Journal of Theoretical and Applied Information Technology
15th September 2020. Vol.98. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTHORSHIP AUTHENTICATION OF POLITICAL ARABIC ARTICLES BASED ON MODIFIED TF-IGF ALGORITHM

HEBA M . KHALIL [1]*, AHMED TAHA [1], TAREK . EL-SHISTAWY [2]

[1] Computer Science Department, Faculty of Computers & Artificial Intelligence, Benha University, Benha, Egypt, {heba.khalil, ahmed.taha}@fci.bu.edu.eg
[2] Information System Department, Faculty of Computers & Artificial Intelligence, Benha University, Benha, Egypt, t.shishtawy@bu.edu.eg
* Corresponding author's Email: heba.khalil@fci.bu.edu.eg

ABSTRACT

Recently, authorship forensic analysis for political articles has become very important. It is the process in which a linguist attempts to identify the author of an anonymous text based on the vocabulary used and the linguistic style of the writer. The most existing studies of authorship forensic analysis focus on the English language, while researches concerning the Arabic language is rare. In this research, we present a new methodology that enhances authorship forensic analysis focusing on the Arabic language. The basic idea is to extract the unique vocabulary terms identifying the author (or a political group) and used for recognition of unknown authors. In the current work, a Term Frequency- Inverse Group Frequency (TF-IGF) is proposed, which is a modification of the traditional TF-IDF method. Our approach is tested with large political dataset and determine the performance of Authorship forensic analysis method based on vocabulary words.The experimental results show that the average accuracy for recognizing groups has increased from 89.33 % when using TF-IDF, to 92% with the proposed TF-IGF. Further improvement is achieved when representing the vocabulary terms in its Arabic lemma form, rather than its root form. The results show that the accuracy is improved from 89.33 % to 92%.
Keywords— Authorship Forensic Analysis, TFIDF, Term Weighting, Linguistic Style.

1. INTRODUCTION

There are many categories for authorship forensic analysis. The scope of authorship forensic analysis problem does not include only identifying the author of an unknown text, there are several author analysis tasks for authorship problem, including author attribution and verification, author profiling, and plagiarism detection. In author attribution, the author for an unknown text is identified. However in author verification, a decision is given If or not a specified text was written by a particular author. In this paper, our study scope will focus on author authentication for political Arabic articles. In addition, author profiling is concerned with extracting some special information (such as the author age, education, sex, etc.) from his written text. Finally, plagiarism detection can be considered as an application of author attribution problem. The main goal is to find the similarity between two text documents.

The method of deciding the authorship of the document is an increasing problem in verification knowledge and in computer science. It has many valuable applications, such as art history and counter-terrorism research [1]. The authorship authentication problem means determining the author of a given text from a set of authors based on the vocabulary words or the writing style of the author [2]. The problem of authentication of authorship has its historical roots in the existence of many texts with a suspect attribution to well-known authors[3]. In the Arab world, political articles are different from other articles due to their subjectivity. The author's convictions and political affiliation heavily influence his political article. In the Arab world, there are many different political ideologies such as Liberal, Islamic Sunni, Islamic Shia, Arab Nationalists, Communists, Socialists, etc. [4]. The capacity to identify an article's political ideology immediately plays a critical role about safety and security in many countries.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Each author has a particular style of writing which can be calculated using stylistic or vocabulary words. In Arabic authorship authentication, most of the previous studies deal with linguists to identify unique Stylistic Features (SF) of text. The only difference was the type of features such as lexical, structural, syntactic, and content-specific. The two measures do not change over time, and can therefore be regarded as unique identifiers [5]. Most current authentication methods fall in one of the two methods: the Bag-Of - Words (BOW) method, and the Stylistic Features (SF) method. In the method of vocabulary, the feature vector for each document is calculated on the basis of the word frequencies within the document, and it is rare of these words in other authors' documents. The SF process focuses on SF computing to detect the differences between the various authors' writing styles. Typically this method depends on the usage of lexical, textual, syntactic, and content-specific features. BOW method uses thousands of features and needs some techniques for reducing features like stemming (reducing words to their roots), stopping word removal etc.

The main aim of this paper is to introduce an Arabic Authorship identification method on political articles written in Modern Standard Arabic (MSA). The method will rely on vocabulary words, which detect the crucial words for each author. The contributions of the paper can be highlighted as follows. The method proposed a new modified TF-IGF algorithm for authorship authentication. The area of authorship attribution is still primarily known for the Arabic language, to the best of our understanding. Moreover, the use of vocabulary words features is better than extracted features, as extracted features are extensive, and this makes some classifiers unable to experiment. Finally, the system suggested has achieved a high accuracy of authentication tasks for a large number of articles.

2. BACKGROUND AND RELATED WORKS

In this section, background information about the authorship identification problem is introduced. We discuss some details about the Arabic language. Then, some of the previous work on authorship identification is presented, especially in the Arabic language. The topic of authorship identification is still primarily understudied for the Arabic language, with only few publications. The development in Arabic language research is challenging because the available resources are little.

Arabic language

In the world Arabic language is one of much widely-spoken languages. In many other countries it is the native language in the Arab nations as well as a secondary language. Arabic‘s alphabet is made up of 28 letters addition to special characters and punctuation marks. In addition the direction of writing in Arabic is start from right to left [6], [7]. The writing shape of the letter changes depending on its position within the word. So, whether the letter arrives first, in the middle, or at the end of the word, it changes the letter shape. In addition, Arabic has diacritics, symbols placed above or below the letters to reduce the pronunciation letter or to give short vowels. Most Arabic-text study performed learning algorithms only designed for English text without making any notable changes[8]. One big issue connected with classification of Arabic text is the absence of standardized and infrequently reported Arabic datasets. In addition, they usually do not publish their data to be used by other researchers. So confidence is not strong enough in the outcomes obtained from these experimental studies. Researchers also claimed, however, that classification of Arabic texts is a very difficult activity because of language difficulty [9].

Arabic stemmers and lemmatizer

In linguistic morphology, stemming refers to reduce words to their word base or root form. It is a necessary process for reducing the number of words that have the same root in the most morphological languages, like Arabic.
The designer of any stemmer should have many information such as prefixes, suffixes, infixes, and all words in the language. In Arabic, the most common stemming techniques include the Khoja stemmer [10] and the light10 stemmer [11]. Lemma captures semantic similarities between words, referring to the set of all word forms having the same meaning [12]. The need for representing Arabic words at the lemma level became very important for many applications.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Authorship analysis

Analysis of authorship is a problem in which the aim is to determine the author of a particular text given a set of text documents written by different authors. We discuss the previous work of the authorship authentication problem (attribution). Most earlier work on authentication of Arabic authorship considers documents as a bag of words where the text is viewed as a vector of weighted frequencies for each of the individual words or tokens. Our paper ueses this method to solve the problem of Arabic article authentication.

Abooraig, R., et al. [4] 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 and a combination of classifiers. The Support Vector Machine (SVM)[13] classifier achieved the best accuracies with selection features. This method depended on features that not sufficiently strong to detect unknown authors. Jindal [14] presented an author identification system in Arabic. This method is based on extracting root-words in Arabic to extract the features. The system reached to a precision of 91.2% and a recall of 90.9% and geting an F-measure of 91.02%. Ehsan Ali [15] proposed the Stylometric Authorship Balanced Attribution (SABA) method. This approach is carried out by combining three methods: the analytical method, the Burrows-delta method and Winnow 's algorithm. The SABA approach often uses a more powerful range of attributes relative to frequent word tool. So far this leads to higher Stylometric prediction, having more accuracy for authorship recognition and prediction of artistic writing style. The successful attributes are defined by the frequent word, pair, and triple, while these are attributes with multiple meanings. The final results detected that, during the final phase of the experiment, the SABA method produces superior predictive accuracy and even delivers a completely correct result. Stefan Ruseti *et al.* [16] obtained from the documents character trigrams, suffixes, POS bigrams and trigrams, syntactic complexity, word length, and structure and percentage of direct speech to represent the document vector. It has been found that the accuracy is improved when different features of the program are implemented. They tested with the Sequential Minimal Optimization (SMO) algorithm[17] and achieved an average accuracy of 77 % in the detection of authors. In[18], Amasyali M. F. et al. presented a new method using the n-gram model for Turkish text. They demonstrated whether or not the modeling of Turkish text with n-grams is a good method for deciding the author's language, text type, and author's gender. On bi-gram and tri-gram models, four separate classifiers were used. They noticed that Naïve Bayes classifier offered the best outcome in defining the author of the document, and the bi-gram model is more effective in deciding the author of the document than tri-gram. Shrestha et al. [19] proposed a technique for the usage of short text CNNs with character n-grams for Authorship Attribution (AA) and produced a detailed analysis of common strategies. They observed that CNNs [20]] giving higher performance for AA tweets. . In addition, using n-grams rather than primarily character sequences can also increase results. They have gained some insights into what our current design is learning.

3. PROPOSED METHOD

The main purpose of the proposed method is the author recognition. The proposed method used traditional TF-IDF and modified TF-IDF, which is called TF-IGF. Besides, it uses traditional TF-IDF and TF-IGF Any types of author writing may be recognized. For eg, if the author has any more regular terms in his writings than the other writers, then such characteristics will help to identify the author. The proposed method consists of some phases are detailed below and shown in Figures 1.

Table.1 .sample of Arabic stop words

Sample of Arabic stop words				
هذه	امام	ايضا	نفسه	في
نحو	الذى	التى	هناك	كل
لدى	الان	اما	فيها	لم
قبل	المقبل	الا	وهى	له
فان	جميع	اكد	وهو	من
هذا	انها	حيث	ومن	هو
مساء	انه	يمكن	ولم	هى
مع	الى	منها	وفى	كما
ما	حين	بين	لكن	لها
لا	حول	اللذين	كلم	منذ

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

قد	دون	حوالى	فيه	وقد
اذا	ذلك	خلال	وقال	ولا

found in Arabic corpus. Table (1) shows some Arabic stop words.

Pre-processing stage

The text is divided at this stage, and unnecessary terms are eliminated. It includes the following steps: tokenisation, removal of the stop-word, extraction of the root, and stemming. Tokenization is the mechanism by which text is separated into tokens. The Arabic Tokenizer uses White Space in tokenization, since space is still the only way to distinguish terms in Arabic. It is a crucial step to make the detection of unwanted tokens very easy. Stop-words are common words

Effect of Word Representation

After cleaning the Arabic corpus from unwanted terms, some versions are extracted from Arabic corpus based on root and lemma. Arabic corpus based on root contains words without any additions. This process is applied to the Arabic corpus's words to delete affixes (suffixes and prefixes) letters and to extract the words in the root form. Lemma is the second version of the Arabic corpus that defined as the set of all forms of words having the same meaning.

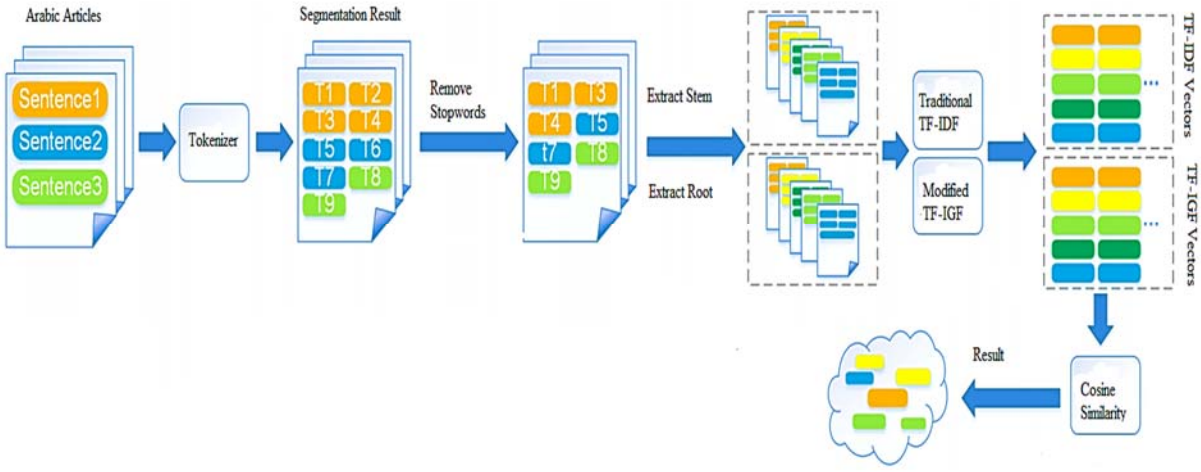

Figure.1 the framework of the proposed method.

Traditional (TF-IDF) method

Term Frequency Inverse Document Frequency (TF-IDF) is a popular method to extract word weighting. The primary purpose of TFIDF is to detect rare words in the corpus by assigning them a high weight than commonly found words in a corpus. TF-IDF is a combination of two algorithms: Term frequency (TF) and Inverse Document Frequency (IDF). TF Is determined by recording a word's amount of occurrences in an article. It also is the simplest method to weight each term. The weight of term t on article a is calculated using equation (1).

$$TF_{t,a} = f(t,a) \qquad (1)$$

Where, $f(t,a)$ represents the frequency of occurrence of term t on article a.

IDF takes the occurrences of terms in the collection of articles into account. The background of this type of weighting is to value a rarely appearing term in the whole collection of articles more. For the statistical measure which supports the assumption that a more frequent term is viewed less valuable in the group. The IDF factor of term t is calculated using equation (2).

$$IDF_t = \log(\frac{N}{af_t}) \qquad (2)$$

Where N is the total number of articles and af_t represents the number of articles containing term t. TF-IDF algorithm is calculated by multiplying both TF and IDF together, as shown in equation (3).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$TFIDF_{t,a} = TF_{t,a} * IDF_t \quad (3)$$

It is shown from equation (3) that the calculations of TF-IDF are based on the importance of each term in each article. The accuracy in TF-IDF is not high because the deduced relation here depends on the number of term occurrences in each article, not in all the groups.

Modified (TF-IGF) method

The TF-IGF is proposed as an extension of inverse document frequency (IDF) to enhance its accuracy. The main idea in TF-IGF algorithm is not only considered the number of occurrences of a term in an article but also, it takes into consideration the proportion of the total number of term occurrences in the group's articles to the total number of distinct terms in the same group. The research is based on the assumption that each group uses words from a list of lexical terms characterizing it. TFIGF has a value ranging from 0 (when the term is not used by the group at all) and one (when the group only uses the term).

$TF_{t,a}$ is the same as what is defined earlier in equation (1) while the dominator denotes to normalization here. Euclidean norm is used for normalizing and is calculated as seen in equation (4). The normalization of articles is very important because a small article would contain very little information and a huge article having the same term frequency.

$$NTF_{t,a} = \frac{TF_{t,a}}{\sum_{t=1}^{m} TF_{t,a}^2} \quad (4)$$

Where, $TF_{t,a}$ represents the number of occurrences of a term t in an article. The denominator is known as the Euclidean norm of the articles where, $TF_{t,a}$ is a frequency of t^{th} term in an article a. The modification made to the IDF equation, as shown in equation (5) leads to the new general modified formula of TF-IDF, as shown in equation (6).

$$IDF_{modified} = IGF_{t,a} = 1/T_t \quad (5)$$

$$TFIGF_{t,a} = NTF_{t,a} * IGF_{t,a} \quad (6)$$

$T_t$is the total number of term occurrences in all group's articles, and it can be calculated as shown in equation (7). A term rarely appearing in the group of articles is considered valuable. The importance of each term is assumed to have low importance to the number of articles containing the term.

$$T_t = \sum_{a=1}^{G} TF_{t,a} \quad (7)$$

Where, $TF_{t,a}$ is a frequency of t^{th} term in all groups and G is the number of all articles.

Vector Space Model

Vector Space Model (VSM) is an algebraic algorithm developed to represent the documents as a series of terms and conditions. VSM is converted from a full-text form to a vector that has a particular sequence of frequency. It represents article a as a vector of terms, $a = (t_{1j}, t_{2j}, t_{3j}, t_{4j}, \ldots, t_{ij})$, where t_{ij} is the weight of the ith term in jth article vectors is calculated using equation (8). In the current implementation, Islamic groups or organizations are represented as vectors in a common vector space. Each group is considered as a point in this space whose location is determined by the n-lexical terms relative importance.

$$\begin{bmatrix} \vec{v}_1 = \{TFIGF_{1,1} + TFIGF_{2,1} + \cdots + TFIGF_{n,1}\} \\ . \\ . \\ \vec{v}_m = \{TFIGF_{1,m} + TFIGF_{2,m} + \cdots + TFIGF_{n,m}\} \end{bmatrix} \quad (8)$$

Where, $\vec{v}_m$ represents the number of vectors over the group.

Cosine Similarity

To calculate the similarity, the functional tool is used to measure the similarity between the tested articles victories and the main categories victories. When the functional tool is used for similarity measure, the output is a numeric value in a range of zero and one. When the output is zero, this means that article vectors are dissimilar. However, when the output is one, this means the article vectors are exactly similar. In this method, cosine similarity is used to measure the similarity. Usually, similarity measures are implemented on vectors. The Similarity measure of any two-article vectors $\vec{v}_l$ and $\vec{v}_2$ is computed and using the following equation 9. The algorithm 1 is defined the all process to authenticate an article.

$$\cos(\theta) = \vec{v}_1 . \vec{v}_2 \quad (9)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Note: The similarity between two vectors is the same if they have zero angular distance between them, but they are entirely different if one is orthogonal to the other.

Algorithm 1 Process for authentication an article

Input : Arabic text articles
Output : category name
1**: procedure AUTHENTICATE (***article***)**
2**:** *terms* =tokenize(*article*);
3: for all *t* ∈ *terms* do
4: *weight* = $\boldsymbol{TF - IGF(t)}$;
5: *articleVector.put(t, weight);*
6: *categoriesVectors* = LoadCategoryVectors();
7: for all *c* ∈ *categoriesVectors* do
8*:* *similarity* = cos_Similarity (*c, articleVector*);
9: *result*.put(*c.name, similarity*);
10: save *result*;

4. EXPERIMENTAL RESULTS

DATASET

One of the major limitations of Arabic research is the lack of resources that might help measure the system's performance. Since a standard Arabic corpus is missing, we used the corpus that is presented in[4]. In this experiment, an Arabic corpus was used to maximize the performance, and to measure the method upper and lower accuracy bounds. The Arabic corpus contains five political ideologies with 600 Arabic articles in each ideology. We divided each ideology to 150 articles in the testing data and 450 for training data. The percentage of the testing data is 25%, and training data is 75% of all the data. This corpus is extracted from social networks and other websites (i.e., Tweeter, Facebook, and Yahoo, etc.). This collection relates to the political articles of Arabic and is component of five ideologies: Arab Nationalist (Nat), Brotherhood (Bro), Islamic Shia (Shi), Liberal (Lib), and Socialist (Soc). The purpose of the experiment is to measure how much accuracy can be achieved with the proposed method TF-IGF for articles.

Theoretical Evaluation of Accuracy

To measure the accuracy of our proposed approach of modified TF-IGF, we compare it with the traditional TF-IDF. Both methods use the cosine similarity measures to determine the closest class of each article. So for comparing the efficiency, we need to define a method that indicates the classification correctness. We propose here equations, which can be defined for a tested article as:

$$NTF_{t,a_{tested}} = \frac{TF_{t,a_{tested}}}{\sum_{t=1}^{m} TF_{t,a_{tested}}^2} \quad (10)$$

Where, $TF_{t,a_{tested}}$ represents the number of occurrences of a term t in an article tested, $TF_{t,a_{tested}}$ is a frequency of t^{th} term in an article tested a_{tested}.
The IGF is calculated for the article tested as the following equation:

$$IGF_{t,a_{tested}} = \frac{1}{T_t} \quad (11)$$

Where, T_t is total number of term occurrences in all group's articles as shown in equation (5).
After calculating $NTF_{t,a_{tested}}$ and $IGF_{t,a_{tested}}$ for article tested, the value of $TFIGF_{t,a_{tested}}$ is calculated as the following equation.

$$TFIGF_{t,a_{tested}} = NTF_{t,a_{tested}} * IGF_{t,a_{tested}} \quad (12)$$

The vector space for the article tested is calculated as the following equation.

$$\vec{v}_{tested} = \{TFIGF_{1,tested} + TFIGF_{2,tested} + \cdots + TFIGF_{n,tested}\} \quad (13)$$

We propose here Cosine Similarity value, which can be defined for a certain method as:

$$\cos(\Theta) = \vec{v}_{tested} . \vec{v}_m \quad (14)$$

Where, $\vec{v}_m$ is a number of vectors of all classes.
This equation requires measuring the cosine similarity values of both $\vec{v}_1$ with $\vec{v}_2$. Then the absolute difference between the two values is calculated. The result value of cosine similarity determines if both $\vec{v}_1$ and $\vec{v}_2$ belong to the same class or a different class. The cosine similarity value becomes very small if both $\vec{v}_1$ and $\vec{v}_2$ belong to different classes. It is expected that the cosine similarity value is high if both $\vec{v}_1$ and $\vec{v}_2$ belong to the same classes.

Measuring Performance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To evaluate the performance of the proposed method, the accuracy is used. This measure is used widely in literature [13] It is defined as follows:

$$\text{Accuracy} = \frac{\text{TP}}{\text{N}} \quad (15)$$

Where TP (number of true results of author authentication determined by the approach), N (total number of articles).

Table 2 shows the cosine similarity values of each ideology with the other ideologies based on root form. It is noted that the cosine similarity values between ideologies based on the root are very high. On the other hand, the cosine similarity values between ideologies are very low based on the lemma, as shown in Table 3. The accuracy of liberal ideology is the worst value based on the root as its terms are more similar to Arabic nationalist. Through the experiment, we found that using data in root form causes high similarity between ideologies and achieves the worst accuracy. On the other hand, using data in lemma form causes low similarity between ideologies and achieves the highest accuracy.

Table2: cosine similarity values for all ideologies based on root

	Arab Nationalist	Brotherhood	Islamic Shia	Liberal	Socialist
Arab Nationalist	1	0.40	0.34	0.39	0.39
Brotherhood	0.40	1	0.36	0.36	0.36
Islamic Shia	0.34	0.36	1	0.31	0.32
Liberal	0.39	0.36	0.31	1	0.36
Socialist	0.39	0.36	0.32	0.36	1

Table 3: cosine similarity values for all ideologies based on lemma

	Arab Nationalist	Brotherhood	Islamic Shia	Liberal	Socialist
Arab Nationalist	1	0.09	0.07	0.096	0.09
Brotherhood	0.093	1	0.081	0.082	0.075
Islamic Shia	0.072	0.081	1	0.063	0.066
Liberal	0.096	0.082	0.063	1	0.078
Socialist	0.090	0.075	0.066	0.078	1

We compare our proposed method TF-IGF with traditional method TF-IDF with the same Arabic corpus, as shown in Table 4. In our experiment, the average accuracy of the proposed method TF-IGF is 92% based on the lemma, while it is 89.33% for traditional TF-IDF based on lemma. The difference gape between the two methods is that TF-IDF does not care about the frequency of a term in all the groups. It is found that the traditional TF-IDF can achieve the highest average accuracy of 84.6% based on the root. Although the proposed method TF-IGF gave the worst average accuracy of 61.33% based on the root. It is noted that in our proposed method, using the lemma is better than

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: results of proposed method and compared method

Ideology	TF-IDF Accuracy based on root	TF-IDF Accuracy based on lemma	TF-IGF Accuracy based on root	TF-IGF Accuracy based on lemma
Arab Nationalist	86.6%	83.3%	60%	100%
Brotherhood	60%	80%	73.33%	100%
Islamic Shia	100%	100%	46.6%	86.66%
Liberal	90%	90%	26.6%	73.33%
Socialist	86.6%	93.3%	100%	100%
Average accuracy	84.6%	89.33%	61.33%	92%

the root because the lemma is more accurate, and it takes the context of the word in mind. So the terms in lemma form determine more rare words that producing high performance in the classification process. On the other hand, the terms in root form achieve the lowest accuracy because it uses many words that are commonly used with other ideologies.

5. CONCLUSION

In this paper, the main aim is the authentication of Arabic political articles. This paper aims to present a novel method that enhances authorship forensic analysis focusing on the Arabic language which is still largely understudied despite its importance. This work is very important to used in security field. The proposed approach will rely on vocabulary words, which detect the crucial words for each author.Two methods are used for the authentication, traditional TF-IDF, and modified TF-IGF. The basic idea is to extract the unique vocabulary terms identifying the author (or a political group) and used for recognition of unknown authors. Two versions of Arabic corpus have been extracted, one based on root form and the other based on lemma form. The result of the proposed method that depends on modified TF-IGF is the best one of 92% average accuracy based on lemma form. On the other hand, traditional TF-IDF achieves a high average accuracy of 89.3% based on root form.

REFERENCES

[1] M. Kestemont, J. Stover, M. Koppel, F. Karsdorp, and W. Daelemans, "Authenticating the writings of Julius Caesar," *Expert Syst. Appl.*, vol. 63, pp. 86–96, 2016.

[2] 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. Mech. its Appl.*, vol. 495, pp. 49–58, 2018.

[3] A.-A. Mahmoud, A. Ahmed, and H. Ismail, "An extensive study of authorship authentication of Arabic articles," *Int. J. Web Inf. Syst.*, vol. 13, no. 1, pp. 85–104, Jan. 2017.

[4] R. Abooraig, S. Al-Zu'bi, T. Kanan, B. Hawashin, M. Al Ayoub, and I. Hmeidi, "Automatic categorization of Arabic articles based on their political orientation," *Digit. Investig.*, vol. 25, pp. 24–41, 2018.

[5] H. Sayoud, "Author discrimination between the Holy Quran and Prophet's statements," *Lit. Linguist. Comput.*, vol. 27, no. 4, pp. 427–444, May 2012.

[6] E. Amer, H. M. Khalil, and T. El-shistawy, "Hierarchical N-Gram Algorithm for Extracting Arabic Entities," in *Proceedings of the 10th International Conference on Informatics and Systems*, 2016, pp. 56–60.

[7] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," *J. King Saud Univ. - Comput. Inf. Sci.*, 2019.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] N. Durrani and S. Hussain, "{U}rdu Word Segmentation," in *Human Language Technologies: The 2010 Annual Conference of the North {A}merican Chapter of the Association for Computational Linguistics*, 2010, pp. 528–536.

[9] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic Natural Language Processing and Machine Learning-Based Systems," *IEEE Access*, vol. 7, pp. 7011–7020, 2019.

[10] M. Sawalha and E. Atwell, "Comparative Evaluation of {A}rabic Language Morphological Analysers and Stemmers," in *Coling 2008: Companion volume: Posters*, 2008, pp. 107–110.

[11] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light Stemming for Arabic Information Retrieval," 2007.

[12] T. El-Shishtawy and F. El-Ghannam, "An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes," *CoRR*, vol. abs/1203.3. 2012.

[13] S. Larabi Marie-Sainte and N. Alalyani, "Firefly Algorithm based Feature Selection for Arabic Text Classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 3, pp. 320–328, 2020.

[14] V. Jindal, "A Personalized {M}arkov Clustering and Deep Learning Approach for {A}rabic Text Categorization," in *Proceedings of the {ACL} 2016 Student Research Workshop*, 2016, pp. 145–151.

[15] T. K. Mustafa, A. A. Abdul Razzaq, and E. A. Al-Zubaidi, "Authorship Arabic Text Detection According to Style of Writing by Using (SABA) Method," *Asian J. Appl. Sci. (ISSN 2321-0893)*, vol. 5, no. 2 SE-, May 2017.

[16] S. Ruseti and T. Rebedea, "Authorship Identification Using a Reduced Set of Linguistic Features," in *CLEF*, 2012.

[17] F. Elghannam, "Text representation and classification based on bi-gram alphabet," *J. King Saud Univ. - Comput. Inf. Sci.*, 2019.

[18] M. F. Amasyal\i and B. Diri, "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender," in *Natural Language Processing and Information Systems*, 2006, pp. 221–226.

[19] P. Shrestha, S. Sierra, F. González, M. Montes, P. Rosso, and T. Solorio, "Convolutional Neural Networks for Authorship Attribution of Short Texts," in *Proceedings of the 15th Conference of the {E}uropean Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 669–674.

[20] N. Habash, O. Rambow, and G. Kiraz, "Morphological Analysis and Generation for Arabic Dialects," in *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 2005, pp. 17–24.