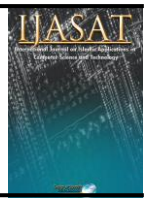


Was the Quran Written by the Prophet? -A Stylometric Investigation Using the Interrogative Form

H. Sayoud, H. Hadjadj
EDT, Faculty of Electrical Engineering
USTHB University, Algiers
halim.sayoud@gmail.com, hadjadj.has@gmail.com

Abstract
In this investigation, we try to see whether the holy Quran could have been written or dictated by the Prophet (Pbuh), thanks to a stylometric discriminative analysis of the corresponding Author styles. The originality of this research work lies in the use of a new set of linguistic features based on 26 interrogative features and a special fusion, which we called logarithmic feature fusion or LFF. The experiments have shown that the proposed features, with their fusion, are interesting. Furthermore, the application of discrimination on the Quran and Hadith has shown a great difference in Author style between the two books, which confirms that the Quran could not be written or dictated by the Prophet.

***Keywords*:** Quran, Hadith, Natural language processing, Interrogative features, Author classification, Stylometry.

1. Introduction

Authorship discrimination is a specific research field of stylometry, which consists of checking whether two different texts are written by the same author or not (Sayoud, 2012).
Stylometry, which can be defined as the statistical analysis of the literary style (Holmes, 2003), is widely used nowadays in several disputed problems.

Research work on authorship attribution usually appears in several types of debates ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite this interest, the field itself is somewhat in confusion regarding which are the best practices and techniques (Juola, 2006).

The literature shows various available techniques, which determine the author of a document. According to the literature, it appears that many works are reported for the English and Greek languages, but there are few serious works in Arabic language, especially for religious purpose, where there exist several old books that are supposed to belong to some authors and for which the authorship is put in doubt.

The application of this research work deals with a religious enigma, concerning the authorship of the holy Quran. In fact, many efforts to find a human source for the Quran do exist assuming for instance that the Quran could be written by the prophet Muhammad (Al-shreef, 2009). So, the key question is: “Was the Quran written by the Prophet? Or only transmitted to him by Allah?”. To respond to that question, it is important to handle the problem with delicacy and scientific rigor.

That is, in this research work, we propose a new set of features, called interrogative features, which are used for identifying the author of a text in Arabic. Those interrogative features (used alone) were never used before (to our knowledge) and are proposed for the first time in stylometry.

2. Related work

Stylometry deals with the relationship between the writing style of a text and the most likely corresponding authorship. The field of authorship recognition predates modern computing stylometry, with research starting in the late 19th century. For example, in 1887 Mendenhall (Shaker, 2010) proposed that word length can be used to distinguish works by different authors. The usage of the features of a text was later extended by Yule to include the lengths of sentences (Yule, 1938).

Authorship studies are currently the most popular application of stylometry. Many studies have been reported during the last years as described in (Juola, 2012), (Stamatatos, 2009), (Sayoud, 2015) and (Hoshila, 2016), where many debates were reported and several types of features and techniques were proposed too.

As noted in previous research (Juola 2006), a vast variety of approaches have been tried for authorship attribution, but no specific approach has emerged as the best one in this field.

Generally, learning approaches appear to be useful and successful for the purpose of authorship analysis. In Navinder et al (Navinder, 2015) an SVM classifier was applied to a Punjabi corpus consisting of poetry written by 10 different poets. Results showed an accuracy ranging from 70% to 80%. In the work of Otoom et al (Otoom, 2014), an intelligent system for author attribution based on a hybrid feature set was proposed, where five popular classification algorithms were used. In this case, an average accuracy of 80% was achieved.

One of the purposes of stylometry is authorship attribution related to religious disputes. Holmes (Mills 2003) stated that the area of stylistic analysis is the main contribution of statistics to religious studies. For example, early in the nineteenth century, Schleiermacher disputed the authorship of the Pauline Pastoral Epistle-1 Timothy (Mills, 2003). As a result, other German speaking theologians, namely F.C. Baur and H.J. Holtzmann, initiated similar studies of New Testament books (Mills, 2003). Since then, several investigations have been done on different pieces of religious texts and with different analysis techniques.

One can find recent works of author discrimination in Arabic (Baraka, 2014), but very few are applied to the Quran. In 2012, for instance, Sayoud presented a series of author discrimination experiments between the Quran and Hadith (Sayoud, 2012). Once, he used the two books in their entirety and another time, he segmented the books into 4 segments each. In both experiments he showed that the authors of the two books were different.

Furthermore, he tried to make an author discrimination between the Quran and the Hadith in a segmental form. Differently, in this research work, we try to make an authorship analysis of the two religious books: Quran and Hadith, by using new linguistic features that are exclusively based on the interrogative style.

3. Dataset

The used dataset is composed of the holy Quran and a subset of the Bukhari Hadith. The Quran is the central religious book of Islam, which Muslims consider to be a revelation from God (Allah) and that has been sent down by God too. The Hadith is the oral statements and words of the prophet Muhammad (Pbuh). In this research work, we used a subset of the Bukhari Hadith, considered as the most confident book of the Hadith. The two books are segmented into text segments of about 2900 words (see table 1).

For machine learning and classification purposes, the dataset can be decomposed into 2 parts: training part and testing part, and since the two books have different sizes (29 texts for the Quran and 8 texts for the Hadith), a logical rule would be: 4 text segments for the training of the Hadith book and 8 text segments for the training of the Quran book. And the remaining text segments could be kept for the testing step, for instance.

Table 1. Size of the different text segments in terms of tokens

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		

4. Proposed features and why Interrogative features?

Over time, many stylometric features have already been investigated and applied in the literature along with a wide variety of analysis models. However, there is no agreement among researchers regarding which features yield the best performances.
The stylometric features that are commonly used, are Character N-grams, Function words, Vocabulary richness, Lexical richness, Distribution of syllables, Word frequency, Word length distribution, Word collocations, Sentence length, etc.

In our investigation, a mixture of 26 interrogative features is proposed. All those features are original and used for the first time in stylometry. Interrogative features are elements that play a key role in the formation of interrogative constructions (Ginsburg, 2009).

In fact, an important step consists in retrieving distinctive features that exhibit the writing style and represent a certain authorship individually, and the interrogative features could be specific to a particular author.

Furthermore, Arabic is one of the world's greatest languages. Its graceful script, magnificent style and rich vocabulary give to the language a unique character and specificity. However, when we read the two books, we notice that the Interrogative form is widely used within the books. The interrogative form is one of the specificities that characterizes the Arabic language and has a great variety and richness in the literary style.

During the preparation of this research work, there were no published works in stylometry using interrogative features, especially in Arabic, despite the importance of the interrogative style in all languages of the world. This fact encourages the idea of establishing a research in Arabic authorship discrimination using those interrogative features.

According to the importance of the interrogative style in our investigation, a set of 26 interrogative features was extracted and collected from the two religious books. This set contains five types of interrogative features as follows:

- Interrogative words/verbs with the particle Hamza (أ / aa): this type of interrogation is commonly used in old Arabic and this particle could represent the question mark "?" or an auxiliary verb in the interrogative form. In our approach, we employed 22 interrogative elements of this type.
- Wondering verb "يسألونك", meaning "they ask you" in English: this type represents an indirect questioning.
- The starting particle أم meaning "or/ isn't it" in English. When it is employed at the beginning of a sentence, it refers to a form of question.
- The interrogative expression "من ذا" meaning "who?" in English: this interrogative expression is interesting since it gives a feeling of a strong personality when speaking.
- The interrogative expression "ما لكم" meaning "what is the matter with you? / why?" in English. This interrogative expression is also interesting for its strong style when speaking.

5. Authorship Discrimination Approach and Experiments

In our approach several steps are performed, as shown in Figure 1, namely: data pre-processing, feature extraction, LFF fusion, classification and author discrimination decision, while the dataset can also be organized into training and testing for a machine learning classification.

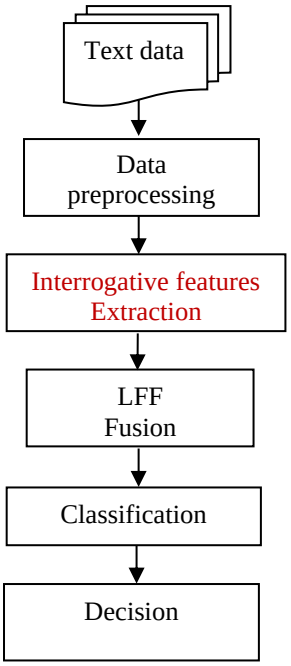


Figure 1. Authorship discrimination method

The Feature extraction process is an important step in author identification. In our case, we recall that we have proposed 26 interrogative features for characterizing the author style. Furthermore, we have defined a new type of fusion, which we called the *Logarithmic Feature Fusion* (LFF). This fusion form is defined as the logarithmic sum of the normalised features frequencies (see equation 1).

In fact, some previous works and experimental results showed that fusion is a competitive technique in terms of classification accuracy and computational complexity (Bota, 2020). For visual analysis purpose, a graphical representation of the LFFs for the 37 text segments is displayed in figure 4. The Logarithmic Feature Fusion (LFF) is given as follows:

$$Fusion = \log\left(\sum_{i} feature_i \,/\max\left(feature_i\right)\right) \quad (1)$$

with i representing the feature index (i=1,2,...26).

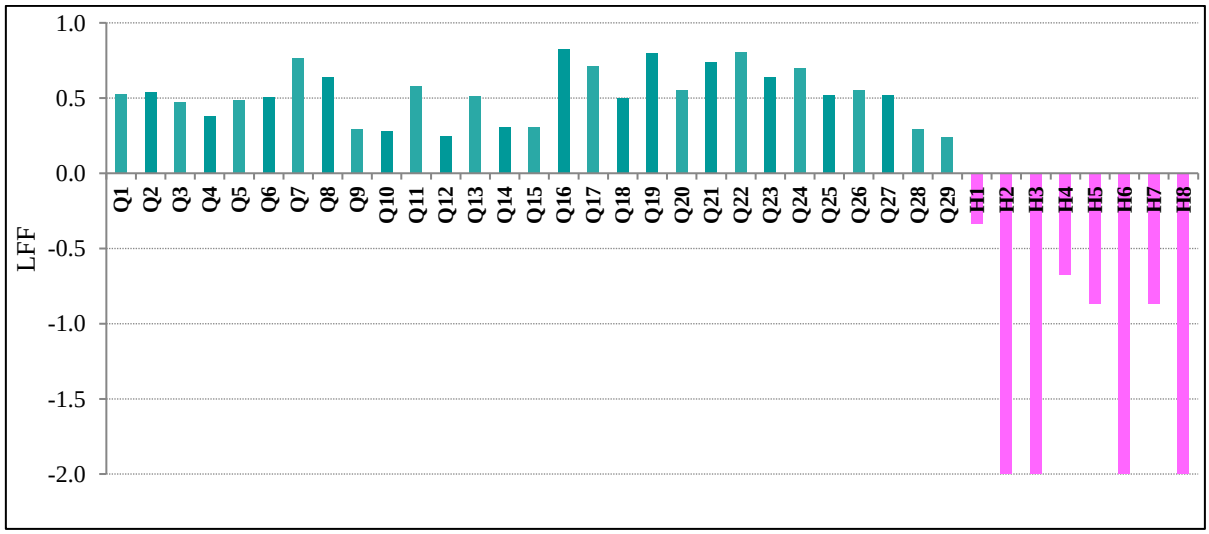


Figure 2. LFF values of the Quran segments (in green) and Hadith segments (in pink)

We can notice, in figure 2, that all the Quran LFFs are positive while all the Hadith LFFs are negative. This figure shows a sharp difference between the Quran segments in green, which present positive logarithmic values, and the Hadith segments in pink, for which the logarithmic values are negative.

Consequently, a simple algorithm of logical filtering can manage to discriminate the two types of documents (i.e. all Quran documents have positive LFF, while all Hadith ones have negative LFF). That is, there is no need to use complex machine learning techniques for that classification, since the only LFF sign is sufficient to make the discrimination efficiently. In this context, all text segments of the holy Quran (in green) appear stylistically different from the Hadith segments (in pink).

6. Discussion

This research work has addressed the problem of author discrimination between the holy Quran and Hadith, to check whether the two books could have been written by the same Author or not.

The originality of this research work lies in the use of a new set of features based on 26 interrogative features exclusively and a special fusion; and to the knowledge of the authors, it is the first time that those features are used alone in stylometry. Furthermore, a special fusion, which we called logarithmic feature fusion or LFF has been proposed and applied in a visual analytical way.

The different experiments of authorship attribution have led to the following important points:

- The proposed feature set is reliable for performing an author identification task in Arabic;
- Concerning our application of author discrimination, the experimental results of AA have clearly revealed that the structures of the interrogative styles of the Quran and Hadith are different.

From this last result, an important arising question would be: Could an author possess two interrogative styles completely different for the same topic? If so, what could be the reason to change his questioning style? And in what proportions would it be possible to do that variability?

Actually, we do not see any other explanation except the fact that the two studied books should have two different Authors, and consequently, we can deduce that the holy Quran could not be authored or invented by the Prophet.

This research investigation confirms well what is stated in the verse (4:105) of the holy Quran:
[إِنَّا أَنزَلْنَا إِلَيْكَ الْكِتَابَ بِالْحَقِّ لِتَحْكُمَ بَيْنَ النَّاسِ بِمَا أَرَاكَ اللَّهُ ۚ وَلَا تَكُن لِّلْخَائِنِينَ خَصِيمًا]

Translation: “Surely, We have sent down to you (O Muhammad) the Book (this Quran) in truth that you might judge between men by that which Allah has shown you (i.e. has taught you through Divine Inspiration), so be not a pleader for the treacherous” (4:105).

So, once again, sciences and technology come together strengthening and supporting the truthfulness of this holy book in a clear and undoubtful way. All praise be to Allah.

References

Al-Shreef, A.A. (2009). Is the Quran Muhammad's invention? *Encyclopedia of miracles in Quran and Sunnah.* http://www.quran-m.com/firas/en1/index.php?option=com_content &view=article&id=294:is-the--quran-muhammads-invention- &catid=51:prophetical&Itemid =105.

Baraka,R., Salem,S., Abu-Hussien,M., Nayef.N.,& Abu-Shaban.W. (2014). Arabic Text Author Identification Using Support Vector Machines. Journal of Advanced Computer Science & Technology Research, 4(1), 1-11.

Bota, P., Wang, C., Fred, A., & Silva, H. (2020). Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet?. Sensors, 20(17), 4723.

Eder, M. (2010). Does size matter? authorship attribution, short samples, big problem. Digital humanities, 30(2), 132-135.

Ginsburg, J. R. (2009). The interrogative features. PhD thesis, the University of Arizona.

Holmes, D. I., & Kardos, J. (2003). Who was the author? An introduction to stylometry. Chance, 16(2), 5-8.

Hoshila, D.R.,Shireen, P. , & Sameerchand, P. (2016). Authorship Attribution Using Stylometry and Machine Learning Techniques. Journal of Intelligent Systems Technologies and Applications, 96(50), 113-125.

Islahi, A. (1989). Fundamentals of Hadith interpretation. Translated by T. M. Hashmi. Lahore: Al-Mawrid. Retrieved from www.monthly-renaissance.com /Download-Container.aspx?id=71.

Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval. 1(3),1-104.

Juola, P. (2012). Large-scale experiments in authorship attribution. English Studies, 93(3), 275–283.

Mills, D.E. (2003). Authorship attribution applied to the Bible. Master thesis, Texas Tech University.

Nasr, S. H. (2013). Encyclopædia Britannica Online. http://www.britannica.com/eb/article-68890/Quran, Last access in 2013.

Navinder, K., & Amandeep,V. (2015). Authorship Attribution of Punjabi Poetry using SVM Classifier, International Journal of Advanced Research in Computer Science and Software Engineering, 5(5).

Otoom, A.F., Abdallah, E.E., Hammad, M., Bsoul, M., & Abdallah, A.E. (2014). An intelligent system for author attribution based on a hybrid feature set, International Journal of Advanced Intelligence Paradigms,6(4).

Sayoud, H. (2012). Author Discrimination between the Holy Quran and Prophet's Statements. LLC journal, Literary and Linguistic Computing Journal, Oxford-University Press. Volume 27, No. 4, 2012, pp 427-444.

Sayoud, H. (2015). "A Visual Analytics based Investigation on the Authorship of the Holy Quran", 6th International Conference on Information Visualization Theory and Applications, pp. 177-181, Berlin, March 11-14, 2015.

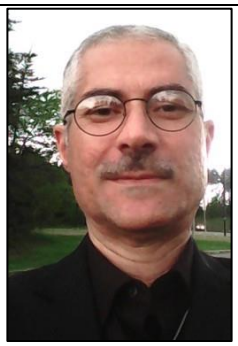
Shaker, K., Corne, D. (2010). Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis, Workshop on Computational Intelligence.

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, 6039, (pp.163–175) Publisher: Spie.

Stamatatos, E. (2009). A survey of modern authorship attribution methods, Journal of the American Society for Information Science & Technology, 60(3), 538–556.

Yule, G.U . (1938). On sentence length as a statistical, characteristic of style in prose with application to two cases of disputed authorship, Biometrica 30, 363-390.

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
	Ms. Hassina Hadjadj is a PhD student at the USTHB University who works in the field of NLP applied to theological documents.