AUTHORSHIP DISCRIMINATION ON QURAN AND HADITH USING DISCRIMINATIVE LEAVE-ONE-OUT CLASSIFICATION

Halim Sayoud

http://sayoud.net

USTHB University

halim.sayoud@uni.de

ABSTRACT

In this survey, we try to make an investigation of authorship discrimination on two ancient religious books: Quran and Hadith, which should be fair and significant. The proposed approach is based on the Leave-One-Out (LOO) cross-validation technique based on support vector machine. The two documents are segmented into distinct text segments of 2900 tokens each, and the used features are composed of character-tetragrams, which are known to be quite efficient in stylometry. The cross-validation technique consists in 37 different experiments of authorship attribution that are carried out in a rotating manner, excluding every time one new sample (i.e. Leave-One-Out dynamic configuration).

In every singular experiment, the attribution score was 100%, which lead to an overall cross-validation accuracy of 100% between the two books.

This investigation shows that the two analysed books are stylistically different with a quite great significance, and confirms the theory of two different Authors. This important conclusion confirms what has been stated by the Prophet: the Quran was only sent down to him (by God), and he was only the narrator but not the author. This conclusion also denies the assumptions and claims of some persons claiming that the Quran was only an invention of the Prophet.

KEYWORDS

Natural language processing, Authorship discrimination, Stylometry, Leave-One-Out, Cross-validation, Quran.

INTRODUCTION

Stylometry is a research field related to author identification by exploring the writing style (Hu 2016). It solved many problems and disputes regarding the actual author of a piece of text. It has been widely used in intelligence and security purposes, in forensics and in religious investigations. Moreover it was also used for a goal of curiosity, such as in Shakespeare's disputed documents (Rudman 2016).

That is, in most cases we used and accepted a single experimental validation for getting the decision of authorship. Even more, many works in this field do limit their experiments to a single training / testing corpus to use, and then the obtained scores of classification are mentioned and accepted without any confidence parameter for assessing the results consistency. So, such results are not significant enough, even if the proposed precision formulas were quite-interesting.

Fortunately, some statisticians provided interesting tools and ways to evaluate the consistency of a classification result. This is roughly called cross-validation and, actually, several techniques do exist in the literature.

One of the most interesting one is the so called "Leave-One-Out" terchnique, which was proposed by Lachenbruch in 1967 (Lachenbruch 1967).

In this investigation we propose to use this cross-validation technique to get a fair accuracy of classification and discrimination on two sets of text segments to be classified, using an SMO-SVM classifier.

Our interest is focused on two important religious books, namely: the holy Quran (words of God) and the Hadith (statements of the Prophet).

As stated in the holy Quran and confirmed by the Prophet, the Quran represents the words of God. It was only sent down to the Prophet (by God), but not written by him.

However, some doubts claimed that the Quran could be only an invention of the Prophet, which means that it could be written by him (according to those claims).

Now, to get a scientific response to that question, we thought that it could be interesting to use stylometry for analysing the two books and see whether the two writing styles are similar or not, since the genre and theme are the same.

THE LEAVE-ONE-OUT METHOD

The Leave-One-Out Method is a jackknife method for evaluating the classification accuracy (Vehtari 2016). It was proposed by Lachenbruch in 1967 (Lachenbruch 1967). His approach was based on discriminant analysis; it has been named the leave-one-out (L-O-O) method (Huberty, 1994). This technique has two steps:

-First, the template is built in the samples with one observation removed,

-Then the resulting estimate parameters (of the training) are used to classify the single removed observation. The main process is repeated M times so that each observation was removed and classified once (see Figure 1), where M represents the number of samples (Kroopnick, 2010).

Eventually, the proposed measure of good classification is given by the number of times that the removed observation was correctly classified (Huberty, 1994) (Kroopnick, 2010).

To evaluate the L-O-O method, Lachenbruch conducted a small Monte Carlo simulation with 300 replications for a two group discriminant analysis. His results showed the efficiency of Lachenbruch's L-O-O technique (Kroopnick, 2010).

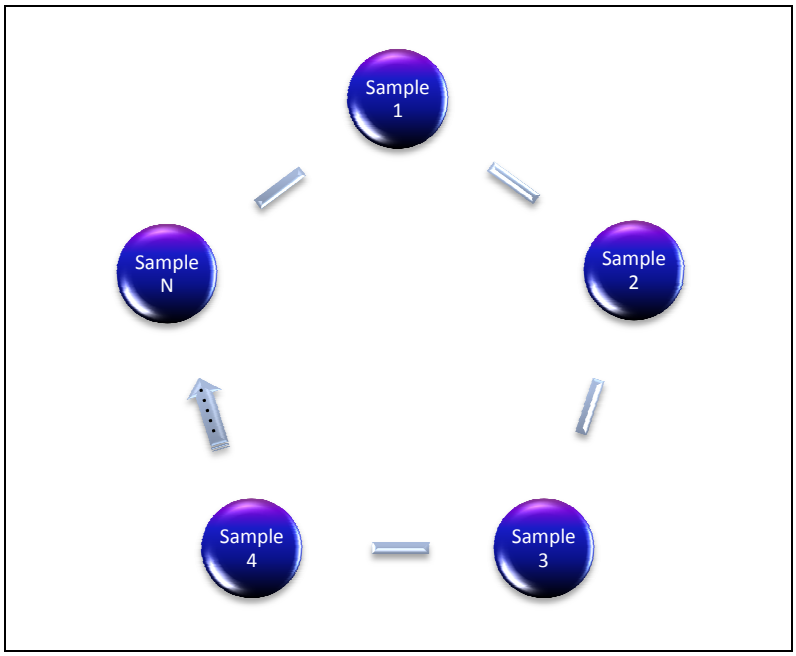

Figure 1.a. Set of the N samples to classify (Sample 1, Sample 2, … Sample N).

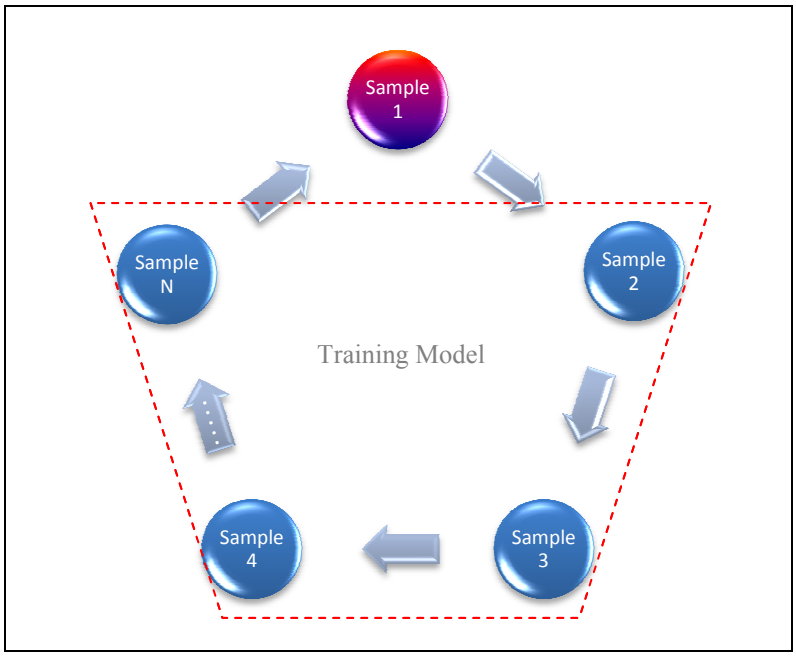

Figure 1.b. The Leave-One-Out algorithm applied to Sample 1 (start of the algorithm).

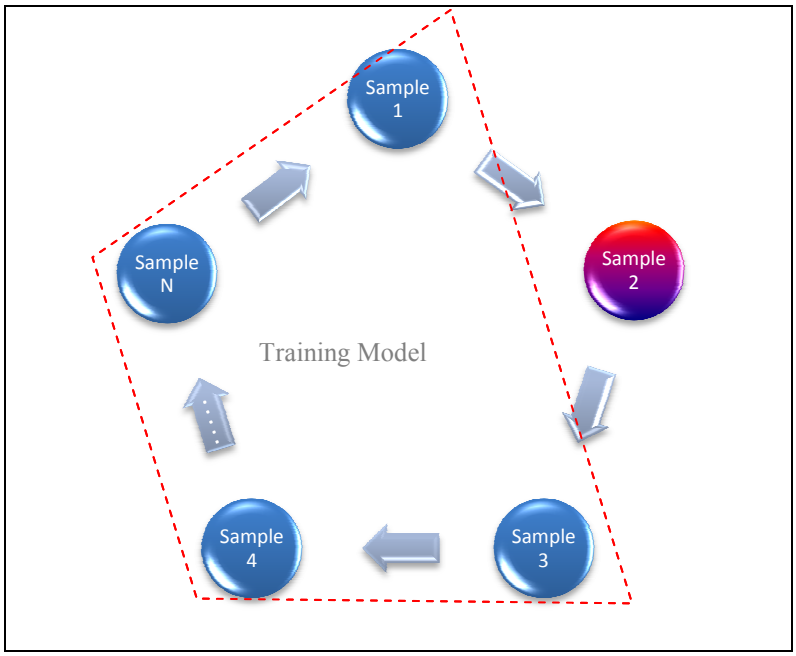

Figure 1.c. The Leave-One-Out algorithm applied to Sample 2 and moving to the next sample.

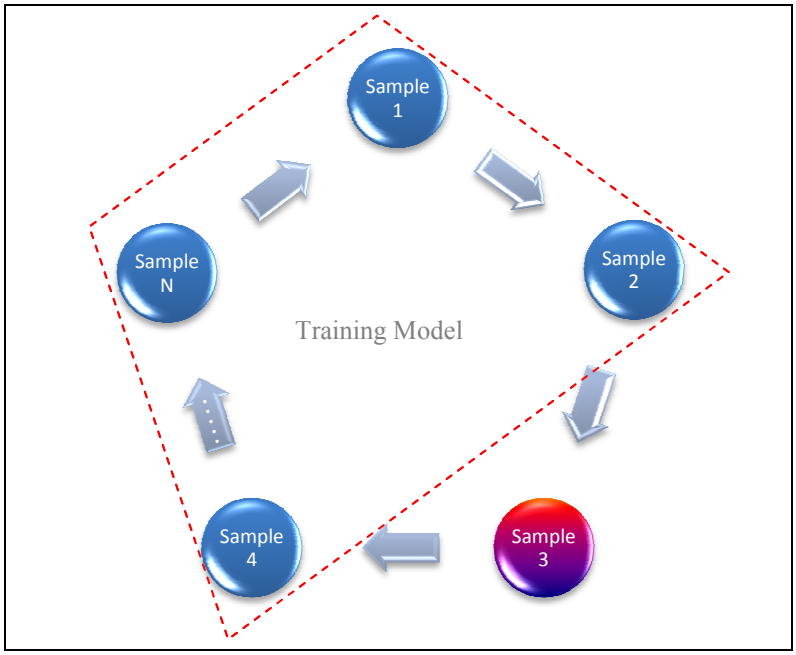

Figure 1.d. The Leave-One-Out algorithm applied to Sample 3 and moving to the next sample.

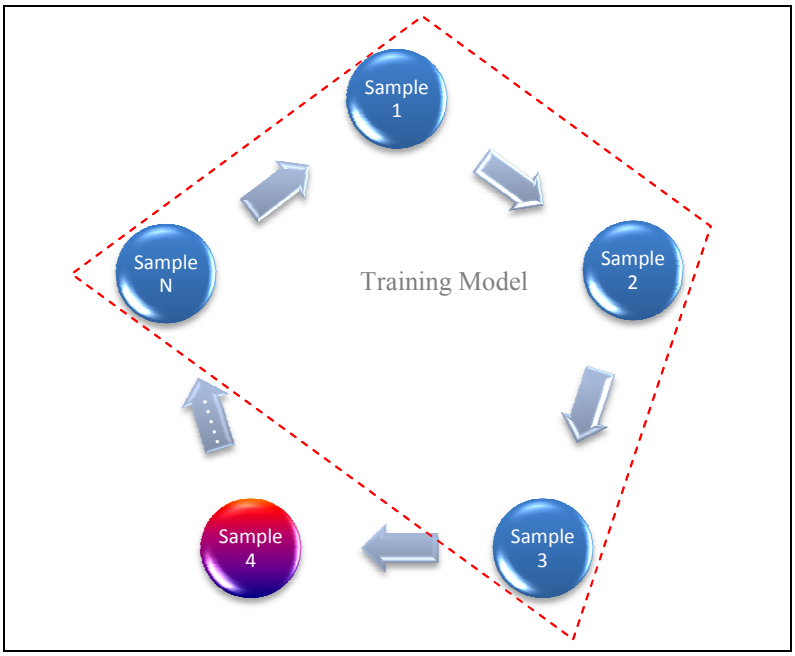

Figure 1.e. The Leave-One-Out algorithm applied to Sample 4 and moving to the next sample.

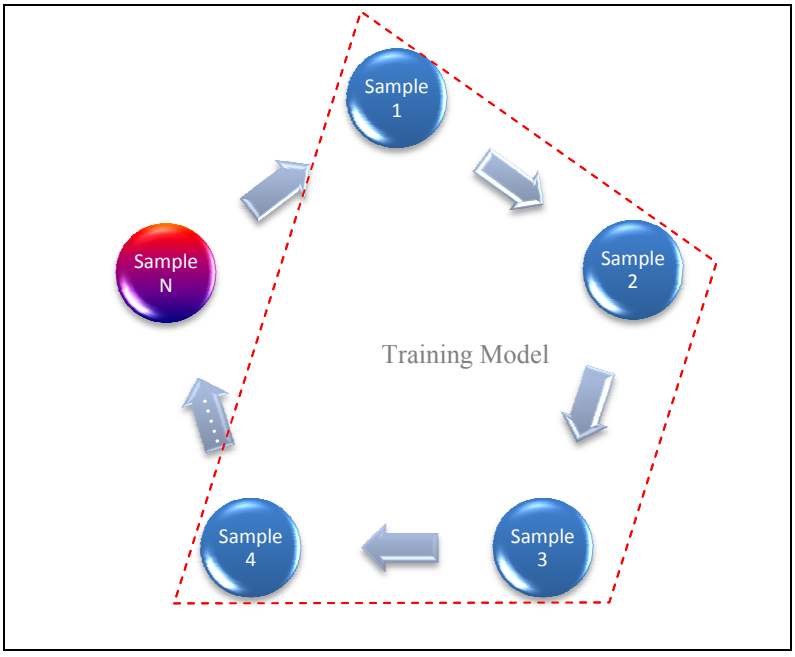

Figure 1.f. The Leave-One-Out algorithm applied to Sample N (end of the algorithm).

ABOUT THE FEATURES

In the literature, one can find several linguistic features that are proposed in the field of authorship attribution [Ranatunga 2013]. One can quote four main types as follows:

Vocabulary based Features: In general, the typical words, an author is used to write, can reveal his or her identity. The problem with such features is that the data can be faked easily. A more reliable method would be able to take into account a large fraction of the words in the document [Juola, 2006] as the average sentence length.

Syntax based Features: One reason that function words perform well is because they are topic-independent [Juola, 2006]. A person's preferred syntactic constructions can be cues to his authorship. One simple way to capture this is to tag the relevant documents for part of speech or other syntactic constructions (Stamatatos, 2001) using a tagger.

Orthographic based features: This feature could be interesting because one weakness of vocabulary-based approaches is that they do not take advantage of morphologically related words. A person who writes of "work" is also likely to write of "working", "worker", etc. [Juola, 2006].

Characters based features: Some researchers [Peng, 2003] have proposed to analyze documents as sequences of characters. This type of parameter can replace several other high-level linguistic features. Furthermore, several experiments showed that character n-gram is quite reliable in authorship attribution [Stamatatos, 2009].

In our investigation, we chose to use the last one since it has been shown that they are extremely pertinent, especially character trigrams and tetragrams. So, in this investigation we have used character-tetragrams.

ABOUT THE CLASSIFIER

In the literature, one can find different types of classifiers that are employed in discrimination, such as: statistical models, neural networks, support vector machine (SVM), linear regression, simple distances, etc. However some previous researches showed that the SVM is one of the best classifier in discrimination, especially in biometrics and stylometry. Actually, one can quote the works of Ouamour et al. in 2016 (Ouamour 2016), in speaker discrimination, and the previous works of Ouamour et al. in authorship attribution (Ouamour 2013), which clearly showed the superiority of the SVM over the other investigated classifiers. Hence, concerning the task of speaker discrimination (Ouamour 2016), the authors implemented nine different classifiers, namely: Linear Discriminant Analysis, Adaboost, Support Vector Machines, Multi-Layer Perceptron, Linear Regression, Generalized Linear Model, Self Organizing Map, Second Order Statistical Measures and Gaussian Mixture Models. Experiments of speaker discrimination were conducted on Hub4 Broadcast-News. Results showed that the best classifier is the SVM, which outperformed all other classifiers in this research work. Again, concerning the task of authorship attribution (Ouamour 2013), the authors investigated the authorship of several short historical texts that are written by ten ancient Arabic travelers: called AAAT dataset.
Several experiments of authorship attribution are conducted on these Arabic texts, by using seven different classifiers, namely: Manhattan distance, Cosine distance, Stamatatos distance, Camberra distance, Multi-Layer Perceptron (MLP), Sequential Minimal Optimization based Support Vector Machine (SMO-SVM) and Linear Regression. Results showed that the best performances of authorship attribution were given by the SVM (accuracy of 80%), which outperformed, once again, the other investigated classifiers.
For this reason, and knowing the good performances of the SVM in discrimination, we have decided to use this classifier for the task of authorship discrimination.

ABOUT THE DATASET

In this section, we will give a description of the two religious books, where the application of author discrimination has been made, namely: the Quran and Hadith.

Quran Description

The Quran (in Arabic: القرآن) is the central religious text of Islam [Nasr 2015], which is believed to be a revelation from God (الله, Allah) and which has been written by God too [Nasr 2015]. It is widely regarded as the finest piece of literature in the Arabic language.

Islam holds that the Quran was verbally revealed by God to Muhammad through the angel Gabriel (Jibril), gradually over a period of approximately 23 years. The beginning of the apparition of the Quran was in the year 610 (after the birth of Christ).

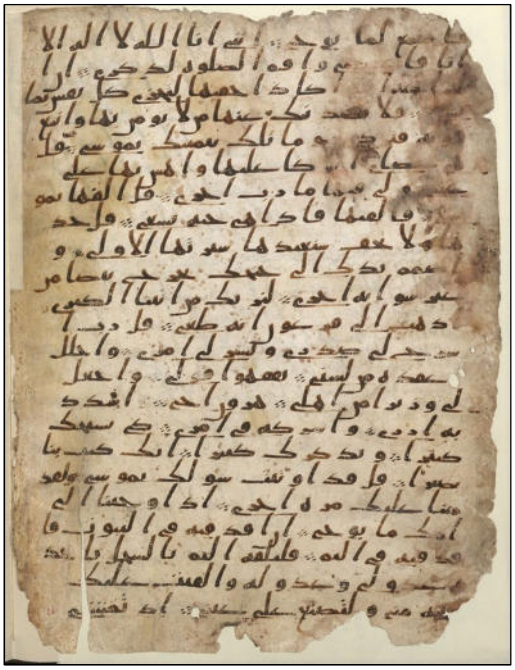

Figure 1: Old page of the holy Quran dating from the period of the Prophet's companions. Courtesy of Birmingham University.

Hadith Description

The Hadith (in Arabic: الحديث) is the oral statements and words said by the Prophet Muhammad (PBUH) [Islahi 1989]. Hadith are collections of the reports claiming to quote what the prophet Muhammad said. 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 research work, we used the Bukhari Hadith, which is considered as the most confident book of the Hadith.

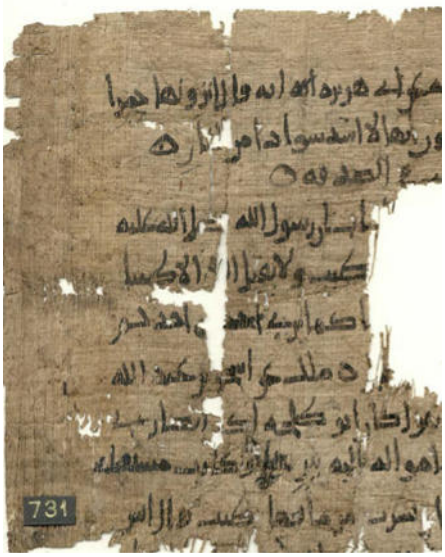

Figure 2: Old page of the Hadith. The fragment has been dated to Mālik's own day in the second half of the second century AH. Courtsy of the Austrian National Library of Vienna.

Dimension of the two religious books

The two books are analyzed in terms of words, tokens and average number of A4 pages. Table 1 gives those statistical characteristics.

Table1: Detailed description of the dataset

Book	Size in terms of token	Size in terms of words	Number of A4 page
1 st book: The Holy Quran	87341	13473	315
2 nd book : The Hadith (Sahih El-Bukhari)	23068	6225	87

According to these size details, the two religious books seem relatively consistent, since the average number of pages is 315 for the Quran book and 87 for the Hadith book. However, since the two books do not have the same size, it is necessary to segment those two books into segments of more or less the same size, in order to avoid unbalanced results.

TEXT SEGMENTATION

A text segmentation is applied in order to construct individual documents with the same size. In fact, when comparing two books with different sizes, it is difficult to know if a specific part of the book is similar to another one or different. That is why a smart segmentation has been proposed and applied to the different books.
The sizes of the segments are more or less in the same range: we obtain 29 different text segments for the Quran and 8 different text segments for the Hadith, with approximately the same size. So, we get 37 different text segments of about 2900 words each in the whole dataset. Table 2 gives the number of words (tokens) contained in each text.
It has been shown in previous research works conducted by Eder [Eder 2010] and Signoriello [Signoriello 2005] that the minimum number of words per text should be about 2500 words in order to obtain a good AA result. So, our chosen configuration, namely: 2900 words per segment, seems to be correct and suitable to the different AA experiments.

Table2: Size of the different text segments in terms of tokens (number of words in the text)

Quran text segments		Hadith text segments	
Text segment designation	**Size in terms of tokens**	**Text segment designation**	**Size in terms of tokens**
Q1	2901	H1	2919
Q2	2903	H2	2898
Q3	2898	H3	2908
Q4	2907	H4	2897
Q5	2906	H5	2908
Q6	2897	H6	2904
Q7	2905	H7	2907
Q8	2901	H8	2727
Q9	2905	/	/
Q10	2906	/	/
Q11	2895	/	/
Q12	2899	/	/
Q13	2904	/	/
Q14	2906	/	/
Q15	2900	/	/
Q16	2896	/	/
Q17	2900	/	/
Q18	2901	/	/
Q19	2906	/	/
Q20	2902	/	/
Q21	2899	/	/
Q22	2900	/	/
Q23	2903	/	/
Q24	2903	/	/
Q25	2909	/	/
Q26	2900	/	/
Q27	2886	/	/
Q28	2900	/	/
Q29	2894	/	/

The segmented dataset is decomposed into 2 rotating parts (Leave-One-Out configuration): the training part containing all the text samples except one, and the testing part consisting in that removed one.

EXPERIMENTS OF AA USING THE L-O-O TECHNIQUE

We recall that there are 37 text segments (segment size of 2900 words each),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 the employed classifier is the SMO-based SVM.

Since there are 37 samples, we will also have 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 following table, we represent the scores of good classification corresponding to our 37 cross validation experiments.

Table 3: Results of AA using the L-O-O technique

Experiment Number	Tested document	Accuracy
1.	Q1	100%
2.	Q2	100%
3.	Q3	100%
4.	Q4	100%
5.	Q5	100%
6.	Q6	100%
7.	Q7	100%
8.	Q8	100%
9.	Q9	100%
10.	Q10	100%
11.	Q11	100%
12.	Q12	100%
13.	Q13	100%
14.	Q14	100%
15.	Q15	100%
16.	Q16	100%
17.	Q17	100%
18.	Q18	100%
19.	Q19	100%
20.	Q20	100%
21.	Q21	100%
22.	Q22	100%
23.	Q23	100%
24.	Q24	100%
25.	Q25	100%
26.	Q26	100%
27.	Q27	100%
28.	Q28	100%
29.	Q29	100%
30.	H1	100%
31.	H2	100%
32.	H3	100%
33.	H4	100%
34.	H5	100%
35.	H6	100%
36.	H7	100%
37.	H8	100%

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

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

Accirding to table 3, the average accuracy of all L-O-O experiments is **100%**.

DISCUSSION

Two ancient religious Arabic books (Quran and Hadith) were analysed by a discriminative authorship analysis using a Leave-One-Out validation

The features consist in character-tetragrams, while the used classifier is based on an SMO-SVM.

The dataset is composed of 37 text documents, where the size of a single segment is about 2900 tokens.

As we could see in the results section, the accuracy of every cross-validation step (i.e. for all the 37 L-O-O experiments) was 100%, leading to an average cross-validation score of 100% too.

From these results, one can deduce the following important conclusions:

- Firstly, the two books Quran and Hadith possess two different author styles;
- The segments of every book are quite similar in terms of style within a single book;
- The L-O-O cross validation technique shows that this result (discrimination score of 100%) is quite significant, since the same score has been obtained 37 times during the tests of cross-validation and with different configurations.

Consequently and according to this investigation, the two ancient books: Quran and Hadith appear to have two different styles and should probably come from two different Authors.

This important conclusion confirms what has been stated by the Prophet: the Quran was only sent down to him (by God), and he was only the narrator but not the author. This conclusion also denies the assumptions and claims of some persons claiming that the Quran was only an invention of the Prophet. So, how could he write that religious book while the scientific analysis of the Quran and the Hadith (statements of the Prophet) are completely different and where the L-O-O technique has shown the clear significance of those statistical results?

It appears obvious and clear now that the two analysed books come from two different authors and that the Quran could not be written by the Prophet, but it was probably only transmitted to him.

ACKNOWLEDGEMENTS

The author of this manuscript wish to warmly thank all those who helped him conducting this research work. He also welcomes all the comments of the readers and apologises for any unintentional mistake that may appear in this paper.

REFERENCES

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

Hu 2016 Xianfeng Hu, Yang Wang, Qiang Wu. Stylometry and Mathematical Study of Authorship. Book title: » New Trends in Applied Harmonic Analysis. Springer 2016, Part of the series Applied and Numerical Harmonic Analysis pp 281-300.

Huberty 1994 Huberty, C. J. (1994). Applied discriminant analysis. New York: Wiley.

Islahi 1989 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.

Juola 2006 P. Juola 2006. Authorship Attribution. Now Publishing, USA 2006.

Kroopnick 2010 Marc H. Kroopnick, Jinsong Chen, Jaehwa Choi, C. Mitchell Dayton. Assessing Classification Bias in Latent Class Analysis: Comparing Resubstitution and Leave-Out Methods, Journal of Modern Applied Statistical Methods. May, 2010, Vol. 9, No. 1, 2-331 pp52 – 63.

Lachenbruch 1967 Lachenbruch, P. A. (1967): An almost unbiased method of obtaining confidence interval for the probability of misclassification in discriminant analysis. Biometrics (December): 639-645.

Nasr 2007 S. H. Nasr, Encyclopædia Britannica Online. http://www.britannica.com/eb/article-68890/Quran, 2007.

Ouamour 2013 S. Ouamour, H. Sayoud. Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features. CyberC – International Conference on Cyber-enabled distributed computing and knowledge discovery CyberC conference - October 10-12, 2013.

Ouamour 2016 S. Ouamour, H. Hamadache, H. Sayoud. Title: Speaker Discrimination Using Several Classifiers and a Relativistic Speaker Characterization. International Conference on Image and Signal Processing, ICISP'2016, Quebec Canada, May 30- June 01, 2016. No: 21. pp 203-212.

Peng 2003 F. Peng, D. Schurmans, V. Keselj, and S. Wang, "Language independent authorship attribution using character level language models," in Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 267–274, Budapest: ACL, 2003.

Rudman 2016 Joseph Rudman, 2016, Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats. Journal of Early Modern Studies, n. 5 (2016), pp. 307-328.

Signoriello 2005 Signoriello, Domenic, Samant Jain, Matthew Berryman, and Derek Abbott. 2005. *Advanced text authorship detection methods and their application to biblical texts*. Proceedings of SPIE (2005), Volume: 6039, Publisher: Spie, Pages: 163–175

Stamatatos 2001 E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," Computers and the Humanities, Vol. 35, No. 2, pp. 193–214, 2001.

Stamatatos 2009 E. Stamatatos 2009. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, Vol. 60, No. 3, pp. 538-556, 2009, Wiley.

Vehtari 2016 Aki Vehtari, Andrew Gelman, Jonah Gabry, 2016. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, pp 1–20, Springer 2016.