AUTHORSHIP IDENTIFICATION OF SEVEN ARABIC RELIGIOUS BOOKS -A FUSION APPROACH-

H. Sayoud[1], H. Hadjadj[2]
USTHB University, Algiers
[1]halim.sayoud@gmail.com, [2]hadjadj.has@gmail.com

Abstract— In this paper, we conduct an investigation of automatic authorship attribution on seven Arabic religious books, namely: the holy Quran, Hadith and five other books written by five 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).

The authorship characterization is based on four different features: character trigrams, character tetragrams, word unigrams and word bigrams. The task of authorship identification is ensured by four conventional classifiers: Manhattan distance, Multi-Layer Perceptron, Support Vector Machines and Linear Regression. Furthermore, a fusion approach has been proposed to enhance the performances of authorship attribution, with two fusion techniques.

The novelty of this research work lies in the following points: the proposal of a new type of fusion and the proposal of a new optimal rule dealing with unbalanced text documents.

A particular application is dedicated to the authorship discrimination between the Quran and Hadith, in order to see if the two books could have the same author or not.

Results show good authorship attribution performances with an overall score ranging from 96% and 99% of correct attribution by using the conventional classifiers. This score reaches 100% of correct attribution by using the proposed fusion techniques.

Concerning the application of discrimination, results have revealed that the Quran and Hadith books are stylistically different and should belong to two different authors.

Keywords— *Artificial Intelligence, Computational linguistics, Pattern Recognition, Authorship attribution, Fusion approach, Automatic Text classification, Author discrimination.*

I. Introduction

Stylometry or author recognition 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].

H. Sayoud, H. Hadjadj. Authorship Identification of Seven Arabic..., HDSKD journal, Vol. 6, No. 1, pp. 137-157, December 2021. ISSN 2437-069X. DOI 10.5281/zenodo.6353805.

Stylometry (or author recognition) can be divided into several research fields:

- Authorship Attribution [3], or identification, which consists in identifying the author(s) of a text;
- Authorship verification [4], which consists in checking if a text claimed to be written by somebody is really written by himself;
- Authorship discrimination [5], which consists in checking if two texts are written by the same author or not;
- Authorship Indexing [6], which consists in segmenting a multi-author text into several homogeneous segments and giving the identity of each author in those homogeneous segments;
- Plagiarism detection [7] [8], which consists in checking if a piece of text has been picked from another author.

In practice, retrieving the real author of a piece of text has raised several questions and problems for centuries. The 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 [9].

Furthermore, the recent development of improved statistical techniques in conjunction with the large availability of digital corpora, have made the automatic and objective inference of authorship a practical and easy task. That is why, this research field has seen an explosion of scholarship, resulting in several related works [11] [16] [17].

Research works on authorship attribution usually appear at 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 with a certain sense of best practices and techniques [9].

As mentioned above and concerning the different existing related works, despite the large utilization of stylometry in the occidental languages, there are not a lot of articles (relatively) related to Arabic text categorization [10], especially for religious texts.

One can find a couple of recent works of author discrimination in Arabic [11]: for instance in 2012, Sayoud presented a series of author discrimination experiments between the holy Quran and Hadith [5]. Once, the author 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 showing an experiment of author discrimination between the holy Quran and Hadith by using a hierarchical clustering. Results were interesting since they sharply showed two important clusters representing the two corresponding authors: Quran author and Hadith author.

In this investigation, we are interested in conducting a stylometric analysis on these two religious books in a larger textual corpus and with several authors. So, in order to enlarge the dataset and increase the number of authors, we have decided to use 7 different books and then 7 different

authors (*Quran, Hadith and 5 other religious books*). These experimental conditions are theoretically more consistent for the discrimination/attribution task.

Hence, we will try to make some experiments of Authorship Attribution (AA) on seven Arabic religious books, which are the holy Quran (*the divine book of God in the Islamic religion*) [12], the Hadith (*the statements of the Prophet Muhamad*) [13] and five other religious books. We note that the genre of the different books is the same and that the topic (*ie. Religion*) is the same too.

This choice/combination has several reasons: firstly, we want to check if the Quran was written by the Prophet or not; second the experiments of authorship attribution should be more consistent if we make the attribution tests with seven books instead of two books only (statistically speaking); and finally, we would like to see what would be the best features and classifiers for Arabic text classification.

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). In the knowledge of the author, it is the first time that it has been applied in stylometry with the proposed forms (*i.e. FC and FF*).

Concerning the structure of our manuscript, it is organized as follows:

- Section 2 gives an overall description of the corpus and the different investigated books.
- Section 3 describes the different proposed method of AA.
- Section 4 describes the different experiments of authorship attribution.
- In section 5, a further investigation of authorship discrimination between the Quran and Hadith is presented.
- Finally, a general conclusion is given at the end of the manuscript with some useful references.

II. Corpus of the Seven Religious Books

As cited previously, there are seven different books written by seven different authors: the holy Quran, Hadith and 5 other books written by 5 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-1* (*Seven Arabic Books – dataset One*). These books are described as follows:

-1st book: the holy Quran (*author: God (Allah)*), it is considered as the divine book of Islam [12]. The Quran is considered to be written by Allah (God) and only sent down to the Prophet Muhammad fourteen centuries ago. This divine book has been delicately conserved by the different scholars over the time. The holy Quran is considered as the first reference of Islam since it is supposed to contain the authentic statements of God (Allah);

Fig. 1: Old pages of the *holy Quran*

-2nd book: the Hadith (*author: the Prophet Muhammad*) contains the statements of the Prophet Muhammad in different situations [13]. Muhammad was born in Mecca in the 6th century, became Prophet at the age of 40 and died at the age of 63. In this investigation, we used the Bukhari Hadith book, which is considered one of the most trusted compilation of the Hadith;

Fig. 2: Old pages of the *Hadith*

-The 5 other books: represent books and texts collections written by 5 religious scholar, namely: *Mohammed al-Ghazali al-Saqqa* [14], *Yusuf al-Qaradawi* [15], *Omar Abdelkafy* [16], *Aaidh ibn Abdullah al-Qarni* [17] and Amr Mohamed Helmi Khaled [18].

Those seven books are preprocessed and segmented into different and distinct text segments. Every segment is about 2900 tokens each. Here are the numbers of segments by book:

Table 1: Books specifications of *SAB-1* dataset.

Book/Author	Number of segments by book*	Big/ Small parameter#	Training set size	Testing set size
1st book: the holy Quran	30 segments	Big	7	23
2nd book: the Hadith	8 segments	Small	4	4
3rd book: books of Alghazali	39 segments	Big	7	32
4th book: books of AlQuaradhawi	13 segments	Small	4	9
5th book: books of Abdelkafy	10 segments	Small	4	6
6th book: books of Aid Alkarny	23 segments	Big	7	16
7th book: books of Amrokhaled	9 segments	Small	4	5

**Each segment is composed of 2900 tokens.*

#Big/Small is a logical parameter (i.e. binary value).

The corpus is decomposed into 2 parts: training part and testing part, and since the different books have different sizes, a balancing rule has been established: 4 text segments are used for the training of small books and 7 text segments are employed for the training of big books. The main reasons for this choice are explained here below.

The choice of the training dataset size is defined by a particular logical (*binary*) parameter we called Big/Small, which gives a qualitative estimation on the size of the book. That is, if the size of the book is over 20 segments, then it is considered as a big dataset otherwise it is considered small. The value or the threshold 20 is equal to the half size of the biggest dataset (*ie. 39 segments for Alghazali book, which implies a threshold of 39/2 ≅20*). This scheme permits us to have different possible sizes for the training dataset.

By observing the small books, we notice that *"4 text segments"* should be a good choice for the small books. In fact, the value 4 is equal to the half size (*50%*) of the smallest book (*ie. the smallest book contains only 8 segments*).

By observing the seven books, we notice that *"7 text segments"* should be a good choice too for the big books. In fact, the value 7 is equal to the maximum size of the training set for the smallest book (*ie. a maximum of 7 segments for the training, since we require at least 1 segment for the testing*).

These two training rules could be applied to the different books with regards to the parameter Big/ Small. But even though, the value 7 is a limit that we cannot exceed (and could be seen as a fixed choice), we cannot say that the value 4 is optimal for small texts: why not 3 or 5 text segments, for instance.

In order to check if this choice was judicious or not, (*experimentally speaking*), we did some experiments of authorship attribution on another corpus consisting of 7 different books (*from a second different dataset called SAB-2*) denoted by A, B, C, … G and where the sizes of the books are very similar to those of the previous one: SAB-1 dataset (*see table 2*). The second dataset was split into two subsets almost randomly: some texts were selected for the training and others were

selected for the testing. The used classification technique was based on the Manhattan Centroid distance.

Table 2: Features of the second dataset: *SAB-2**

		Case 1	Case 2	Case 3
Book designation	Big/ Small dataset	Training set size	Training set size	Training set size
Book A	Big	7	7	7
book B	Small	3	4	5
book C	Big	7	7	7
book D	Small	3	4	5
book E	Small	3	4	5
book F	Big	7	7	7
book G	Small	3	4	5

** Note that the corpus SAB-2 will no longer be utilized in the next sections.*

Hence, three cases are investigated:

- Case 1: 3 text segments are used for the training of small books and 7 text segments are employed for the training of big books;
- Case 2: 4 text segments are used for the training of small books and 7 text segments are employed for the training of big books;
- Case 3: 5 text segments are used for the training of small books and 7 text segments are employed for the training of big books.

The different results of authorship attribution, got on this second dataset, are summarized in the following table:

Table 3: Results of authorship attribution got on the second dataset: ***SAB-2****

			Score of correct attribution in % (experiments conducted on another corpus)							
Training size			Char. Bi-gram	Char. Tri-gram	Char. Tetra-gram	Word	Word Bi-gram	Word tri-gram	Word Tetra-gram	Average performance in %
Case	Big	Small								
Case 1	7	3	74.74	83.83	89.89	94.94	94.94	32.32	33.33	63.88
Case 2	7	4	76.84	89.47	91.57	93.68	97.89	54.73	31.57	69.47
Case 3	7	5	76.92	85.71	89.01	95.6	97.8	35.16	32.96	65.38

** Note that the corpus SAB-2 will no longer be utilized in the next sections.*

According to table 3, case 2 (corresponding to 4 training texts for the small books) seems to be the most interesting case. That is, by observing the average performance given by Manhattan distance, we can easily see that the best average score is 69.47%, which corresponds to the second case (*ie. 4 text segments for small books and 7 ones for big books*). According to this result, the chosen training configuration seems to be judicious and interesting for the authorship attribution experiments conducted on the first dataset.

However, we should note that we cannot expand this result to other classifiers like machine learning ones, especially those which need a great amount of training data, such as neural networks or support vector machines, for instance.

III. 3. Authorship Attribution Methods

Several experiments of authorship attribution are conducted on the 7 segmented religious books.

For a purpose of feature selection and evaluation [19], four types of characteristics are employed: character-trigram, character tetra-gram, word and word-bigram. Two of these features are based on characters and the two others are typically lexical.

Also, 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.

Furthermore, a Fusion approach is proposed to try enhancing the attribution accuracy of the conventional classifiers/features.

III.1 Conventional Classifiers

The 4 conventional classifiers are described here below.

- Manhattan distance

This distance [5] is very reliable in text classification. The corresponding distance between two vectors *X* and *Y* is given by the following formula:

$$d_{X,Y} = \sum_{i=1}^{n} |X_i - Y_i| \quad (1)$$

where n is the length of the vector.

In this investigation, the different samples of the training are employed to build the centroid vector, which will be used, as reference, to compute the required distance with the previous formula (*also called KNN method*). Manhattan distance is simple to implement and very efficient for text classification.

- Multi-Layer Perceptron (MLP)

The MLP (*Multi-Layer Perceptron*) is a classical neural network classifier that uses the errors of the output to train the network weights [20]. The MLP can use different back-propagation schemes to ensure the training of the classifier. It is trained by the different texts of the training set, whereas the remaining texts are used for the testing task. Usually the MLP is efficient in supervised classification, however some bad training cases could be observed with local minima, which may lead to some classification errors.

- Sequential Minimal Optimization based Support Vector Machine (SMO-SVM)

In machine learning, support vector machines (*SVMs*) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, which 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 margin 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, SVMs 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 [21]. Concerning the Sequential Minimal Optimization (*SMO*) algorithm, it is used to speed up the training of the SVM [22].

- Linear Regression

Linear Regression is the oldest and most widely used predictive model. The method of minimizing the sum of the squared errors to fit a straight line to a set of data points was published by Legendre in 1805 and by Gauss in 1809. 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 [23] [24].

III.2 The Fusion approach

In order to enhance the authorship attribution performance, we have proposed the use of several classifiers, which are combined in order to get a lower identification error: this combination is

technically called Fusion [25]. We have proposed two types of fusion: a Feature-based Decision Fusion and a Classifier-based Decision Fusion.

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 [26] [27] [28]:

- 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 [29], but it is not complicated in implementation.

In this investigation, we propose the use of the third technique, namely the decision level based fusion. As mentioned previously, 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;
- Word Bigram.

Those three features were chosen because of their good performances on SAB2 (three best accuracies of the experiment) as reported in table 3.

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 previous works. The FDF Fusion consists in fusing the outputs of the different classifiers according to a specific vote provided by their different decisions: each decision concerns one feature *Fj* (*see figure 3*).

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

$$Decision = D_f, \ with \ f = argmax_j(freq(D_j)) \quad (2)$$

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

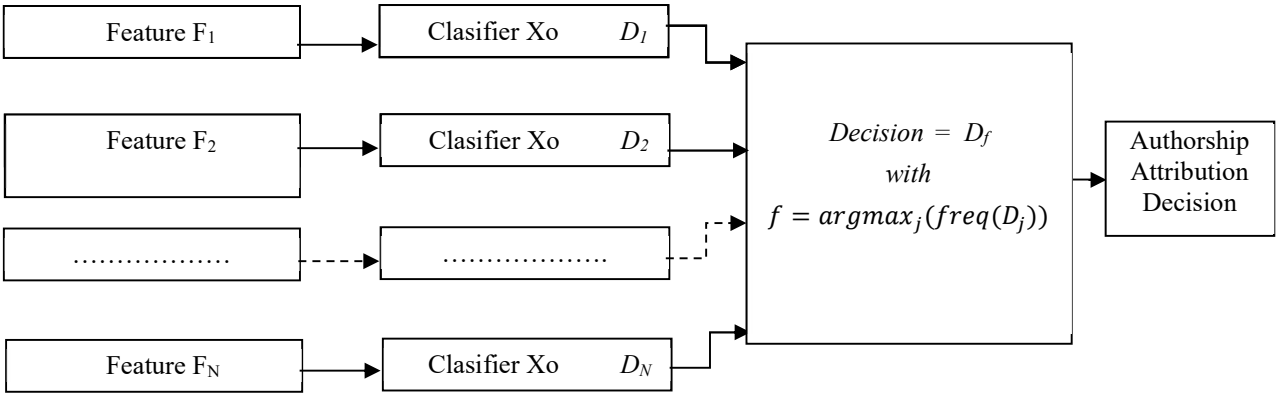

Fig. 3: Principle of the Feature-based Decision Fusion (FDF)

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

- Manhattan centroid;
- 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 (see table 3).

The CDF Fusion consists in fusing the outputs of the different classifiers according to a specific vote provided by their different decisions: each decision concerns one classifier *Cj* (*see figure 4*).

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

$$Decision = D_f, \ with \ f = argmax_i(freq(D_i)) \quad (3)$$

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

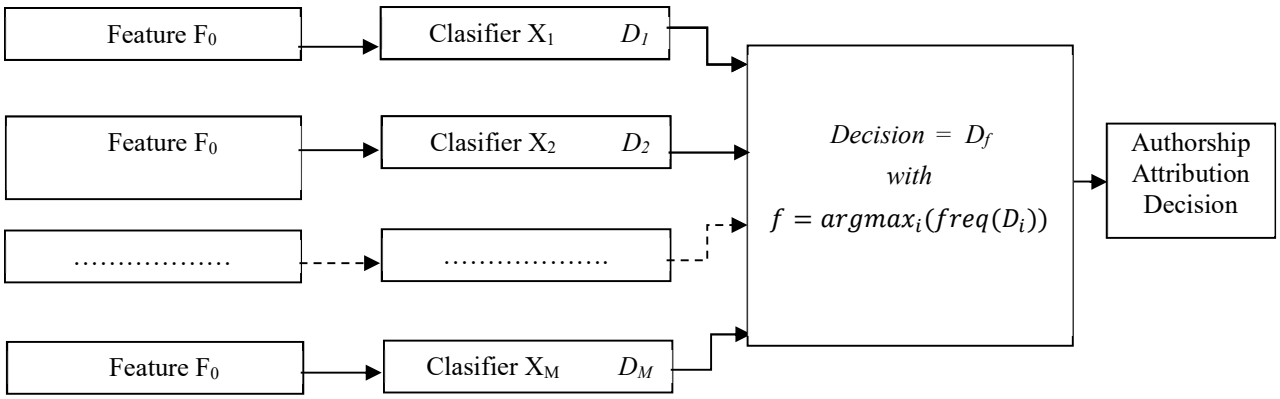

Fig. 4: Principle of the Classifier-based Decision Fusion (CDF)

All the results of the fusion approach are represented in tables 8 and 9, summarizing the corresponding AA scores of the first and second fusion techniques respectively.

IV. Experiments of Authorship Attribution

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-1***. We also recall that several features and several classifiers are used in the experiments of authorship attribution, by using the JGAAP-5.2 tool.

IV.1 Experiments of authorship attribution using conventional features and classifiers

In this section we report the different results obtained by using conventional classifiers and features. The different experimental results are organized into 4 tables (*table 4, 5, 6 and 7*):

- Table 4 displays the different results obtained with the Character-trigram feature;
- Table 5 displays the different results obtained with the Character-tetragram feature;
- Table 6 displays the different results obtained with the Word (*Word-unigram*) feature;
- Table 7 displays the different results obtained with the Word-bigram feature**.**

The corresponding tables (*table 4, 5, 6 and 7*) display the errors of authorship attribution given by the 4 classifiers: Manhattan centroid, MLP, SMO-SVM and Linear Regression. Furthermore, a column untitled “Total identification error” summarizes the overall error of attribution for the 7 books. This indication gives us an interesting idea on the overall performances of authorship attribution (*corresponding to a specific feature*).

Table 4: Identification Error in % using the feature: *Character-trigram*, on *SAB-1* dataset.

		Total Identification error on the 7 books	The holy Quran book	The Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
Date / Century			Ancient: 6th century	Ancient: 6th century	Recent: 20th century	Recent: 20th century	Recent: 20th century	Recent: 20th century	Recent: 20th century
Classifier	Manhattan centroid distance	4.2%	0%	0%	12.5%	0%	0%	22.2%	0%
	MLP classifier	3.1%	0%	0%	0%	16.7%	0%	22.2%	0%
	SMO-SVM classifier	4.2%	0%	0%	0%	33.3%	0%	22.2%	0%
	Linear Regression	4.2%	0%	0%	6.25%	16.7%	0%	22.2%	0%

In table 4, one can notice that, with Character-trigrams, the best classifier is the MLP, which gives an error of only 3.1% (look at the 1st columbn), the other classifiers have the same performances (*total identification errors of 4.2%*). The two authors: Abdelkafy and Alquaradawi present some problems of authorship attribution, with respectively 16.7% and 22.2.% in the case of the MLP. These two authors are often confused with other authors. Note that the Quran and Hadith books are attributed without any error (*error of 0%*).

Table 5: Identification Error in % using the feature: *Character-tetragram*, on *SAB-1* dataset.

		Total Identification error on the 7 books	The holy Quran book	The Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
Date / Century			Ancient 6th century	Ancient 6th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century
Classifier	Manhattan centroid distance	1.05%	0%	0%	0%	0%	0%	11.1%	0%
	MLP classifier*	2.1%	0%	0%	6.25%	16.7%	0%	0%	0%
	SMO-SVM classifier*	3.1%	0%	0%	12.5%	16.7%	0%	0%	0%
	Linear Regression*	2.1%	0%	0%	6.25%	16.7%	0%	0%	0%

**: 500 most frequent features only.*

In table 5, we can see that the best classifier is Manhattan distance, which gives an error of only 1.05%, the other classifiers present different performances (*total identification errors ranging between 2.1% and 3.1*). The three authors: Aaid-Alkarni, Abdelkafy and Alquaradawi present some problems of authorship attribution depending on the choice of the classifier. These two first ones are often confused with other authors. As previously, we can note that the Quran and Hadith books are attributed without any error (*error of 0%*).

Table 6: Identification Error in % using the feature: *Word*, on *SAB-1* dataset.

		Total Identification error on the 7 books	The holy Quran book	The Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
Date / Century			Ancient 6th century	Ancient 6th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century
Classifier	Manhattan centroid Distance	1.05%	0%	0%	6.25%	0%	0%	0%	0%
	MLP classifier*	1.05%	0%	0%	0%	16.7%	0%	0%	0%
	SMO-SVM classifier*	2.1%	0%	0%	0%	0%	33.3%	0%	0%
	Linear Regression*	2.1%	0%	0%	6.25%	16.7%	0%	0%	0%

**: 500 most frequent features only.*

In table 6, we can see that the best classifiers are the MLP and Manhattan distance, which give an error of only 1.05%, the other classifiers present the same performances (*total identification errors of 2.1%*). The two authors: Aaid-Alkarni and Abdelkafy present some problems of authorship attribution depending on the choice of the classifier. These two particular authors are often confused with other authors.

Table 7: Identification Error in % using the feature: *Word Bigram*, on *SAB-1* dataset.

		Total Identification error on the 7 books	The holy Quran book	The Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
Date / Century			Ancient 6th century	Ancient 6th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century	Recent 20th century
Classifier	Manhattan centroid distance	1.05%	0%	0%	0%	0%	3.1%	0%	0%
	SMO-SVM classifier#	3.1%	0%	0%	12.5%	16.7%	0%	0%	0%
	MLP classifier#	4.2%	0%	0%	12.5%	33.3%	0%	0%	0%
	Linear Regression#	4.2%	0%	0%	12.5%	16.7%	0%	0%	20%

#: 600 most frequent features only.

In table 7, we can see that the best classifier is Manhattan distance, which gives an error of only 1.05%, the other classifiers present different performances (*total identification errors ranging between 3.1% and 4.2%*). The three authors: Aaid-Alkarni, Abdelkafy and Alghazali present some problems of authorship attribution depending on the choice of the classifier. Again, these two first ones are often confused with other authors. Once again, we can note that the Quran and Hadith books are attributed without any error (see the confusion matrices in tables 8.a to 8.d).

Although there is not a great difference between the 7 books vocabularies since they all talk about religion, it is possible that some linguistic features cause a discrimination between two groups of books: the ancient books (Quran and Hadith), which belong to the 7th century, and the contemporary books, which belong to the 20th or 21th centuries.

In tables 8.a, 8.b, 8.c and 8.d, representing the confusion matrix, one can see the different cases of misclassification for the first classifier.

Table 8.a: Confusion matrix for Manhattan distance with character-trigrams (Number of attributions per author).

↗	Quran's Author	Hadith's Author	Aaid	Abdelkafy	Alghazali	Alquaradawi	Amrokhaled
Quran book	23	0	0	0	0	0	0
Hadith book	0	4	0	0	0	0	0
Aaid's books	0	0	14	0	0	2	0
Abdelkafy's books	0	0	0	6	0	0	0
Alghazali's books	0	0	0	0	32	0	0
Alquaradawi's books	0	0	0	2	0	7	0
Amrokhaled's books	0	0	0	0	0	0	5

Table 8.b: Confusion matrix for Manhattan distance with character-tetragrams (Number of attributions per author).

↗	Quran's Author	Hadith's Author	Aaid	Abdelkafy	Alghazali	Alquaradawi	Amrokhaled
Quran book	23	0	0	0	0	0	0
Hadith book	0	4	0	0	0	0	0
Aaid's books	0	0	16	0	0	0	0
Abdelkafy's books	0	0	0	6	0	0	0
Alghazali's books	0	0	0	0	32	0	0
Alquaradawi's books	0	0	0	1	0	8	0
Amrokhaled's books	0	0	0	0	0	0	5

Table 8.c: Confusion matrix for Manhattan distance with words
(Number of attributions per author).

↗	Quran's Author	Hadith's Author	Aaid	Abdelkafy	Alghazali	Alquaradawi	Amrokhaled
Quran book	23	0	0	0	0	0	0
Hadith book	0	4	0	0	0	0	0
Aaid's books	0	0	15	0	0	1	0
Abdelkafy's books	0	0	0	6	0	0	0
Alghazali's books	0	0	0	0	32	0	0
Alquaradawi's books	0	0	0	0	0	9	0
Amrokhaled's books	0	0	0	0	0	0	5

Table 8.d: Confusion matrix for Manhattan distance with word-bigrams
(Number of attributions per author).

↗	Quran's Author	Hadith's Author	Aaid	Abdelkafy	Alghazali	Alquaradawi	Amrokhaled
Quran book	23	0	0	0	0	0	0
Hadith book	0	4	0	0	0	0	0
Aaid's books	0	0	16	0	0	0	0
Abdelkafy's books	0	0	0	6	0	0	0
Alghazali's books	0	0	0	0	31	1	0
Alquaradawi's books	0	0	0	0	0	9	0
Amrokhaled's books	0	0	0	0	0	0	5

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*).

IV.2 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 9 and 10 the corresponding results of those two fusion techniques respectively.

Table 9: Error of identification using the *feature-based fusion (FDF)*

Total Identification error on the 7 books	Holy Quran book	Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
0%	0%	0%	0%	0%	0%	0%	0%

Table 10: Error of identification using the *classifier-based fusion (CDF)*

Total Identification error on the 7 books	Holy Quran book	The Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquaradawi's book	Amro-Khaled's book
0%	0%	0%	0%	0%	0%	0%	0%

As we can see in tables 9 and 10, the authorship attribution error is equal to zero for every author. 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.

IV.3 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.

Another important conclusion, one can deduce, is that the two religious books Quran and Hadith appear to have two different Authors.

V. Application of Author Discrimination between the Quran and Hadith

V.1 Purpose of this discrimination

In this section, we will try to respond to the following question: Was the Quran written by the prophet? In fact, it is well known that Muhammad was only the narrator who recited the sentences of the Quran as written by Allah (*God*), but not the author.

Certain doubts about the origins of the Quran tried to find a human source for this book continuously. Such suppositions say that the Quran could be an invention of the prophet Muhammad [30].

So, the purpose of this section is to analyze some of the experiments presented in this paper, in order to see whether the two concerned books could statistically belong to the same author or not: i.e. authorship discrimination task [31] [32] [33]. Furthermore, a Leave-One-Out (LOO) technique is used to make the discrimination more significant.

V.2 Experiments of author discrimination using LOO and LTO techniques

In this experiment, there are 37 text segments, where 29 segments are taken from the holy Quran and 8 are taken from the Hadith. We used the feature character-tetragram by keeping only the 500 most frequent features, and employed the SMO-based SVM classifier.

Since there are 37 samples, the LOO technique will lead to 37 experiments of rotating classification, where in every experiment one sample is removed and put in testing set, in order to be identified through the remaining samples that represent the training model.

In the LOO technique, with 37 different combinations, every combination/experiment got a score of 100% of correct attribution. Similarly, for the LTO technique, with 19 different combinations, every combination/experiment got a score of 100% of correct attribution too.

For simplification, it could be more interesting to compute the average accuracy, corresponding to the overall performances of classification. This entity can be evaluated by using equation 4.

$$\textbf{Average Accuracy} = \frac{\sum_{i=1}^{N} CrossVal_i}{N} \quad (4)$$

where N represents the number of cross-validation experiments (denoted by CrossVal).

According to the previous results, the average accuracy of all LOO experiments is equal to 100%, and, the average accuracy of all LTO experiments is equal to 100%.

Since there are no cross-errors of attribution between the Quran and Hadith texts (*LOO accuracy of 100%*) and according to the previous section, we can state that these 2 books are completely different in style each other and also different from all the other investigated books. Consequently, it appears that the Quran and Hadith should have two different authors or at least two different author styles.

VI. Discussion and Conclusion

As described in this paper, several experiments of authorship attribution have been conducted on seven Arabic religious books, namely: the holy Quran, Hadith and 5 other books written by 5 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*).

To conduct these experiments, several features have been proposed: character tri-grams, character tetra-grams, word uni-grams and word bi-grams. On the other hand, several classifiers have also been employed: Manhattan distance, Multi-Layer Perceptron, Support Vector Machines and Linear Regression. Furthermore, we have proposed and implemented 2 fusion methods called FDF and CDF to enhance the AA performances.

Results have shown good authorship attribution performances with an overall score ranging from 96% and 99% of correct attribution (*depending on the features and classifiers that are employed*) without the use of fusion.

However, this score reaches 100% of correct 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.

The second part of this research work was dedicated to the authorship discrimination between the Quran and Hadith books, to check whether these two books could be written by the same author or not. It presents several novelties compared to previous works [5] in the same topic, such as the proposal of a new optimal rule dealing with unbalanced text documents, the use of LOO and LTO cross-validation techniques, the Quran and Hadith segmentation (more segments and shorter texts), etc.

The related results (*of this second part*) have shown that the Quran texts and Hadith texts, used in this survey, are different with a cross-validation discrimination accuracy of 100%, and should then belong to two different authors or at least two different styles. This result confirms completely what has already been found in the works referenced in [5] and [34]. Finally, the

present research work is a computational linguistics investigation and not a religious one, but could help (*in the opinion of the author*) shedding a little of light on certain historical questions.

References

1. D. J. Signoriello, S. Jain, M. J. Berryman, D. Abbott, "Advanced text authorship detection methods and their application to biblical texts," Proceedings of SPIE (2005), Volume: 6039, Publisher: Spie, pp. 163–175, 2005.
2. M. Eder, "Does size matter? : autorship attribution, short samples, big problem," In Digital humanities 2010 conference, pp.132-135, London, 2010.
3. Sarwar, R., Urailertprasert, N., Vannaboot, N., Yu, C., Rakthanmanon, T., Chuangsuwanich, E., & Nutanong, S. (2020). $ CAG $: Stylometric Authorship Attribution of Multi-Author Documents Using a Co-Authorship Graph. IEEE Access, 8, 18374-18393, 2020.
4. Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., ... & Stein, B. (2020). Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In CLEF, 2020.
5. H. Sayoud, Author Discrimination between the Holy Quran and Prophet's Statements. LLC journal, Literary and Linguistic Compting, pp 427-444, Vol. 27, No. 4, 2012.
6. Zangerle, E., Mayerl, M., Specht, G., Potthast, M., & Stein, B. (2020). Overview of the style change detection task at PAN 2020. CLEF, 2020.
7. Muangprathub, J., Kajornkasirat, S., & Wanichsombat, A. (2021). Document Plagiarism Detection Using a New Concept Similarity in Formal Concept Analysis. Journal of Applied Mathematics, 2021.
8. Zouhir, A., El Ayachi, R., & Biniz, M. (2021). A comparative Plagiarism Detection System methods between sentences. In Journal of Physics: Conference Series (Vol. 1743, No. 1, p. 012041). IOP Publishing, 2021.
9. P. Juola. Foundations and Trends in Information Retrieval. Vol. 1, No. 3 (2006) 233–334. DOI: 10.1561/1500000005. Now Publishing, USA.
10. L. Fodil, S. Ouamour, H. Sayoud, "Theme Classification of Arabic Text: A Statistical Approach", TKE'2014 conference : Terminology and Knowledge Engineering, 19-21 June 2014, Berlin, Germany.
11. R. Baraka, S. Salem, M. Abu-Hussien, N. Nayef, W. Abu-Shaban, "Arabic Text Author Identification Using Support Vector Machines". Journal of Advanced Computer Science and Technology Research, Vol.4 No.1, March 2014, pp 1-11.
12. I. A. Ibrahim. A brief illustrated guide to understanding Islam. Library of Congress, Catalog Card Number: 97-67654, Published by Darussalam, Publishers and Distributors, Houston, Texas, USA. Web version: http://www.islam-guide.com/contents-wide.htm, ISBN: 9960-34-011-2.
13. A. A. Islahi, 1989. Fundamentals of Hadith Interpretation – an English translat. of "Mabadi Tadabbur-i-Hadith" by T. M. Hashmi. Lahore: Al-Mawrid. www.monthly-renaissance.com/DownloadContainer.aspx?id=71.
14. Mohammed al-Ghazali, From Wikipedia, the free encyclopedia (website). Last visit in 2021. http://en.wikipedia.org/wiki/Mohammed_al-Ghazali
15. Alqaradawi. Last visit in 2013. https://www.al-qaradawi.net/. Last visit in 2021.
16. Abdelkafy. Last visit in 2013. https://en.wikipedia.org/wiki/Omar_Abd_al-Kafi. Last visit in 2021.
17. Aaidh ibn Abdullah al-Qarni. From Wikipedia, the free encyclopedia (website). http://en.wikipedia.org/wiki/Aaidh_ibn_Abdullah_al-Qarni. Last visit in 2021.
18. Amr-Khaled. Last visit in 2013. http://amrkhaled.net. Last visit in 2021.
19. B. Hawashin, A. Mansour, S. Aljawarneh, "An Efficient Feature Selection Method for Arabic Text Classification", International Journal of Computer Applications 2013, IJCA Journal published by foundation of somputer Science, New York, USA, Volume 83 - Number 17, pp 1-6 .

20. H. Sayoud, "Automatic speaker recognition – Connexionnist approach", PhD thesis, USTHB University, Algiers, 2003.
21. I.H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations," In Nikola Kasabov and Kitty Ko, editors, Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems, pp. 192-196, Dunedin, New Zealand, 1999.
22. S.S. Keerthi, S.K. Shevade, C. Bhattacharyya & K.R.K, "Murthy, Improvements to Platt's SMO Algorithm for SVM Classifier Design", Neural Computation 13, pp. 637–649, 2001.
23. Bibas, K., Fogel, Y., & Feder, M. (2019, July). A new look at an old problem: A universal learning approach to linear regression. In 2019 IEEE International Symposium on Information Theory (ISIT) (pp. 2304-2308). IEEE, 2019.
24. Ying, G. S., Maguire, M. G., Glynn, R., & Rosner, B. (2017). Tutorial on biostatistics: linear regression analysis of continuous correlated eye data. Ophthalmic epidemiology, 24(2), 130-140, 2017.
25. Liu, Z., Pan, Q., Dezert, J., Han, J. W., & He, Y. (2017). Classifier fusion with contextual reliability evaluation. IEEE transactions on cybernetics, 48(5), 1605-1618, 2017.
26. A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," IEEE Transactions on Circuits and Systems for Video Technology, Volume 14, number 1, January, 2004.
27. B.V. Dasarathy (1994), "Decision Fusion". IEEE Computer Society Press, Los Alamitos, CA, 1994.
28. P. Verlinde, "A Contribution to Multimodal Identity Verification using Decision Fusion," PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, September 17th 1999.
29. Y. Stylianou, Y. Pantazis, F. Calderero, P. Larroy, F. Severin, S. Schimke, R. Bonal, F. Matta, , and A. Valsamakis,"GMM- Based Multimodal Biometric Verification," Final Project Report 1, Enterface'05, July 18th - August 12th, Mons, Belgium, 2005.
30. A. Al-Shreef. Is the Holy Quran Muhammad's invention? (website). Last visit in 2013. http://www.quran-m.com/firas/en1/index.php?option=com_content&view=article&id=294:is-the-holy-quran-muhammads-invention-&catid=51:prophetical&Itemid=105
31. D. E. Mills. Authorship Attribution Applied to the Bible. Master thesis, Graduate Faculty of Texas, Tech University, 2003. Linguistic. Computing Journal, Oxford-University Press. Published in 2012. Citation reference: doi: 10.1093/llc/fqs014, Volume 27, No. 4, 2012, pp 427-444.
32. Khondaker, M. T. I., Khan, J. Y., Alam, T., & Rahman, M. S. (2020). Agree-to-Disagree (A2D): A Deep Learning-Based Framework for Authorship Discrimination Task in Corpus-Specificity Free Manner. IEEE Access, 8, 162322-162334, 2020.
33. Sayoud, H. (2018, June). Visual Analytics Based Authorship Discrimination Using Gaussian Mixture Models and Self Organising Maps: Application on Quran and Hadith. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 158-164), Montreal. Springer, Cham., 2018.
34. H. Sayoud, Segmental analysis based authorship discrimination between the holy quran and prophet's statements. Digital Studies journal, Canada, 2015. http://www.digitalstudies.org/ojs /index.php/digital_studies.