al. [16] collect and manually define a corpus of articles and comments written in Modern Standard Arabic, from various political contexts in the Arab world. They used the stylometric features approach to determine forensic authorship attribution. They used a combination of features with various types such as statistical features and syntactic features. They divided there features to two groups, features based on sentence and features based on words. They extracted four different versions from Arabic datasets F1, F2, F3 and F4. They trained the extracted features with a combination of classifiers and the best result was ranged from 72% to 87%. The Support Vector Machine (SVM) classifier achieved the best accuracies with selection features. This method depended on features that not sufficiently strong to detect unknown authors with low achieved results. Ouamour et al [17] proposed an approach that used the SVM classifier. They added using a Sequential Minimal Optimization model on SVM to enhance the training of SVM. This method depended on extract features such as vocabulary richness, characters, word extracted by n-gram and characters extracted by n-gram. This method used only two documents for each author to train it. They used a set of features and the better accuracy results was 80%. The disadvantage of this method was that it depended only on lexical features to detect the style of authors. Also they used a very small data and obtain a badly results. Fuchon et.al [18] introduced a new approach depended n gram model based on character level. They thought that using character level n gram help to detect important interword and interphrase features. The advantage of this approach was that it can applied in any language and did not have any parse sentence. The method depended mainly on character n gram level to extract the features from documents of authors. The classifier obtained the author of unknown documents of each language. This method evaluated three languages datasets and obtain the result ranging from 70% to 90%. Also in this method detecting the best value of n in n gram model still a main problem. Filiz et al [19] introduced a proposed method that depended on four types of machine learning models. Each model using the vector of features which extracted with bi-gram and tri-gram. They added some statistical features such as number of words in sentence, number of characters in words and vocabulary richness features. They depended on function words and part of speech tagging for extracting the detected features. This work collected many features to help the classifier to obtain the best results. They divided the features to four groups and used each group with multilayer perceptron, random forest, SVM and KNN classifiers. Each group of features achieves a different performance ranged from 60% to 84%. The disadvantage of this method was that it used n gram and can't detect the best value of n. Luyckx et.al [20] also depended on machine learning models to solve the author attribution problem. They used the type of SVM classifier called multiclass SVM. They used different types of features such as, characters feature that introduced by

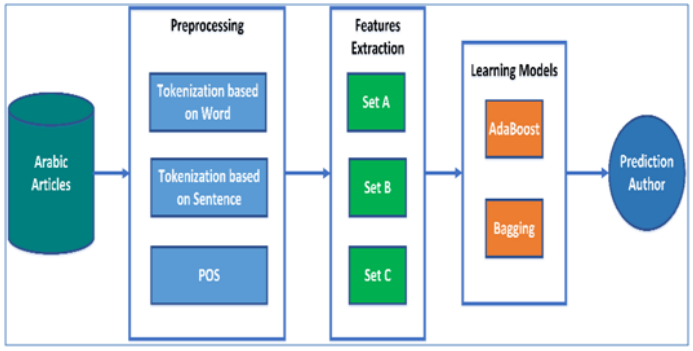
character n-grams level, lexical features that represented by n-grams word, and the syntactic feature that depended on using POST. The multiclass SVM able to using the both of large and small datasets very good. Jochim et al [21] presented machine learning approach for solving author attribution problem. They depended on a Support Vector Machine (SVM) classifier which was a type of machine learning models. They collected a collection of different types of features which were extracted from various articles to detect the author of an unknown article. The advantage of using SVM classifier was that it can easily get a very large numbers of features. This method achieved precision ranged from 60% to 80%.

3- PROPOSED METHOD

Throughout this section, we clarify the key components of the framework that we have implemented to test the ensemble process for Arabic forensic authorship authentication. The forensic authorship authentication method's main phases consist of pre-processing input texts, detecting the features and then training and authenticating. Fig. 3 illustrates this method.

The phase of Pre-processing

Each article in train and test data is preprocessed before extracting features from Arabic articles. The following preprocessing steps have been taken: With each article, the sentence identification and tokenization method is applied. To address differences in the representation of Arabic articles, normalization seems to be very necessary. The elimination of nonletters and stop words is maintained because, unlike text mining activities, they can provide authorial evidence. A Stanford POS tagger named StanfordCoreNLP to classify and evaluate each word's POS tagging.



Features extraction vectors

Articles are viewed as numeric vectors that match all the features obtained. We used some different types of features and divided them to three sets. A complete overview of the series of features are represented as the following:

Set A: statistical features

1.F_{smal}: this feature represents the frequency of words smaller than 3 characters normalized by dividing it by the total number of words in the given article. This feature leads to represent an article as a numeric vector in a way that its entries indicate the number of words with lengths smaller than 3. This feature is type of lexical features as shown in table 1.

2.F_{bet}: in this feature words length is used. It represents the total number of words between 3-5 characters normalized by dividing it by the all words in the given article. This feature is type of lexical features.

3.F_{long}: which represents the frequency of words longer than 5 characters normalized by dividing it by the total number of words in the given article. This feature is type of lexical features.

4.F_{dig}: which represents the frequency of digital words normalized by dividing it by the total number of words in the given article. This feature is type of lexical features.

5.F_{voc}: this feature is called vocabulary richness feature. It is calculated by detecting the total different words then dividing it by the total number of words in the given article.

6.F_{slen}: in this feature sentence length is used. It detects the type of sentence, if it long or short sentence. It is calculated by detecting the total number of words in the sentence and dividing it by the total number of sentences in the given article.

Table.1. *Features of set A*

Feature Name	Feature Description
F_{small}	Total number of words smaller than 3 characters /total number of words
F_{bet}	Total number of words between 3-5 characters /total number of words
F_{long}	Total number of words longer than 5 characters /total number of words
F_{dig}	Total number of digital words /total number of words
F_{voc}	Total number of unique words /total number of words.

Set B: linguistic features

1.F_{adj}: in this feature all types of adjectives are detected such as definite adjective and indefinite adjective. It calculates by dividing total number of adjectives by the total number of words in the given article. It is type of POS features as shown in table 2.

2.F_{conj}: This feature is referred to Conjunction Counting (CC) which represents the frequency of conjunction words identified by part of speech tagging and dividing it by the total number of words in the given article.

3.F_{adv}: this feature is defined as normalized number of adverbs. It represents the total number of adverbs words detected by Stanford part of speech tagging normalized by dividing it by the total number of words in the given article.

4.F_{quan}: in this feature all types of Quantities are detected such as (كل, كلا, معظم, جميع). It calculates by dividing total number of Quantities words by the total number of words in the given article.

5.F $_{mod}$: it defined as Normalized Number of Modal Verbs such as (عسى, نعم, بئس). It represents the total number of modal verbs detected by Stanford part of speech tagging normalized by dividing it by the total number of words in the given article.

6. Verbs Tense Mood: after Applying complex NLP techniques to identify the verb tense which is one of effective features to identify the author's style writing. When the verbs tense is extracted, two features are detected (F pres, F past). Each feature is calculated by dividing total number of verb type by all numbers of words in article.

7.F $_{viet}$: it represents the total number of Vietnamese words detected by Stanford part of speech tagging normalized by dividing it by the total number of words in the given article.

8.F $_{prep}$: which represents the total number of prepositions words detected by Stanford part of speech tagging normalized by dividing it by the total number of words in the given article.

9. Personal Pronouns: personal pronouns have three types of features called First Person, Second Person and Third Person. To calculate each one of them, we divided the total number of each feature by all the words of article. It is type of grammatical person features.

***Table2 .Features** of set B*

Feature Name	Feature Description
F $_{adj}$	Total number of Adjectives/total number of words
F $_{conj}$	Total number of Conjunction /total number of words
F adv	Total number of Adverbs /total number of words
F quan	Total number of Quantities /total number of words
F mod	Total number of Modal Verbs /total number of words
F pres	Total number of Verbs (Present) /total number of words
F past	Total number of Verbs (Past) /total number of words
F viet	Total number of Vietnamese /total number of words
F prep	Total number of Prepositions /total number of words

Set C: linguistic features based sentences

1.F_{neg}: Sentence Contain Negative (SCN). This feature counting the number of sentence which contains negative word then dividing it by the total number of sentences in the given article. It is type of content specific features as shown in table 3.

2.F $_{ques}$: Question Sentence Frequency (QSF): The feature is adopted to counting the sentences which written in question forms. Question form means sentence contain question marks and/or question words only identified by part of speech tagging. It normalized by dividing it by the total number of sentences in the given article.

3.F $_{com}$: Comparative Form Sentence: The feature is adopted to counting the sentences which contain comparative form identified by part of speech tagging. It normalized by dividing it by the total number of sentences in the given article.

4.F $_{cert}$: Certain Sentence Frequency (CSF): The feature is adopted to counting the sentences which contain certain word. It normalized by dividing it by the total number of sentences in the given article.

***Table.3** .Features of set C*

Feature Name	Feature Description
F $_{neg}$	Sentence Contain Negative / total number of sentence
F $_{ques}$	Question Sentence Frequency / total number of sentence
F_{com}	Comparative Form Sentence / total number of sentence
F_{cert}	Certain Sentence Frequency / total number of sentence

4-EXPERIMENTAL RESULTS

We suggest different types of features in the previous section that reflect the style of the author to verify the authorship of Arabic documents. Now, when classifying Arabic articles, we need to evaluate the impact of each feature. To perform this goallearning model is be achieved by using learning model and training articles to detect its authors.

Dataset

The absence of data that could help assess the output of the system is one of the main restrictions of Arabic study. Since there is a lacking in Arabic dataset, we used the dataset that is created by [22]. In this experiment, an Arabic corpus was used to maximize the performance, and to measure the method upper and lower accuracy bounds. The Arabic dataset includes ten different authors collected from the Alwaraq site (http:/www.alwaraq.net). The aim of the experiment is to measure how much accuracy can be reached with the proposed approach for the problem of authentication of Arabic authorship.

Learning models

AdaBoost, short for Adaptive Boosting, is type of ensemble method. It builds a strong classifier from set of weak classifiers to enhance the performance. AdaBoost is used with noisy data as, it is sensitive to it. Also AdaBoost represents to a particular technique of training a boosted classifier. The following equation (1) represents a boost classifier form.

$$F_K(z) = \sum_{t=1}^{K} f_t(z) \qquad (1)$$

Where an object z is used as input for each weak classifier f_t .Then each classifier returns output value which match object's class. The K_{th}classifier is return author one if the input article is similar to author one's writing. Each weak classifier creates an output hypothesis,$h(z_i)$ for each article in the training data. In each iteration t a weak classifier is chosen and given a coefficientα_t. E_t is referred to sum training error which calculating during processing iteration as the following equation (2):

$$E_t = \sum_i^{\cdot} E[F_{t-1}(z_i) + \alpha_t h(z_i)] \qquad (2)$$

Where the boosted model is $F_{t-1}(z)$ which has been carried out to the previous training process.$f_t(z) = \alpha_t h(z)$ is the weak classifier which is added to the final classifier and $E(F)$ represents to some error function. A weight $w_{i,t}$ at each iteration during training process equals the value of error $E[F_{t-1}(z_i)]$. These weights can be used to inform the training of the weak classifier, for instance, decision trees can be grown that favor splitting sets of samples with high weights.

Bagging is a type of ensemble method which is a simple and very powerful. It is designed to enhance the stability and performance of machine learning algorithms that used in statistical classification and regression. It is a special case of the classifier averaging method. Also it minimize variance and helps to avoid overfitting. It can used with any type of classifiers and it is usually used decision tree classifier.

Evaluating Results

The accuracy is used to calculate the efficiency of the system proposed. It is calculated as the following equation.

$$Acc = \frac{T_a}{N_a} \qquad (3)$$

Where T_a (total number of true extracted results), N_a (all numbers of used articles).

To measure the performance of the proposed system, three sets of features are used. The first experiment aimed to measure the performance of using Bagging model with the all sets of features. In second experiment, we measure the effect of using AdaBoost model with our proposed features.

***Table 4**: Results of sets of Bagging*

Author	Bagging Accuracy
Alfarabi	91.6%
Alghazali	88.3%
Aljahedh	93.3%
Almas3ody	95%
Almeqrezi	98.3%
Altabary	88.3%
Altow7edy	91.6%
Ibnaljawzy	91.6%
Ibnrshd	85%
Ibnsena	91.6%
Average Accuracy	91.5%

In Table 4, the results of using bagging model changes from 85% to 98.3% of accuracy. The Bagging model achieves 91.5% in average of accuracy. In table 5, the results of using AdaBoost changes from 88.3% to 98.3% of accuracy. Also, AdaBoost model achieves an average accuracy of 93.6%.

***Table 5:** Results of using AdaBoost model*

Author	AdaBoost Accuracy
Alfarabi	91.6%
Alghazali	95%
Aljahedh	93.3%
Almas3ody	96.6%
Almeqrezi	98.3%
Altabary	91.6%
Altow7edy	93.3%
Ibnaljawzy	95%
Ibnrshd	93.3%
Ibnsena	88.3%
Average Accuracy	93.6%

During the last two experiments, we found that the AdaBoost accuracy is higher than the Bagging classifier. Our proposed features are effective in some used articles and can obtained a high result in authorship authentication. We compared our proposed system with Altheneyan et al.[22]. They obtained 92.03% with using MNB model and 87.40% with MPNB model. Our proposed method achieves the best result of 93.6% with using AdaBoost model.

Conclusion

The primary objective of this paper is to solve the problem of forensic authorship authentication for Arabic articles. Various features were used to recognize each author's writing style. We divided the detected features to three sets, set A, set B and set C. The linguistic features were extracted by using POS analysis that have two types based on words and based on sentences. We used two types of machine learning ensemble method, AdaBoost and Bagging models. To evaluate the performance of the

method, two experiments were applied with different sets of features. The AdaBoost model achieve a high results with all different sets of features. The highest result was achieved is 93.6% with AdaBoost using set A, set B and set C of the features.

Reference

[1] E. Stamatatos, "A survey of modern authorship attribution methods," J. Amer. Soc. Inf. Sci. Technol. , vol. 60, no. 3, pp. 538–556, 2009, doi: 10. 1002/asi.21001.

[2] I. Markov, J. Baptista, and O. Pichardo-Lagunas, "Authorship attribution in portuguese using character n-grams," Acta Polytechnica Hungarica, vol. 14, no. 3, pp. 59–78, 2017.

[3]S. Lahiri and R. Mihalcea, "Authorship attribution using word network features," 2013, arXiv:1311.2978. [Online]. Available: https://arxiv.org/abs/1311.2978.

[4]M. G. Kendall, F. Mosteller, and D. L. Wallace, ‘‘Inference and disputed authorship: The federalist,’’ Biometrics, vol. 22, no. 1, p. 200, Mar. 1966.

[5]E. Dauber, R. Overdorf, and R. Greenstadt, ‘‘Stylometric authorship attribution of collaborative documents,’’ in Proc. Int. Conf. Cyber Secur. Cryptogr. Mach. Learn. , Jun. 2017, pp. 115–135.

[6]P. Szwed, ‘‘Authorship attribution for polish texts based on part of speech tagging,’’ in Proc. Int. Conf., Beyond Databases, Archit. Struct. Cham, Switzerland: Springer, May 2017, pp. 316–328.

[7]P. P. Paul, M. Sultana, S. A. Matei, and M. Gavrilova, "Authorship disambiguation in a collaborative editing environment," Comput. Secur. , vol. 77, pp. 675–693, Aug. 2018.

[8]C. Akimushkin, D. R. Amancio, and O. N. Oliveira, "On the role of words in the network structure of texts: Application to authorship attribution," Phys. A, Stat. MechAppl. , vol. 495, pp. 49–58, Apr. 2018.

[9]J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, I. Batyrshin, D. Pinto, and L. Chanona-Hernández, ‘‘Application of the distributed document representation in the authorship attribution task for small corpora,’’ Soft Comput. , vol. 21, no. 3, pp. 627–639, 2017.

[10]A.-F. Ahmed, R. Mohamed, and B. Mostafa, "Machine learning for authorship attribution in Arabic poetry," Int. J. Future Comput. Commun. , vol. 6, no. 2, pp. 42–46, Jun. 2017.

[11]F. M. Giraud and T. Artières, ‘‘Feature bagging for author attribution,’’ in Proc. CLEF (Online Working Notes/Labs/Workshop), 2012.

[12]E. Ekinci and H. Takçı, "Comparing ensemble classifiers: Forensic analysis of electronic mails," Tech. Rep., 2013.

[13]A. Abbasi and H. Chen, "Applying authorship analysis to Arabic Web content," in Proc. Int. Conf. Intell. Secur. Inform. Berlin, Germany: Springer, May 2005, pp. 183–197.

[14]M. Al-Ayyoub, Y. Jararweh, A. Rabab'ah, and M. Aldwairi, "Feature extraction and selection for Arabic tweets authorship authentication," J. Ambient Intell. Humanized Comput. , vol. 8, no. 3, pp. 383–393, 2017.

[15]Shaker, Kareem, and David Corne. "Authorship attribution in arabic using a hybrid of evolutionary search and linear discriminant analysis." 2010 UK Workshop on Computational Intelligence (UKCI). IEEE, 2010.

[16]Abooraig, Raddad, et al. "Automatic categorization of Arabic articles based on their political orientation." Digital Investigation 25 (2018): 24-41.

[17]S. Ouamour and H. Sayoud, "Authorship attribution of ancient texts written by ten Arabic travelers using a SMO-SVM classifier," in Proc. Int. Conf. Commun. Inf. Technol. (ICCI), 2012, pp. 44–47.

[18]Keselj, Fuchun Pengt Dale Schuurmanst Vlado, and Shaojun Wang. "Language Independent Authorship Attribution using Character Level Language Models."

[19]Türkoğlu, Filiz, Banu Diri, and M. Fatih Amasyalı. "Author attribution of Turkish texts by feature mining." International Conference on Intelligent Computing. Springer, Berlin, Heidelberg, 2007.

[20]Luyckx, Kim. "Authorship attribution of e-mail as a multi-class task." Notebook for PAN at CLEF (2011).

[21]Diederich, Joachim, et al. "Authorship attribution with support vector machines." Applied intelligence 19.1 (2003): 109-123.

[22]Altheneyan, Alaa Saleh, and Mohamed El Bachir Menai. "Naïve Bayes classifiers for authorship attribution of Arabic texts." Journal of King Saud University-Computer and Information Sciences 26.4 (2014): 473-484.