

Authorship Attribution of Seven Arabic Religious Books -A Fusion Approach-

Hadjadj Hassina, Sayoud Halim
USTHB University, Algiers, Algeria

hadjadj.has@gmail.com

halim.sayou@uni.de

Abstract—In this paper, we investigate the authorship of seven Arabic religious books, written by seven religious scholars. The Arabic styles are almost the same (i.e. Standard Arabic) for the seven books. The genre is the same and the topics of the different books are also the same (i.e. Religion). Several experiments of authorship attribution are conducted by using four different features namely: character trigrams, character tetragrams, word unigrams and word bigrams. On the other hand, different classifiers are employed, such as: Manhattan distance, Multi-Layer Perceptron (MPL), SMO-based Support Vector Machines (SMO-SVM) and Linear Regression (LR). Furthermore, a fusion approach has been proposed to enhance the performances of authorship attribution, with two fusion techniques: Feature-based Decision Fusion (FDF) and Classifier-based Decision Fusion (CDF). Results show good authorship attribution performances with an optimal score between 92% and 98% of good attribution. The proposed fusion technique raised this score to 100% of good authorship attribution. Moreover, this comparative survey has revealed interesting results concerning the Arabic language.

Keywords— Authorship analysis; Fusion approach; Natural language processing; Authorship attribution ; Religious books; Text Classification.

I. INTRODUCTION

Stylometry or author recognition is a typical problem in natural language processing. It is a research field that consists in recognizing the authentic author of a piece of text. It is evident that the recognition accuracy is not as high as some biometric modalities that are used in security purposes, but it has been shown that for texts with more than 2500 tokens, the recognition task becomes significantly accurate [1, 2].

Stylometry can be divided into several research fields: Authorship Attribution (referred to as AA) [3], Authorship verification, Authorship discrimination, Authorship Indexing and Plagiarism detection.

Authorship attribution (AA) is research field of stylometry, which consists in identifying the authors(s) of a piece of text by using some techniques of text mining and statistics.

That is; determining the real author of a piece of text has raised several questions and problems for centuries. Problem

of authorship can be of interest not only to humanities researchers, but also to politicians, historians and religious scholars in particular. Thorough investigative journalism, combined with scientific analysis (e.g., chemical analysis) of documents has traditionally given good results [4].

The area of authorship analysis has been researched for many years going back to the early 60s of works such as [3], where the authors were studying the important Federalist Papers case for solving an authorship claim by different authors. In the recent years, there has been growing interest in developing practical applications for authorship identification (Authorship Attribution). These applications focus on many areas such as: email authorship [4], plagiarism detection [5] and forensic cases [6].

Research work on authorship attribution usually appears in several types of debates ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite this interest, the field itself is somewhat in confusion regarding which are the best practices and techniques [7].

An interesting area in identification technologies is Biometric identification which is way to find or verify the identity of who we claim to be, by using physiological or behavioral characteristics [8]. As the human has physiological or behavioural characteristics in every human; he has also linguistic features. Human usage of language, writing, set of vocabulary, unusual usage of words, and particular syntactic and stylistic traits tend to be stable. The big challenge for authorship analysis is locating and learning from such features.

In fact, it is not clear which features of a text should be used to classify an author. So, the principal issue in computer-based author identification is to identify a set of features that represents the author's writing style. These are used to classify the authors of selected unknown texts. A different set of features can be used to identify authors; these include word-level, character-level, syntactic, semantic and lexical features [9].

The literature display several available techniques, which determine the author of a document. According to the literature [10] [11], it appears that many works are reported for the English, Greek and Hebrew languages. But, there are few serious published works in Arabic language, especially for religious texts.

Hence, we will try to make some experiments of Authorship Attribution (AA) on seven Arabic religious books, written by seven religious scholars. We note that the genre of the different books is the same and that the topic (ie. Religion) is the same too.

An interesting new idea is the proposal of the Fusion approach, which we applied in two different forms: Fusion of Classifiers (FC) and Fusion of Features (FF).

The rest of this paper is organized as follows: section 2 presents related works, section 3 gives a description of seven religious books, in section 4, we present the authorship attribution methodology. Section 5 describes the experimental results and finally, section 6 concludes the paper.

II. RELATED WORKS

Many studies have been reported during the last years, where many debates were reported and several types of features and techniques were proposed too.

For instance, Stamatatos conducted a study of the latest advances in automated approaches used in authorship attribution [9]. He examined the characteristics of these approaches for text representation and text classification, and also the evaluation criteria and methodologies used in author identification studies. The survey distinguishes different types of stylometric features to quantify the writing style including character features, lexical features, syntactic and semantic features.

In 2012 Shaker et al. used a hybrid method of evolutionary search and LDA approach [14]. In this survey he investigated the usage of function words that are specific words which are used by the writer in distinct way and which may or may not relate to the subject matter. The approach was tested on Arabic and English documents.

recently, a plethora of models more familiar to machine learning practitioners than linguists such as support vector machines, neural networks, latent Dirichlet allocation, decision trees have been applied to different types of features with success [15] [16] [17].

Seroussi et al. use authorship attribution of informal text such as e-mails with topic modelling [18]. Disjoint Author-Document Topic (DADT) model was suggested that projects authors and documents to two disjoint topic spaces. Latent

Dirichlet Allocation (LDA), Author-Topic (AT) and DADT models are implemented on formal as well as informal.

Ouamour et al. employed several character N-grams [19]. The authors examined the authorship of Arabic books written by ten Arabic travelers. Different types of features were used such as character, character-bigram, character-trigram and character-tetra gram. For the classification, they used Stamatatos distance, Manhattan distance, Multi-Layer Perceptron (MLP) and Support Vector Machines (SVM).

One can find a couple of recent works of author discrimination in Arabic but very few are applied to the Quran. Sayoud presented a series of author discrimination experiments between the holy Quran and Hadith [13]. Once, he used the two books in their entirety and another time, he segmented the books into 4 segments each. In both experiments he showed that the authors of the two books are different. Later on, he published another article describing an experiment of author discrimination between the holy Quran and Hadith by using a hierarchical clustering. Results were interesting since they sharply showed two main clusters representing the two corresponding authors: Quran author and Hadith author.

In 2015 Sayoud presents an experiment of author discrimination between the holy Quran and Hadith by using a hierarchical clustering [17], where seven types of NLP features are extracted. Results were interesting since they sharply showed two main clusters representing the two corresponding authors: Quran author and Hadith author

III. CORPUS OF THE SEVEN RELIGIOUS BOOKS

As cited previously, there are seven different books written by seven religious scholars. We recall that the Arabic styles are almost the same (i.e. Standard Arabic) for the 7 books, the genre of the books is the same and the topics are also the same (i.e. Religion). We called this dataset: **SAB-2** (Seven Arabic Books – dataset two). These books are described as follows:

1stbook: text collection of Alghazali (Author: Mohammed al-Ghazali al-Saqqā): it contains some articles and dissertations of Alghazali. This author is a contemporary Egyptian religious scholar, who is born in 1917 and died in 1996. Sheikh al-Ghazali held the post of Chairman of the Academic Council of the International Institute of Islamic Thought in Cairo.

2ndbook: text collection of Alqaradawi (Author: Yusuf al-Qaradawi): it contains some articles and dissertations of Alqaradawi. This author is a contemporary Egyptian/Qatari religious scholar, who is born in 1926. He is the head of the European Council for Fatwa and Research, an Islamic scholarly entity based in Ireland. He also serves as the chairman of International Union for Muslim Scholars (IUMS).

3rdbook: text collection of Abdelkafy (Author: Omar Abdelkafy). This text collection contains some articles and dissertations of Dr. Omar Abdelkafy, who was born in Almenia, Egypt on May 1, 1951. He memorized the Holy Quran completely when he was ten years old. Dr. Abdelkafy also memorized Sahih Al-Bukhary and Muslim with full references. Abdelkafy studied Islamic Theology and Arabic Linguistics from clever scholars and started serving the Islamic Dawah in 1972.

4thbook: text collection of Al-Qarni (Author: Aaidh ibn Abdullah al-Qarni). This text collection contains some articles and dissertations of Shaykh Aaidh ibn Abdullah al-Qarni, who was born in 1960. He is a Saudi religious scholar and author of a famous book. Al-Qarni is best known for his distinguished book “La Tahzan” (in English: Don’t Be Sad), which had a lot of success over the time.

5thbook: text collection of Amr Khaled (Author: Amr Mohamed Helmi Khaled).

Several articles and dissertations of Amr Khaled have been collected into a unique text. This author was born in 1967 in Egypt. He is an Egyptian Muslim activist and television preacher. He is often described as “the world’s most famous and influential Muslim television preacher”.

6thbook: text collection of Hassan (Author: Mohamed bin Ibrahim Al-Hassan): it contains some articles and dissertations of Hassan. This author is a contemporary Egyptian religious scholar, who is born in 1926 in Egypt.

7thbook: text collection of Al-Arifi (Author: Mohamed Al-Arifi): it contains some articles and dissertations of Al-Arifi. This author was born in 1970. He is a Saudi author and scholar. He is a graduate of King Saud University, and Member of the Muslim World League and the Association of Muslim Scholars

IV. AUTHORSHIP ATTRIBUTION METHOD

In our approach different steps are performed, as shown in Figure 1, namely: data preprocessing, text segmentation, feature extraction, classification and author discrimination decision, while the data set is collected. In the second step, preprocessing is applied to our dataset. After that, text segmentation is used in order to construct individual texts with the same size.

In the following step, the data is organized into training and testing. Thereafter, the features are extracted from the data during both training and testing. In the fourth step, a classification model is constructed from the training data, and used for the testing process. During the training process, the feature vectors are introduced in association with the author

classes. Finally, the testing process is performed and evaluated according to the decision provided by the classifier.

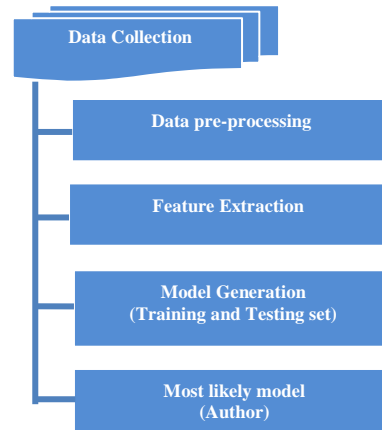


Fig. 1 Typical Procedure for Authorship attribution

A. Data Pre-processing

Data pre-processing is an important step in authorship analysis. Text documents in their original form are not appropriate for direct analysis. So, they must be converted into a suitable input format.

Hence, punctuation marks, diacritics, numbers and non-Arabic letters are removed from the text documents. After that, each text document is formatted according to UTF8 format.

This step of text pre-processing is crucial in determining the quality of the next stages, feature extraction and classification stage.

B. Features Extraction

An important stage is a process of dataset to find distinctive features which exhibit the writing style of each an authorship individually. Assumption that every style of each author has particular features can be accessible to exploit these stylomatic features.

. As we can see from the Table 1, n-gram based approaches can operate at either word level or character level. In using such techniques, a text document or a piece of text is regarded as a sequence of n words (or n characters), where n is the number of words (or characters), in that text.

TABLE I
 CHARACTER AND LEXICAL FEATURES USED IN THIS STUDY

Feature used	Feature usage description	Feature Type
Character Bigrams	Character pairs in sequence.	Character
Character Trigrams	Groups of three successive letters.	
Character Tetra Grams	Groups of four successive letters.	
Words	Words frequencies (white space as separator).	Lexical
Word Bigrams	Word pairs in sequence.	

C. Classification methods

All In our experiments, four different classifiers are used for the automatic authorship classification (into ideally 7 different classes), where every class should represent one particular author. The different classifiers are defined as follows:

- Manhattan centroid distance;
- Multi Layer Perceptron;
- SMO based Support Vector Machines;
- Linear Regression.

D. The Fusion approach

Furthermore, in this investigation, a Fusion approach is proposed to enhance the attribution accuracy of the conventional classifiers/features.

In order to enhance the authorship attribution performance, we have proposed the use of several classifiers and several features, which are combined in order to get a lower identification error: this combination is technically called Fusion [18].

Theoretically, the fusion can be performed at different hierarchical levels and forms. A very commonly encountered taxonomy of data fusion is given by the following techniques [20, 21, 22]:

- Feature level where the feature sets of different modalities are combined. Fusion at this level provides the highest flexibility but classification problems may arise due to the large dimension of the combined (concatenated) feature vectors.

- Score (matching) level is the most common level where the fusion takes place. The scores of the classifiers are usually normalized and then they are combined in a consistent manner.

- Decision level where the outputs of the classifiers establish the decision via techniques such as majority voting. Fusion at the decision level is considered to be rigid for information integration [23], but it is not complicated in implementation.

In this investigation, we propose the use of the third technique, namely the decision level based fusion. Furthermore, two types of combinations are employed: combination of features, called **FDF** or Feature-based Decision Fusion, and combination of classifiers, called **CDF** or Classifier-based Decision Fusion.

- **Feature-based Decision Fusion (FDF):** In the first proposed fusion (combination of several features), three different features are employed: Character-tetragram; Word and Word Bigram.

The fusion technique fuses the different corresponding scores of decision into one decision (the final decision). The chosen classifier is Manhattan centroid because it has shown excellent performances during the previous experiments.

The Feature-based Decision Fusion or FDF (see Fig. 2) consists in fusing the outputs of the classifier according to a specific vote provided by the different decisions: each decision concerns one feature F_j .

The fused decision D_f of N features is given by the following equation:

$$\text{Decision} = D_f, \text{ with } f = \text{argmax}_j(\text{freq}(D_j)) \quad (1)$$

freq denotes the occurrence frequency of a specific decision and $j=1..N$.

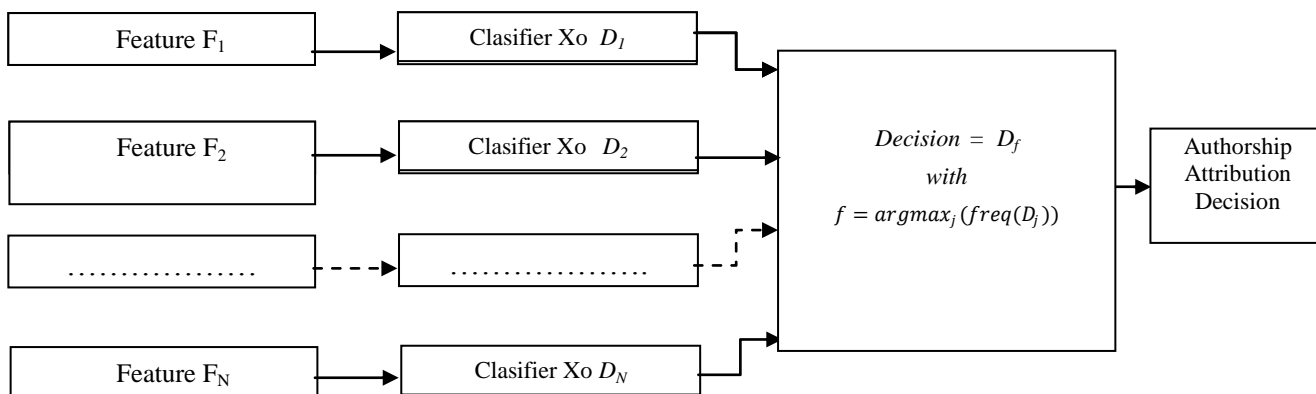


Fig. 2. Principle of the Feature-based Decision Fusion (FDF)

- **Classifier-based Decision Fusion (CDF):** In the second proposed fusion (combination of several classifiers), three different classifiers are employed:

- Manhattan centroid distance;
- SMO-SVM;
- MLP.

As previously, the fusion technique fuses the different corresponding scores of decision into one decision (the final decision). Concerning the choice of the features, the word descriptor has been used because it has been shown that this type of feature presented relatively good performances during our experiments.

It is called Classifier-based Decision Fusion or CDF (see figure 3) and consists in fusing the outputs of the different classifiers according to a specific vote provided by their different decisions: each decision concerns one classifier C_j .

The fused decision D_f of M classifiers is given by the following equation:

$$\text{Decision} = D_f, \text{ with } f = \text{argmax}_i(\text{freq}(D_i)) \quad (2)$$

freq denotes the occurrence frequency of a specific decision and $i=1..M$.

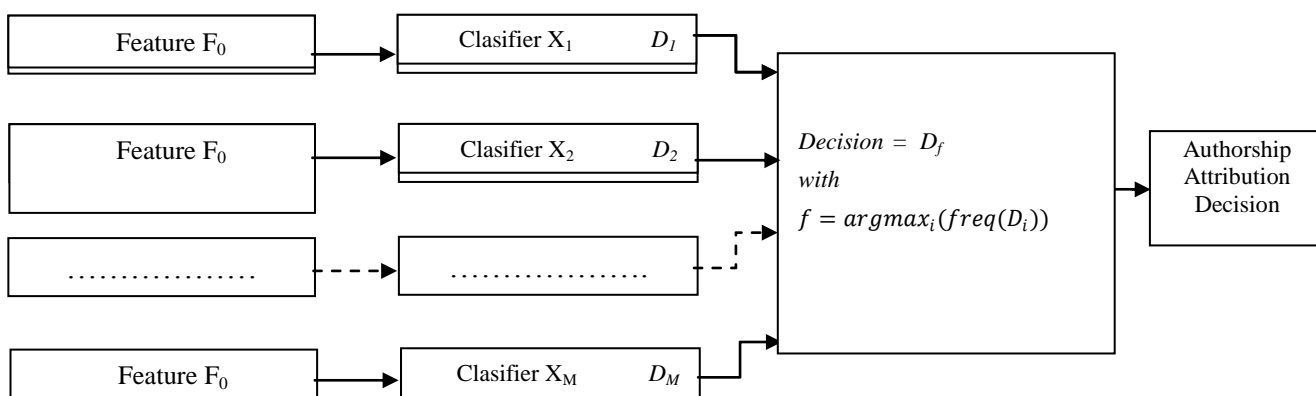


Fig. 3: Principle of the Classifier-based Decision Fusion (CDF)

V. EXPERIMENTAL RESULTS AND ANALYSIS

As mentioned previously, seven Arabic religious books are investigated and analyzed in order to make a classification of the text documents per author: the experimented corpus is called SAB-2. We also recall that several features and several classifiers are used in the experiments of authorship attribution.

Note that score of good authorship identification is calculated, in our investigation, by using the following formula:

$$\text{Score of good authorship identification} = \frac{\text{Number of correctly classified segments}}{\text{Total number of tested examples}} \quad (3)$$

A. Experiments of authorship attribution using conventional features and classifiers

Figure 4 is a graphical representation of Score of good authorship identification by feature. We can see that, generally, the Linear Regression gets the highest score (97.9%); while the MLP gets the lowest one 92.6%.

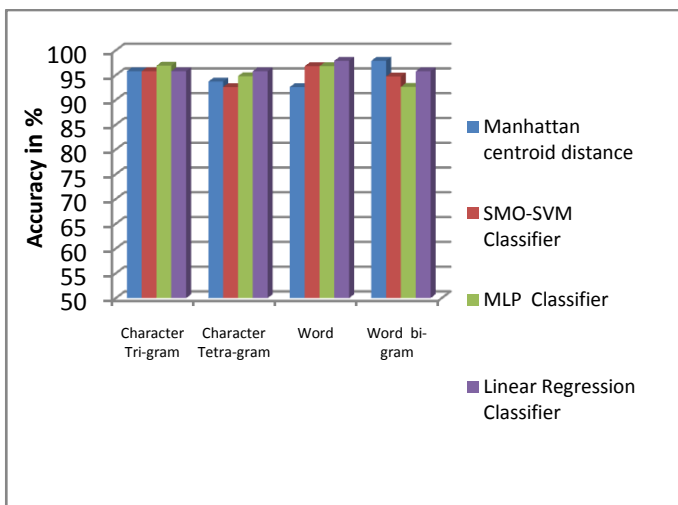


Fig.4: Score of good authorship identification by feature

Note: we notice that Manhattan centroid distance, which is a relatively simple statistical classifier, outperforms the other machine learning classifiers in many cases. However we do know that these last ones are usually better than the distance based classifiers especially for the SVM classifier, which is considered as the state-of-the-art classifier in many research fields. The main possible reason is the low dimensionality of the training dataset, which usually leads to a weak training process (note that some books are too small with only 8 or 9 texts per book: this fact makes difficult to get a big training dataset).

B. Experiments of authorship attribution using fusion techniques

In order to further enhance the authorship attribution performances, two fusion techniques have been proposed and implemented: the FDF and CDF fusion techniques. We can see in tables 2 and 3 the corresponding results of those two fusion techniques respectively.

The four authors: Aaid-Alkarni, Abdelkafy , Hassan and Alghazali presented some problems of authorship attribution depending on the choice of the classifier. Again, the two first ones are often confused with other authors.

In order to further enhance the authorship attribution performances, two fusion techniques have been proposed and implemented: the **FDF** and **CDF** fusion techniques (as explained in the previous section). In Tables 2 and 3 we can see the corresponding results of those two fusion techniques respectively.

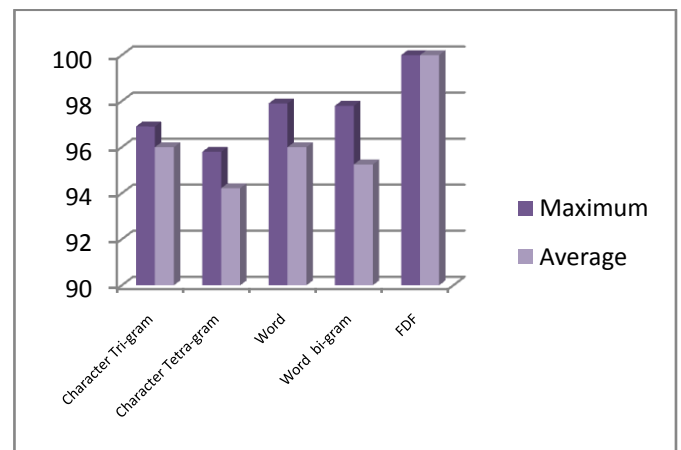


Fig .5: Comparing result (in %) of **FDF** approach with the conventional features

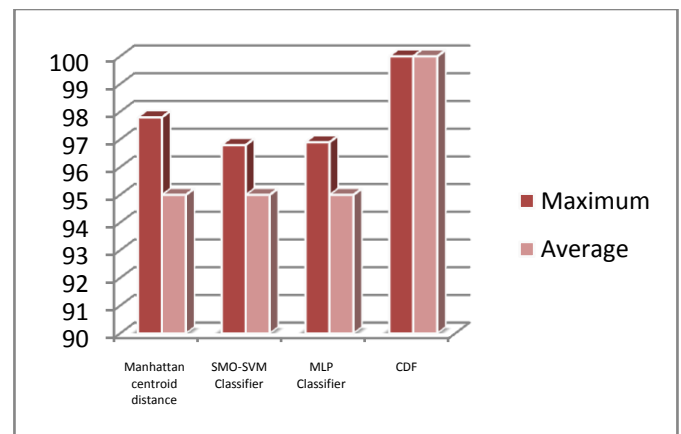


Fig. 6: Comparing result (in %) of **CDF** approach with the conventional classifiers

TABLE 2.
 ERROR OF IDENTIFICATION WITH AND WITHOUT FEATURE-BASED FUSION (FDF)

	Total Identification error on the 7 books	Hassan's book	Al-arifi's book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	A-Khaled's book
Char_trigram	4.2	0%	0%	12.5%	0%	0%	22.2%	0%
Char_tetragram	6.32%	9.09%	0%	18.75%	0%	0%	11.1%	0%
Word	7.37%	4.5%	0%	12.5%	16.7%	0%	11.11%	0%
Word bi_gram	2.1%	0%	0%	0%	16.7%	0%	11.11%	0%
FDF Fusion	0%	0%	0%	0%	0%	0%	0%	0%

TABLE 3.
 ERROR OF IDENTIFICATION USING THE CLASSIFIER-BASED FUSION (CDF)

	Total Identification error on the 7 books	Hassan's book	Al-arifi's book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	A-Khaled's book
Manhattan	7.37%	4.5%	0%	12.5%	16.7%	0%	11.11%	0%
SVM	3.16%	9.09%	0%	0%	16.7%	33.3%	0%	0%
MLP	3.16%	9.09%	0%	0%	16.7%	0%	0%	0%
CDF Fusion	0%	0%	0%	0%	0%	0%	0%	0%

The figure 5 and 6 show that the total identification score is 100%, showing the superior performances of the fusion techniques over the conventional classifiers as expected in theory. This result is very interesting since it shows that a combination of different features and/or classifiers can lead to high authorship attribution performances.

C. Comments

By observing the different experimental results, we can see that the 7 different books have been discriminated (let us say) correctly with regards to the writer/author: the corresponding text segments have been attributed to the correct authors with a

small error of identification. Moreover, by using the fusion approach the attribution error have been reduced to 0%. This important result shows that the classical features and classifiers that are usually employed in English and Greek languages got good results for the Arabic language too and appear to be utilizable for the authorship attribution of texts that are written in Arabic.

The first conclusion we can state is that the fusion approach is quite interesting in multi-classifier or multi-feature authorship attribution.

VI. CONCLUSIONS

In this research work an authorship attribution investigation has been conducted on seven Arabic religious books written by 7 religious scholars. We recall that the genre of the different books is the same and that the topic (ie. Religion) is the same too.

Hence, four different classifiers have been used for the attribution task, by using four different features as described in section 4. Moreover a two 2 fusion methods called **FDF** and **CDF** were proposed to enhance the AA performances.

Results have shown good authorship attribution performances with an overall score ranging from 92% and 98% of good attribution (depending on the features and classifiers that are employed) without the use of fusion.

However, this score reaches 100% of good attribution by using the proposed fusion techniques (**FDF** and **CDF**). This result shows that the fusion approach is interesting and should be strongly recommended for authorship attribution methods that require high degree of accuracy, such as in religious disputes or in criminal investigations.

Finally, this investigation on Arabic language shows that the fusion approach can really improve AA result if it is judiciously performed.

References

- [1] Signoriello, D.J., Jain, S., Berryman, M.J., Abbott, D.: Advanced text authorship detection methods and their application to biblical texts. In: Proceedings of SPIE (2005), vol. 6039, pp.163–175. SPIE (2005)
- [2] Eder, M.: Does size matter? Authorship attribution, short samples, big problem. In: Digital Humanities 2010 Conference, London, pp. 132–135 (2010)
- [3] Mosteller, F. and Wallace, D.L. :Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Springer.13(10) ,1-15 (1964)
- [4] Holmes, D. I. : The evolution of stylometry in humanities scholarship. Literary and linguistic computing 13(3), 111-117 (1998)
- [5] Van Halteren, H. : Linguistic profiling for author recognition and verification. Proc. of the 42nd Annual Meeting on Association for Computational Linguistics. 199 -205 (2004)
- [6] De Vel, O., Anderson, A., Corney, M. and Mohay, G. :Mining e-mail content for author identification forensics. ACM Sigmod Record. 30 (4), 55-64(2001)
- [7] Juola, P., Sofko, J. and Brennan, P.: A Prototype for Authorship Attribution Studies. Literary and Linguistic Computing. 2, 169-178 (2006)
- [8] Jain, A. : Biometric Identification. Communications of the ACM. 43, 91-98 (2000)
- [9] Stamatatos, E.: A survey of modern authorship attribution methods. Journal of American Society for information Science and Technology. 60 (3) , 238-556 (2009)
- [10] Stamatatos, E.: Author identification: Using text sampling to handle the class imbalance problem. Journal of Information Processing & Management. 44 (2), 790-799 (2008)
- [11] Juola, P.: Large-scale experiments in authorship attribution. English Studies. 93(3), 275–283 (2012)
- [12] Sayoud,H. : Investigation of Author Discrimination between two Holy Islamic Books. IET (ex-IEEE) Teknologia Journal. 1(1), X-XII (2010)
- [13] Sayoud, H. : Author Discrimination between the Holy Quran and Prophet's Statements. LLC journal, Literary and Linguistic Computing Journal, Oxford-University Press.7(4), 427-444 (2012)
- [14] Shaker,K.: Investigating Features and Techniques for Arabic Authorship Attribution. Submitted for the degree of Doctor Of Philosophy On compilation of research in the Department Of Computer Science School of Mathematics and Computer Science Heriot-Watt University, (2012)
- [15] Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., Sanchez-Perez, M. and Barron-Cedeno, A.: Overview of the author identification task at PAN 2014. Analysis, 13-31 (2014)
- [16] Kim, Y. : Convolutional neural networks for sentence classification. International Conference on Empirical Methods in Natural Language Processing (EMNLP), Qatar (October 25-29, 2014)
- [17] Sayoud,H. : A Visual Analytics based Investigation on the Authorship of the Holy Quran. 6th International Conference on Information Visualization Theory and Applications. 177-181 Berlin (March 11-14, 2015)
- [18] Seroussi,Y., Zukerman, I. and Bohnert, F. : Authorship Attribution with Topic Models. Assoc. Comput. Linguist. vol. 40, no. 2, 269–310, (2014)
- [19] Ouamour, S., Sayoud, H.: Authorship attribution of ancient texts written by ten Arabic travelers using character N-Grams. in Proceedings of International

Conference on Computer, Information and
Telecommunication Systems (CITS). 1–5 (2013)

- [20] Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Trans. Circuits Syst. Video Technol.* 14(1), 4–20 (2004)
- [21] Dasarathy, B.V.: *Decision Fusion*. IEEE Computer Society Press, Los Alamitos (1994)
- [22] Verlinde, P.: *A Contribution to Multimodal Identity Verification using Decision Fusion*. Ph.D thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, 17 September 1999
- [23] Stylianou, Y., Pantazis, Y., Calderero, F., Larroy, P., Severin, F., Schimke, S., Bonal, R., Matta, F., Valsamakis, A.: *GMM- based multimodal biometric verification*. Final Project Report 1, Enterface 2005, 18 July–12 August, Mons, Belgium (2005)