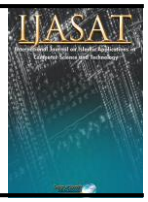


Stylometric Authentication of an Uncredible Extra-Hadith Collection

Halim Sayoud

EDT, Faculty of Electrical Engineering
USTHB University, Algiers
http://sayoud.net
halim.sayoud@gmail.com

Abstract
In this paper, we describe a survey on the stylometric authentication of an uncredible extra-dataset claimed to be a part of the Hadith, but for which religious scholars showed that it was probably not (i.e., fabricated or weak collection). The extra-Hadith collection is analyzed and compared to the genuine certified Hadith book of Bukhari. For that purpose, we present a stylometric approach based on the author style of the Matn (i.e., pure speech of the Prophet - Pbuh). That is, two experiments are conducted and commented: the first experiment is an authorship attribution on 19 text segments; and the second experiment is an automatic document clustering on 15 text segments. In the first experiment, we used character 4-grams and the nearest neighbor classification technique with Manhattan distance. In the 2nd experiment, we used a Hierarchical Clustering with Manhattan distance and Spearman distance. The results of both classification and clustering experiments show a difference in author style between the uncredible extra-Hadith collection (or at least a main part of it) and the genuine Bukhari Hadith. Although the authentication technique is made here at the subset level (i.e., text subsets of about 500 words each), the obtained results give a scientific agreement to the Islamic religious scholars about their evaluation on the doubtful collection: the uncredible collection, or at least a main part of it, does not have the same author style as the genuine Hadith one.

***Keywords*:** Hadith, Author style, Arabic Language, Text Mining, Stylometry, Fabricated Hadith.

1. Introduction

The categories of Hadith include authentic (Saheeh), good (Hassan), weak (Daeef), and fabricated (Mawdu) based on two main characteristics of Isnad (ie, who narrated the Hadith in the chain of reporters) and Matn (ie, the main text of Hadith) (Hakak, 2022).

The authentication of Hadith entirely depends on the authenticity of the chain of narrators reporting Hadith. Hardly any serious attention is paid to the authenticity of Hadith by the authentication of its text. Islamic scholars believe that if the chain of narrators of a hadith fulfils five criteria, the hadith is to be accepted as authentic: 1- continuity in the chain of narrators; 2- integrity of character; 3- infallible retention; 4- freedom from any hidden defect; and 5- safety from any aberrance (Khan, 2010).

Some researchers tried to make a scientific authentication based on chain narrators approach (Shukur, 2011), which is a process of identifying the chain of Hadith transmission from one narrator to another. Others proposed an analysis to produce a hierarchy with different levels of related studies in computational hadith to link with the computational authentication of isnad al-hadith science (Ibrahim, 2016). Although the results are interesting, they are not sufficient to make a clear decision on the authenticity of a Hadith subset.

In the present research work, we present a stylometric approach for the task of Hadith subset authentication. For that purpose, an experimental authenticity evaluation is conducted on two textual corpora: a genuine Bukhari Hadith subset and a doubtful Extra-Hadith collection that is not approved by Islamic scholars.

2. The Bukhari Hadith – A High Confidence Book

About Sahih al-Bukhari Collection

Ṣaḥiḥ al-Bukhari is a collection of hadith compiled by Abu Abdullah Muhammad Ibn Ismail al-Bukhari. His collection is recognized to be one of the most authentic collections of the Sunnah of the Prophet (Pbuh). It contains roughly 7563 hadith (Khan 1997).

About Al-Bukhari

Imam Al-Bukhari is known as the Amir al-Mu'minin in hadith (king in Hadith). His father Ismail was a well-known and famous Muhaddith in his time and had been blessed with the chance of being in the company of Imam Malik, Hammad Ibn Zaid and also Abdullah Ibn Mubarak (Khan 1997).

Imam Al-Bukhari was born on 13th of Shawwal 194 (A.H.). His father passed away in his childhood. At the age of sixteen after having memorized the compiled books of Imam Wakiy and Abdullah Ibn Mubarak, he performed Hajj with his elder brother and mother. After the completion of Hajj he remained in Makkah for a further two years and upon reaching the age of eighteen, he headed for Madinah, compiling the books "Qadhayas-Sahabah wa at-Tabi'in" and "Tarikh al-Kabir." Imam Al-Bukhari also traveled to other key centers of Arabia in search of knowledge like Syria, Egypt, Kufa, Basra, and Baghdad (Khan 1997).

Imam Al-Bukhari first started listening and learning hadith in 205 A.H., and after benefiting from the scholars of his town he started his travels in 210 A.H. His memory was considered to be one of a kind; after listening to a Hadith sentence he could repeat it from memory. It has been known that in his childhood he had memorized 2000 Hadiths (Khan 1997). There are a number of books compiled by Imam Al-Bukhari. However, his Ṣaḥiḥ collection is regarded as the highest authority of the collection of Hadith. He named this book "Al-Jami` al-Musnad as-Sahih al-Mukhtasar min Umuri Rasulullahi sallallahu 'alaihi wa sallam wa Sunanihi wa Ayyamihi", in Arabic: (الجامع المسند الصحيح المختصر من أُمور رسول الله صلى الله عليه وسلّم وسننه وأيامه) (Khan 1997).

After finishing his work, he showed the manuscript to his teachers Imam Ahmad ibn Hanbal for approval, along with Ibn al-Madini, and lastly Ibn Ma`in. It has also been recorded that it took Imam Al-Bukhari a period of 16 years to gather the Hadith and to write the Sahih book (Khan 1997). Furthermore, before he actually placed a Hadith in his compilation he prayed two prayers asking Allah for guidance. He finalized each Hadith in the Rawdah of Masjid an-Nabawi (between the Prophet's grave and his Minbar) and wrote the Hadith in the Mosque. However, he did not place any Hadith in his collection unless he was completely satisfied with that Hadith (Khan 1997).

Al-Bukhari's method of collection

Al-Bukhari imposed some strict conditions that all narrators in the Hadith chain must have met before a Hadith could be included in his book (Khan 1997):

- All narrators in the chain must be just and fair.
- All narrators in the chain must possess strong memory and all the scholars (Muhadditheen) who possess great knowledge of Hadith must agree upon the narrators' ability to learn and memorize.
- The chain must be complete without any missing narrator.
- It must be known that consecutive narrators in the chain met each other. This is a specific condition of Al-Bukhari in order to keep only authentic and confident Hadith.

Imam an-Nawawi relates that all religious scholars in Islam have agreed that Sahih al-Bukhari has gained the status of being the most authentic book after the Quran (Khan 1997).

3. Method of Authorship Identification

In practice, retrieving the real author of a piece of text has raised several questions and problems for centuries. The scientific field related to this authorship problem is called stylometry or author recognition (Sayoud, 2012). It can be of interest not only to humanities researchers, but also to politicians, historians and religious scholars in particular, as it is the case in this survey.

Stylometry 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 (Signoriello, 2005) (Eder, 2010).

Several methods of authorship attribution exist based on statistical techniques, neural networks or both. In this survey we used two classification methods: a nearest neighbor algorithm and a hierarchical clustering. As for the features, we used character 4-grams because of their efficiency in authorship attribution (Ouamour, 2018).

Last, but not the least, during the text collection of our corpus, we only kept the pure speech of the Prophet (i.e., Matn), by discarding all the chain of narrators script (i.e., Isnad), in order to make a fair stylometric analysis.

4. Experiments of Authorship Identification

In these experiments, we want to check the authenticity of an extra-Hadith collection, which was claimed to be a part of the Hadith, but where the Islamic religious scholars consider it as an "Untrue Hadith" collection or "احاديث مشهورة غير صحيحة" (Kelala 2022) (Alsaggaf, 2022), and which is composed of some fabricated and weak Ahadeeth without probably any relation with the Prophet (Pbuh).

That is, two classification methods are employed: a nearest neighbor algorithm and a hierarchical clustering. As for the features, we used character 4-grams because of their efficiency in authorship attribution.

1st Experiment: Automatic Author Classification

In this experiment, the two text datasets are segmented into text segments of 500 words each, which will produce a total of 19 segments corresponding to 11 segments from the genuine Bukhari Hadith and 8 segments from the fabricated one.

The used features are character 4-grams and the classification task is ensured by a Centroid-based Nearest neighbor algorithm (Manhattan distance) with an LOO cross-validation process. The results are displayed in table 1.

The medium LOO Accuracy of the authorship attribution experiment on the 19 text segments is 100% of correct attribution. As a consequence, we can conclude that the author style of the extra-Hadith collection is very different from the Hadith style, which strengthens the affirmation of the religious scholars about the non-authenticity of that doubtful collection or at least a main part of it.

Table 1: Automatic author style classification of all the text segments into "Hadith" and "Different from Hadith".

Ref.	Text Segment	Automatic Author Style Classification
1.	Fabric-1	Different from Hadith
2.	Fabric-2	Different from Hadith
3.	Fabric-3	Different from Hadith
4.	Fabric-4	Different from Hadith
5.	Fabric-5	Different from Hadith
6.	Fabric-6	Different from Hadith
7.	Fabric-7	Different from Hadith
8.	Fabric-8	Different from Hadith
9.	Hadith-1	Similar to Hadith
10.	Hadith-2	Similar to Hadith
11.	Hadith-3	Similar to Hadith
12.	Hadith-4	Similar to Hadith
13.	Hadith-5	Similar to Hadith
14.	Hadith-6	Similar to Hadith
15.	Hadith-7	Similar to Hadith
16.	Hadith-8	Similar to Hadith
17.	Hadith-9	Similar to Hadith
18.	Hadith-10	Similar to Hadith
19.	Hadith-11	Similar to Hadith

2nd Experiment: Automatic Documents Clustering by Author

In this experiment, the two text datasets are segmented into text segments of about 1000 words each, which will produce a total of 15 segments corresponding to 11 segments from the genuine Bukhari Hadith and 4 segments from the fabricated one.

The used features are character 4-grams and the clustering task is ensured by a hierarchical clustering with Manhattan distance and Spearman distance, which will produce a dendrogram representing the possible linkage between the different text documents. The results are displayed in figures 1 and 2.

According to figures 1 and 2, there are 2 main different clusters: a blue cluster, on the right, grouping all genuine Hadith segments, and a red cluster, on the left, grouping all fabricated segments. The results show that the extra-Hadith collection is different from the Hadith style. So, once again, these results strengthen the affirmation of the religious scholars on the doubtful collection or at least on a main part of it.

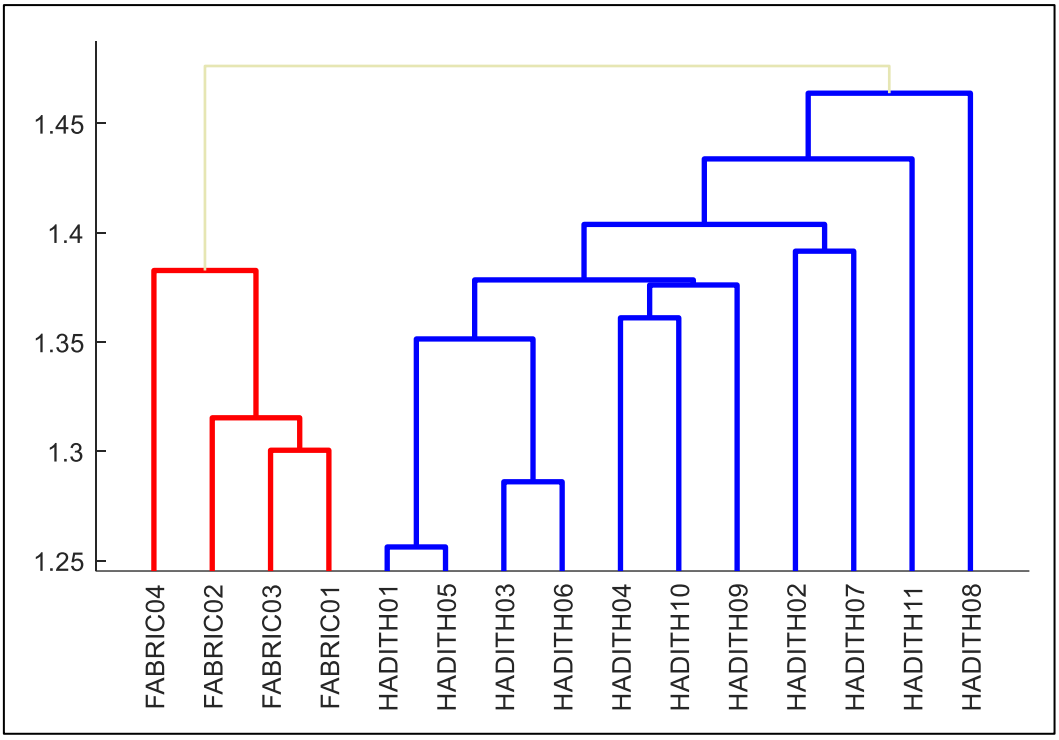


Figure 1: Hierarchical Clustering of the different text segments by using Manhattan distance. In blue: the Bukhari Hadith segments, and in red: the fabricated text segments.

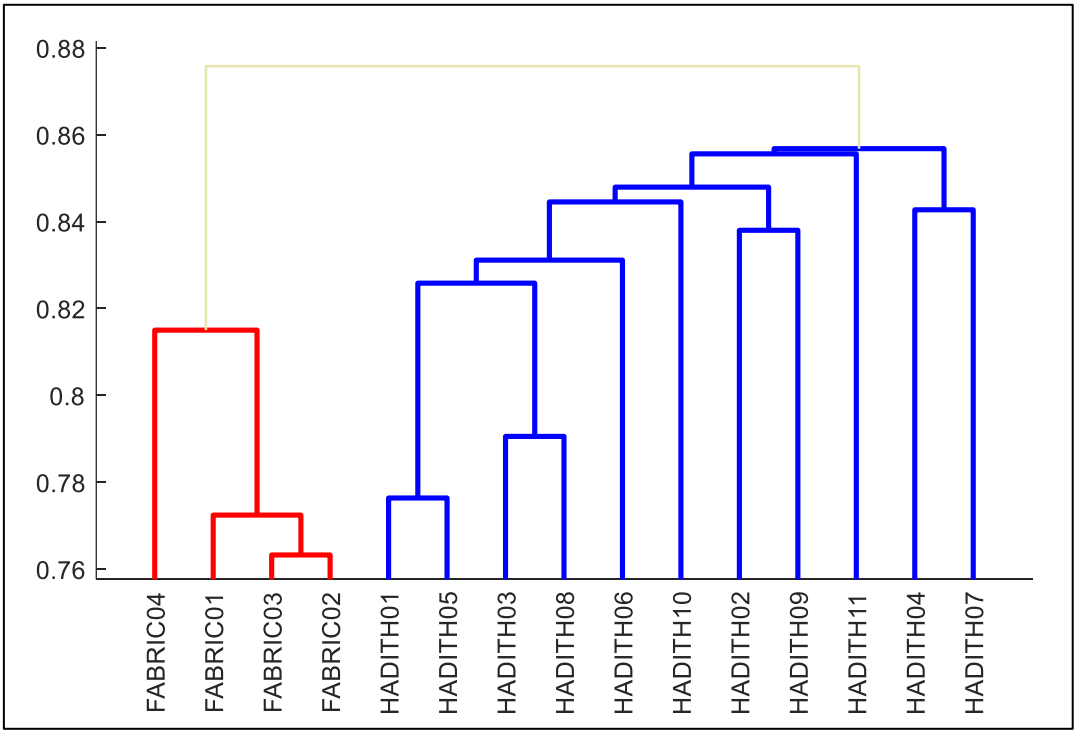


Figure 2: Hierarchical Clustering of the different text segments by using Spearman distance. In blue: the Bukhari Hadith segments, and in red: the fabricated text segments.

5. Conclusion

In this research work, we presented a survey on the stylometric authentication of a doubtful text collection, claimed to be a part of the Hadith, but which was not approved by religious scholars. The doubtful extra-Hadith collection, known as the Untrue Collection " احاديث غير صحيحة", has been segmented into text segments and then analyzed and compared to the genuine Bukhari Hadith. Two types of experiments have been conducted and commented: in the first experiment, one made an authorship attribution on the different text segments; and in the second experiment, one made an automatic clustering on the different text segments. In both experiments, the character 4-grams were used as features. The results of the two experiments (i.e., Automatic author style classification and Automatic hierarchical clustering) have shown that the author style of the doubtful collection, or at least a main part of it, is different from the author style of the genuine Bukhari Hadith.

6. Discussion

In this research work, the stylometric investigation showed that the whole extra-dataset (at a whole) is stylistically different from the Hadith, but it does not mean that every sentence included in that dataset is typically different, since stylometry requires a minimum of data size to work properly. However, even though the proposed authentication is made here at the subset level (i.e., text subsets of at least 500 words each), the obtained results give a scientific agreement to the Islamic religious scholars about their evaluation on the doubtful collection: the uncredible collection, or at least a main part of it, does not have the same author style as the genuine Hadith one.

References

A Alsaggaf A. (2022) الدرر السنية (dorar.net) ALDORAR ALSANIYYAH, أحاديث منتشرة لا تصح - https://www.dorar.net/fake-hadith?page=6, last access in November 2022.

Eder M. (2010) "Does size matter? : autorship attribution, short samples, big problem," In Digital humanities 2010 conference, pp.132-135, London, 2010.

Hakak, S., Kamsin, A., Zada Khan, W., Zakari, A., Imran, M., bin Ahmad, K., & Amin Gilkar, G. (2022). Digital Hadith authentication: Recent advances, open challenges, and future directions. Transactions on Emerging Telecommunications Technologies, 33(6), e3977.

Ibrahim, N. K., Noordin, M. F., Samsuri, S., Seman, M. S. A., & Ali, A. E. B. (2016). Isnad Al-hadith computational authentication: An analysis hierarchically. In 2016 6th international conference on information and communication technology for the muslim world (ICT4M) (pp. 344-348). IEEE.

Kelala, N. أحاديث مشهورة غير صحيحة. https://islamonline.net/أحاديث-نبوية-مشهورة-ومنتشرة-لكنها-غير/ Last access in September 2022.

Khan, M. M. (1997). Sahih-Al-Bukhari: Arabic-English Translation. Darussalam, KSA.

Khan, I. A. (2010). Authentication of hadith: Redefining the criteria. IIIT The International Institute of Islamic Thought (IIIT), Herndon, USA.

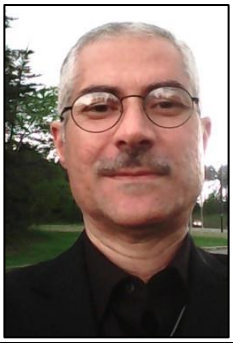
Ouamour, S., & Sayoud, H. (2018). A Comparative Survey of Authorship Attribution on Short Arabic Texts. In International Conference on Speech and Computer (pp. 479-489). Springer, Cham.

Sayoud, H. (2012). Author discrimination between the Holy Quran and Prophet's statements. Literary and Linguistic Computing, 27(4), 427-444.

Shukur Z, Fabil N, Salim J, Noah SA. (2011). Visualization of the hadith chain of narrators. Paper presented at: International Visual Informatics Conference; 2011:340-347; Springer.

Signoriello D. J., Jain S., Berryman M. J., Abbott D. (2005). "Advanced text authorship detection methods and their application to biblical texts," Proceedings of SPIE (2005), Volume: 6039, Publisher: Spie, pp. 163–175, 2005.

Biodata



Pr Halim Sayoud is Full Professor at the USTHB University. He is the head of the EDT research team. Pr Halim Sayoud published about 100 scientific research papers in conferences proceeding or international journals and is also the Editor-in-Chief of the HDSKD international journal. He is particularly interested in the following research fields: Speaker Recognition, NLP, Stylometry, Text categorization, Ancient documents analysis and Artificial Intelligence. Official website: http://sayoud.net
Personal website: http://scholarpage.org/sayoud.html