Fusion Based Authorship Attribution-Application of Comparison Between the Quran and Hadith

Halim Sayoud(✉) and Hassina Hadjadj

USTHB University, Algiers, Algeria
halim.sayoud@uni.de, 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, by using two fusion techniques. The Arabic dialect is the same (i.e. Standard Arabic) for the seven books. The genre is the same and the topic of the different books is 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, we propose two fusion approaches to strengthen the classification performances. Finally, 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 have shown the importance of the fusion techniques in authorship attribution and confirm that the two books (Quran and Hadith) should belong to two different Authors, which implies that the Quran could not be written by the Prophet.

Keywords: Computational linguistics · Fusion approach · Authorship attribution
Automatic text classification · Author discrimination · Quran authorship

1 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].

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

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].

A. Lachkar et al. (Eds.): ICALP 2017, CCIS 782, pp. 191–200, 2018.
https://doi.org/10.1007/978-3-319-73500-9_14

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 [5, 6].

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 [4].

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 [7], especially for religious texts.

One can find a couple of recent works of author discrimination in Arabic [8], but very few ones are applied on the Quran: in 2012 for instance, Sayoud presented a series of author discrimination experiments between the holy Quran and Hadith [9]. 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 [10]. 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.

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*).

2 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:

1stbook: the holy Quran, it is considered as the divine book of Islam [11]. The Quran is considered to be written by Allah (God) and only sent down to the Prophet Muhammad fourteen centuries ago (Fig. 1).

2nd book: the Hadith contains the authentic statements and speeches of the Prophet Muhammad in different situations [12]. In this investigation we used the Bukhari Hadith. Moreover, we removed the Quranic verses present in the Hadith to get only pure statements of the Prophet (Fig. 2).

Fig. 1. Old page of the *holy Quran.*

Fig. 2. Old page of the *Hadith.*

3rd book: text collection of Alghazali (*Author: Mohammed al-Ghazali al-Saqqa*): 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.

4th book: text collection of Alquaradawi (*Author: Yusuf al-Qaradawi*): it contains some articles and dissertations of Alquaradawi. 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*).

5thbook: 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.

6thbook: 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.

7thbook: 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".

Those seven books are preprocessed and segmented into different and distinct text segments, and every segment is about 2900 tokens each.

3 Authorship Attribution Methods

Several experiments of Authorship Attribution (AA) are conducted on the 7 segmented religious books. For a purpose of feature selection and evaluation, 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 [9]; Multi Layer Perceptron NN [13]; SMO based Support Vector Machines [14, 15] and Linear Regression [16, 17].

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 [19, 20, 21]:

- 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 [22], 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 *Manhatan centroid* because it has shown excellent performances during the previous experiments.

The Feature-based Decision Fusion or FDF (*see* Fig. 3) consists in fusing the outputs of the classifier according to a specific vote provided by the different decisions: each decision concerns one feature *Fj*.

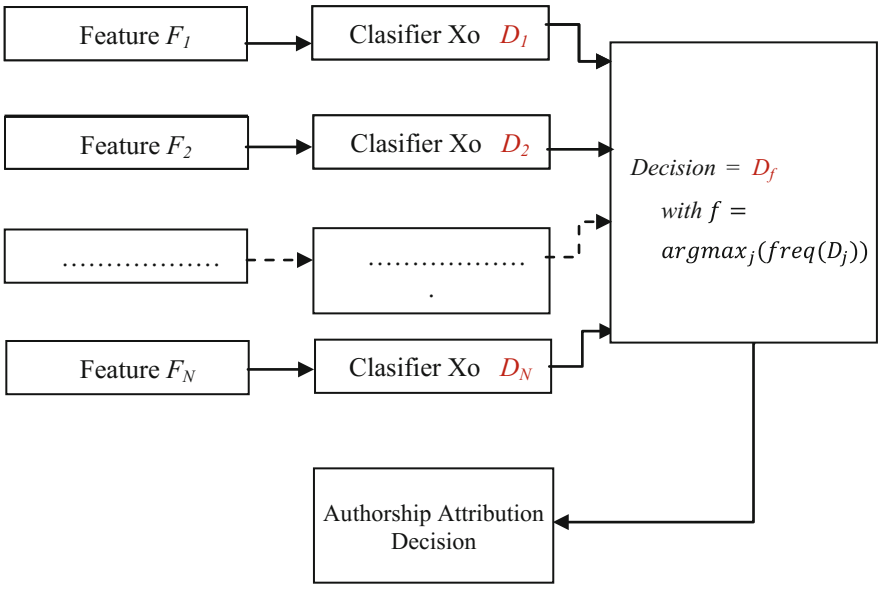

Fig. 3. Principle of the Feature-based Decision Fusion (FDF)

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

$$Decision = D_f, \; with f = argmax_j(freq(D_j)) \tag{1}$$

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

– **Classifier-based Decision Fusion (CDF)**: In the second proposed fusion (*combination of several classifiers*), three different classifiers are employed: Manhattan centroid; SMO-SVM and 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.

The Classifier-based Decision Fusion or CDF (*see* Fig. 4) 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*.

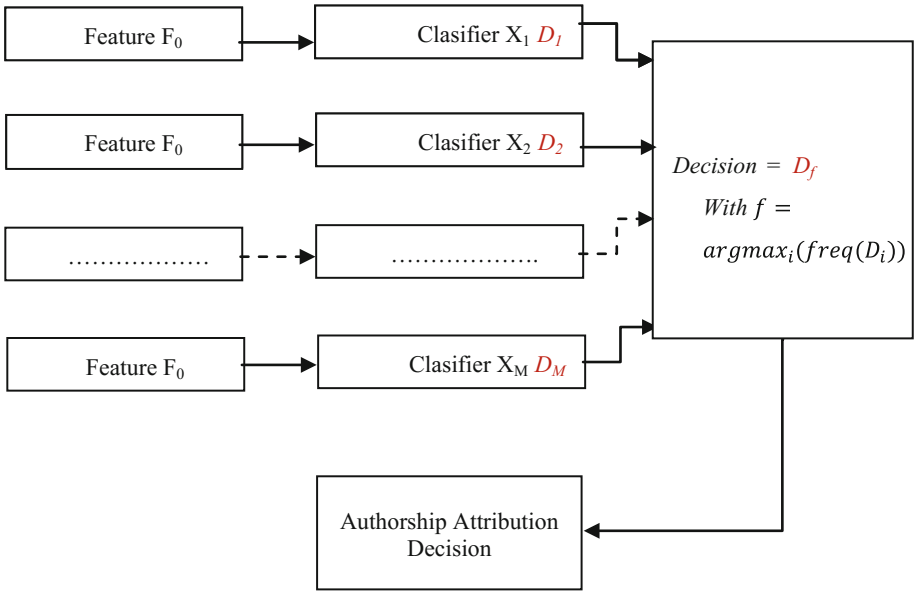

Fig. 4. Principle of the Classifier-based Decision Fusion (CDF)

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

$$Decision = D_f, \; with f = argmax_i(freq(D_i)) \tag{2}$$

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

All the results of the fusion approach are represented in Tables 1 and 2, summarizing the corresponding AA scores of the first and second fusion techniques respectively.

Table 1. Error of identification with and without *feature-based fusion (FDF)*

	Total Identification error on the 7 books	Quran book	Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquar-adawi's book	A-Khaled's book
Char-tetragram	**1.05%**	0%	0%	0%	0%	0%	11.1%	0%
Word	**1.05%**	0%	0%	6.25%	0%	0%	0%	0%
Word-bigram	**1.05%**	0%	0%	0%	0%	3.1%	0%	0%
FDF fusion	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**

Table 2. Error of identification using the *classifier-based fusion (CDF)*

	Total Identification error on the 7 books	Quran book	Hadith book	Aaid's book	Abdelkafy's book	Alghazali's book	Alquar-adawi's book	A-Khaled's book
Manhattan	**1.05%**	0%	0%	6.25%	0%	0%	0%	0%
SVM	**2.1%**	0%	0%	0%	0%	33.3%	0%	0%
MLP	**1.05%**	0%	0%	0%	16.7%	0%	0%	0%
CDF fusion	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**

4 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.

We noticed that the best results were obtained by the Manhattan distance and the MLP, which give an error of only 1.05%, the SVM present an identification error of 2.1%. Furthermore, in other experiments (*not mentioned in this paper*) the Manhattan distance outperformed all the evaluated classifiers, showing the high performances of this last one. The three authors: Aaid-Alkarni, Abdelkafy 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. We also noticed that the Quran and Hadith books are attributed without any error of attribution (*error of 0%*).

We noticed that Manhatan 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 text segments per book*).

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 1 and 2 we can see the corresponding results of those two fusion techniques respectively.

As we can see in the last line of Tables 1 and 2, 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/features 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.

So, the first conclusion one can state is that the fusion approach is quite interesting in multi-classifier or multi-feature authorship attribution. Furthermore, since there are no cross-errors of attribution between the Quran and Hadith texts (*each other*) and more generally, since there was not any cross-error of attribution for the Quran texts or Hadith texts with regards to the 6 other investigated books, we can state that these 2 books are completely different in style each other, and also different from all the other investigated books.

5 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 dialect is the same (*i.e.*

Standard Arabic) for the 7 books, the genre of the books is the same and the topic is 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% to 99% 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.

Concerning the comparison between the Quran and Hadith books, the related results (*of this investigation*) have shown that the Quran texts and Hadith texts are statistically different with a discrimination score of 100% (*i.e. discrimination error of 0%*), with or without using the fusion, and should probably belong to two different Authors, which also implies that the Quran could not be written by the Prophet.

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? Autorship attribution, short samples, big problem. In: Digital Humanities 2010 Conference, London, pp. 132–135 (2010)
3. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, pp. 513–520, August 2008
4. Juola, P.: Authorship attribution. Found. Trends Inf. Retrieval **1**(3), 233–334 (2006). https://doi.org/10.1561/1500000005. Now Publishing, USA
5. Love, H.: Attributing Authorship: An Introduction. Cambridge University Press, Cambridge (2002)
6. McMenamin, G.R.: Forensic Linguistics — Advances in Forensic Stylistics. CRC Press, Boca Raton (2002)
7. Fodil, L., Ouamour, S., Sayoud, H.: Theme classification of arabic text: a statistical approach. In: TKE 2014 Conference: Terminology and Knowledge Engineering, 19–21 June 2014, Berlin, Germany (2014)
8. Baraka, R., Salem, S., Abu-Hussien, M., Nayef, N., Abu-Shaban, W.: Arabic text author identification using support vector machines. J. Adv. Comput. Sci. Technol. Res. **4**(1), 1–11 (2014)
9. Sayoud, H.: Author discrimination between the Holy Quran and Prophet's statements. LLC J. Lit. Linguist. Compt. **27**(4), 427–444 (2012)

10. Sayoud, H.: Authorship classification of two old arabic religious books based on a hierarchical clustering. In: LRE-Rel: language resources and evaluation for religious texts, Lütfi Kirdar Convention & Exhibition Centre Istanbul, Turkey, pp. 65–70 (2012)
11. Ibrahim, I.A.: A brief illustrated guide to understanding Islam. Library of Congress, Catalog Card Number: 97-67654. Published by Darussalam, Publishers and Distributors, Houston (1999). http://www.islam-guide.com/contents-wide.htm, ISBN: 9960-34-011-2
12. Islahi, A.A.: Fundamentals of Hadith Interpretation; Hashmi, T.M. (English Trans.): Mabadi Tadabbur-i-Hadith. Al-Mawrid, Lahore (1989). www.monthly-renaissance.com/DownloadContainer.aspx?id=71
13. Sayoud, H.: Automatic speaker recognition – Connexionnist approach. Ph.D thesis, USTHB University, Algiers (2003)
14. Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J.: Weka: practical machine learning tools and techniques with Java implementations. In: Kasabov, N., Ko, K. (eds.) Proceedings of the ICONIP/ANZIIS/ANNES 1999 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems, Dunedin, New Zealand, pp. 192–196 (1999)
15. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput. **13**, 637–649 (2001)
16. Linear Regression. Last visit in 2013. http://en.wikipedia.org/wiki/Linear_regression
17. Huang, X., Pan, W.: Linear regression and two-class classification with gene expression data. Bioinformatics **19**(16), 2072–2078 (2003)
18. Tchechmedjiev, A., Schwab, D., Goulian, J.: Fusion strategies applied to multilingual features for an knowledge-based word sense disambiguation algorithm: evaluation and comparison. In: CICLING 2013 Conference: 14th International Conference on Intelligent Text Processing and Computational Linguistics, 24–30 March 2013, University of the Aegean, Samos, Greece (2013)
19. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 4–20 (2004)
20. Dasarathy, B.V.: Decision Fusion. IEEE Computer Society Press, Los Alamitos (1994)
21. Verlinde, P.: A Contribution to Multimodal Identity Verification using Decision Fusion. Ph.D thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, 17 September 1999
22. 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)