AUTHOR IDENTIFICATION BASED ON A HYBRID FEATURE SET USING MACHINE LEARNING AND CLUSTERING TECHNIQUES

Hassina Hadjadj
Speech Communication & Signal Processing Laboratory
hadjadj.has@gmail.com

Abstract**—** Author identification of a document can be performed using computational or statistical method. In this paper, we try to identify the author of two ancient Arabic religious books dating from the 6th century: The holy Quran and the Hadith. Authorship identification consists in identifying the author of an anonymously document by using some techniques of Natural Language processing (NLP) and Artificial intelligence. In fact, each author has a unique writing style. Therefore, two series of experiments are undergone and commented. The first experiment deals with authorship identification of the two books using a Manhattan centroid distance and SMO-SVM classifier. Whereas, in the second experiment a Hierarchical Clustering is employed to identify the authors of the two books. Furthermore, three new features are combined to present the author. The results show good authorship identification performances with an accuracy of 100% corresponding to a clear authorship distinction between the two religious books.

Keywords**—** Authorship analysis; Natural language processing; Author identification; Religious books; Quran; Hadith; Text Classifiaction

I. Introduction

Authorship analysis is a typical problem in natural language processing. Authorship identification is a research field of stylometry that is used to identify the author of an anonymous document by using some techniques of text mining.

The area of authorship analysis has been researched for many years going back to the early 60s of works such as (Mosteller, 1964), where the authors were studying the important Federalist Papers case for solving an authorship claim by different authors. The main aim of authorship analysis is to study the characteristics of a piece of writing to draw conclusions about it.
In the recent years, practical applications for author identification have grown in several different areas such as email authorship (Holmes, 1998), plagiarism detection (Van Halteren, 2004) and forensic cases (De Vel, 2001).

Stylometry is part of a broader growth within computer science of identification technologies, including biometrics, cryptographic signatures, intrusion detection systems, and others (Madigan, 2005).
An interesting area in identification technologies is Biometric identification which is way to find or verify the identity of who we claim to be, by using physiological or behavioral characteristics (Jain, 2010).

As the human has physiological or behavioral characteristics; he has also linguistic features. Human usage of language, writing, set of vocabulary, unusual usage of words, and particular syntactic and stylistic traits tend to be stable. The big defy for authorship analysis is locating and learning from such features.

In fact, it is not clear which features of a text should be used to classify an author. So, the principal issue in computer-based author identification is to identify a set of features that represents the author's writing style. These are used to classify the authors of selected unknown texts. A different set of features can be used to identify authors; these include word-level, character-level, syntactic, semantic and lexical features (Stamatatos, 2009).

This research work deals with religious enigma, which has not been solved for fifteen hundred years (Sayoud, 2007) (Sayoud, 2012). Actually, 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. However, in such problems, it is crucial to use rigorous scientific tools and it is more important to interpret them very carefully. Hence, in this paper, we target an interesting area of authorship analysis, which is author identification of the holy Quran and Prophet's statements (Hadith) in order to check if really the Quran was not written by the Prophet Muhammad (i.e. it was only sent to him by God) (Sayoud, 2012).

In this purpose, our main aim is to extract a three new set of features called: interrogative words, Discriminative words and COST parameter, which are capable of identifying the author of an Arabic text. Concerning the classification methods, the different experiments of authorship identification are performed by using : Manhattan distance , SMO-SVM Classifier and hierarchical clustering on the different texts.

The rest of this paper is organized as follows: section 2 presents related works, section 3 gives a description of two religious books to be compared, in section 4, we present the authorship identification methodology. Section 5 describes the experimental results and finally, section 6 concludes the paper.

II. Related works

Many studies have been reported during the last years, where many debates were reported and several types of features and techniques were proposed too.

For instance, Stamatatos conducted a study of the latest advances in automated approaches used in authorship attribution (Stamatatos, 2009). He examined the characteristics of these approaches for text representation and text classification, and also the evaluation criteria end methodologies used in author identification studies. The survey distinguishes different types of stylometric features to quantify the writing style including character features, lexical features, syntactic and semantic features.

In 2012 Shaker et al. used a hybrid method of evolutionary search and LDA approach (Shaker, 2012). In this survey he investigated the usage of function words that are specific words which are used by the writer in distinct way and which may or may not relate to the subject matter. The approach was tested on Arabic and English documents.
recently, a plethora of models more familiar to machine learning practitioners than linguists such as support vector machines, neural networks, latent Dirichlet allocation, decision trees have been applied to different types of features with success (Stamatatos, 2014) (Kim, 2014) (Sayoud, 2015).
Ouamour et al. employed several character N-grams (Ouamour, 2013). The authors examined the authorship of Arabic books written by ten Arabic travellers. Different types of features were used such as character, character-bigram, character-trigram and character-tetra gram. For the classification, they used Stamatatos distance, Manhattan distance, Multi-Layer Perceptron (MLP) and Support Vector Machines (SVM).

Sayoud presented a series of author discrimination experiments between the holy 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 are different. Later on, he published another article describing an experiment of author discrimination between the holy Quran and Hadith by using a hierarchical clustering. Results were interesting since they sharply showed two main clusters representing the two corresponding authors: Quran author and Hadith author.

Seroussi et al. use authorship attribution of informal text such as e-mails with topic modelling (Seroussi, 2014). Disjoint Author-Document Topic (DADT) model was suggested that projects authors and documents to two disjoint topic spaces. Latent Dirichlet Allocation (LDA), Author-Topic (AT) and DADT models are implemented on formal as well as informal.

In 2015 Sayoud presents an experiment of author discrimination between the holy Quran and Hadith by using a hierarchical clustering (Sayoud, 2015), where seven types of NLP features are extracted. Results were interesting since they sharply showed two main clusters representing the two corresponding authors: Quran author and Hadith author.

III. Brief Description of the two old religious Books

Herein, we will give a brief description of the two religious books namely: the holy Quran and Hadith.

A. Holy Quran Description

The holy Quran (author: God (Allah)) is considered as the divine book of Islam (Juola, 2012). The Quran is written by Allah (God) and only sent down to his Prophet Muhammad fourteen centuries ago. This divine book has been delicately conserved by the different scholars over the time. The holy Quran considered as the first reference of Islam since it contains the authentic speech and statements of God (Allah).

Fig. 1. *Old pages of the Quran*

B. Description of the Hadith

The Hadith (author: the Prophet Muhammad) contains the statements of the Prophet Muhammad in different situations (Hashmi, 2012). 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 Sahih EL-Bukhari of Hadith book, which is considered as one of the most confident book of the Hadith.

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

IV. Author identification method

The author identification process consists of five steps namely: Document collection and preprocessing, Document segmentation, features extraction, classification and finally, evaluation (Fig. 3).

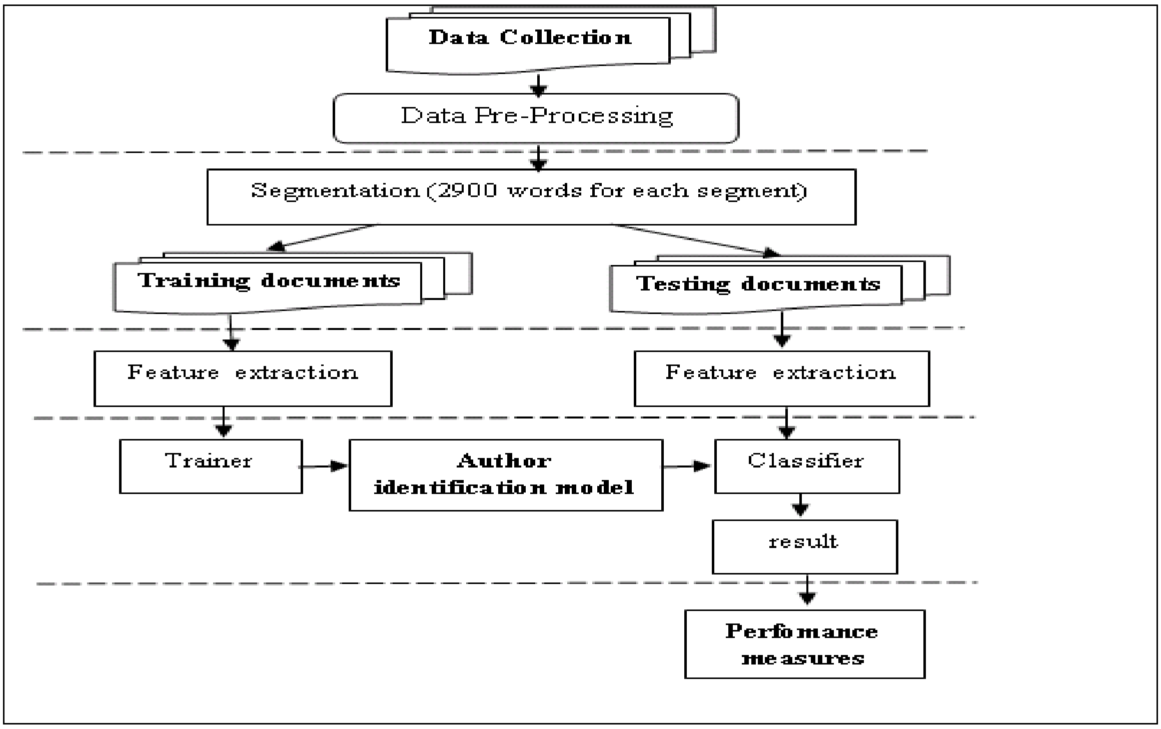

Fig. 3. *Author identification method*

A. Data Pre-processing

Data pre-processing is a very important step in authorship attribution. Text documents in their original form are not appropriate for learning. They must be converted into a suitable input format.

In order to obtain the same structural data and improve classification accuracy, punctuation marks diacritics, numbers and non Arabic letters are removed from text documents. After that, each text document is formatted according to UTF8 format.

B. Document segmentation

A text segmentation is used in order to construct individual texts with the same size. In fact, when comparing two books with various sizes, it is difficult to know if a specific part of the book is similar to another one or different.

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.

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 an accurate authorship identification. So, we decided to choose a size of 2900 words per segment.

Table I summarizes the size of the two books in terms of tokens and number of segments by book.

TABLE I. SIZE OF THE TWO BOOKS

Book	**Total number of tokens in the book**	**Number of segments in the book**
The holy Quran	87341	29
Sahih EL-Bukhari of Hadith	23068	8

The corpus is divided into 2 sets, the training set and the testing one.

The two books have different sizes; we cannot exceed 7 segments, because Sahih EL-Bukhari of Hadith book has only 8 segments. So, we choose 7 text segments of the Quran and 4 segments of the Hadith for the training step and the remaining text segments are used during the testing step.

C. Features extraction

The first step in any classification problem is the features extraction. After collecting the two religious books, we analyzed these books and ran our feature extractor to extract important information in these books. As we presented in the introduction, three original features are proposed: interrogative words (IntW), Discriminative Words (DisW) and COST parameter.

1) Interrogative Words (IntW): In Arabic, there are two ways of asking a question that needs a yes/no answer. These two particles are هَلْ /hal/ and the hamza أ /aa/ which are both equal to all the auxiliary verbs used in English to ask yes/no questions. These question particles should always start the question sentence.

In practice, we noticed that the particle hamza is very commonly used in the Quran book (أفمن, أأنت, أفرأيت, أأنتم…..) ; whereas, in the Hadith, this particle is rarely used.

2) Discriminative words (DisW) : The discriminative words are some words that are very commonly used in only one of the books. In practice, we noticed that the words: الذين (in English: THOSE or WHO in a plural form) and الأرض (in English: EARTH) are very commonly used in the Quran book; whereas, in the Hadith, these words are rarely used.

3) COST parameter: The COST parameter is a cumulative distance measuring the similarity between the ending of one sentence and the ending of the next and the previous ones in the text. It gives an estimation measure on the poetic form of a text. In fact, when poets write a series of poems, they make a termination similarity between the neighboring sentences of the poem, such as a same final syllable or letter. That is, the COST parameter estimates the similarity ratio between successive sentences in term of ending syllables.

D. *Classification methods*

Three different classification methods are employed namely: Manhattan Centroid distance, Sequential Minimal Optimization based Support Vector Machine (SMO-SVM) and Hierarchical Clustering.

1) Manhattan Centroid distance: This distance is very powerful in text mining. The corresponding distance between two vectors X and Y is given by formula (1).

$$d_{x,y} = \sum_{i=1}^{n} |X_i - Y_i| \qquad (1)$$

Where n is the length of the vector.

In this research, 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. Manhattan distance is simple to implement and very efficient for text classification.

2) Sequential Minimal Optimization based Support Vector Machine (SMO-SVM): In machine learning, support vector machines (SVM) 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 gap 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.
Concerning the Sequential Minimal Optimization (SMO) algorithm, it is used to speed up the training of the SVM (Keerthi, 2001).

3) Hierarchical Clustering: Hierarchical clustering is a method of cluster analysis which searches to construct a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. In our case, we used the Agglomerative clustering with an Euclidean distance measure. In order to use the agglomerative algorithm, we need to calculate the Euclidean distance between any two text segments, to find the distance matrix of our data set.

E. Evaluation

The performance giving by Manhattan distance and SMO-SVM classifier is measured in terms of score of good authorship identification which is calculated as the ratio of the number of true attributions over the total number of testing segments.

$$Score\ of\ good\ authorship\ identification = \frac{Number\ of\ correctly\ classified\ segments}{Total\ \text{number}\ of\ tested\ examples} \qquad (2)$$

V. Experimental Result and Experiment

As mentioned previously, Two Arabic religious books are investigated and analyzed in order to identify the author of the text documents. Furthermore, we recall that three original NLP features and several methods of classification are used in the experiments of authorship identification.

A. Experiments of Author Identification using Manhattan centroid distance and SMO based Support Vector Machines

This investigation performs a segmental analysis on the two Arabic religious books: Quran and Bukhari Hadith, for the task of authorship identification. The analysis of the different text segments is performed by using 23 features (combination of our features) and 2 type of classification methods (Manhattan Centroid distance and SMO-SVM classifier) where these two classification algorithms are very using in text mining.

Firstly, we recall that the dataset is divided into two sets: training data and testing data. So, there are 37 different text segments of about 2900 words each, consisting of 8 Hadith segments and 29 Quran segments. 4 segments of the Hadith and 7 other segments of the Quran are used for the training step and the remaining segments (4 Hadith segments and 22 Quran segments) are used for the testing step. Therefore, there are 37 different segments to identify according to 2 referential Authors (Quran Author or Hadith Author).

Table II. Shows the score of good authorship identification of the classification algorithms using the proposed features set above mentioned.

TABLE II. SCORE OF GOOD AUTHORSHIP IDENTIFICATION

Classification Algorithm	**Score of good authorship identification in %**
Manhattan Centroid Distance	100
SMO-SVM Classifier	100

By observing the above table, we can notice that all the 22 Quran segments are attributed to the referential "Quran Author" and all the 4 Hadith segments are attributed to the referential "Hadith Author". That is, the 26 different text segments are classified into 2 main classes: "Quran class" and "Hadith class", with 100% score of good authorship identification. From this result, we can conclude that the proposed features set are powerful for discriminating between the 2 classes with an identification error of 0%. And also we can deduce that the 2 religious books should have 2 different authors (or at least 2 different writing styles).

B. Experiments of Author identification using a Hierarchical clustering

In the second experiment a hierarchical clustering (Sayoud, 2012), using Euclidean distance, has been performed on all text segments by using the same features. Fig. 4 displays the resulting dendrogram.

Fig. 4 shows that the segments have been automatically divided into 2 principal clusters: "cluster Q" (in red) gathering all the text segments of the Quran and "cluster H" (in light blue) grouping all the text segments of the Hadith.

We can notice that the last clustering into one cluster (big line at the top) is inconsistent for two reasons: first, because the corresponding distance of this last cluster is more than 2.5, which is relatively large; and second, because we do not retrieve any link between heterogeneous segments at all.

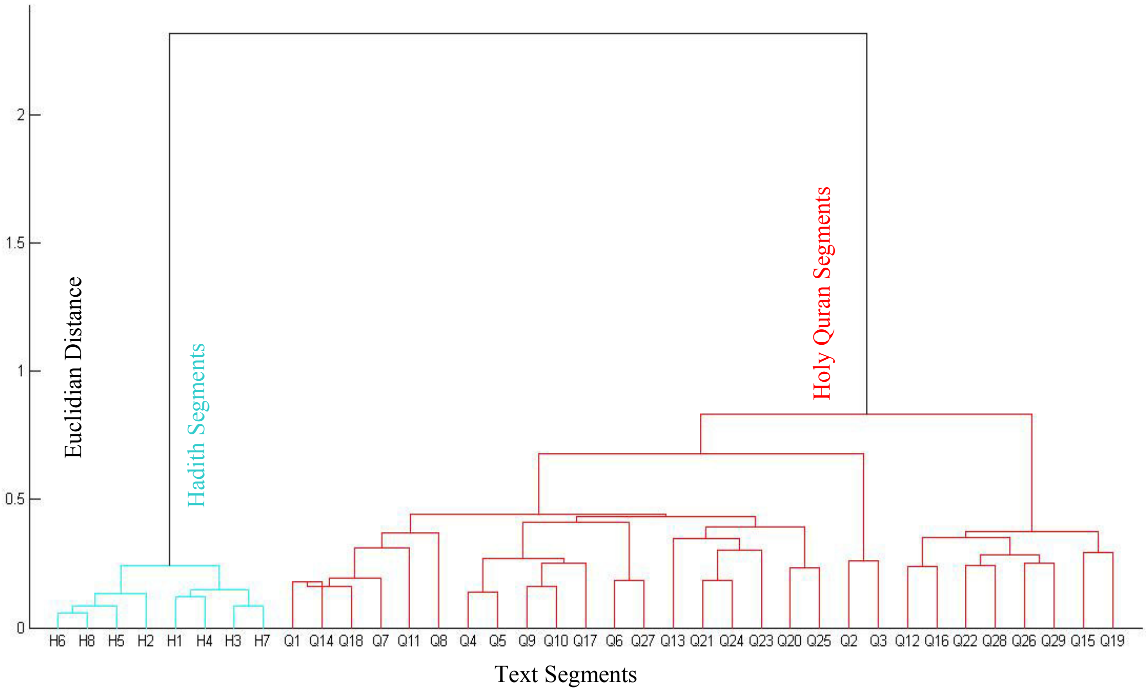

Fig. 4. *Automatic classification by the hierarchical clustering*

VI. Conclusion

As a continuation of a previous research work on the same topic (Sayoud, 2007)(Sayoud, 2012) , this research deals with the problem of authorship identification for two Arabic ancient religious books: the Quran and Sahih EL-Bukhari of Hadith . To have accurate author identification, it is very important to have a strong feature set that can discriminate between the different authors. For this purpose, a hybrid set of three types of writing features (total number of 23 features) are extracted from 37 text segments related to two different authors.

To show the efficiency of the proposed features set, we conducted several experiments:

- The first series of experiments consists in an authorship attribution task, which analyses the different text segments by using Manhattan centroid distance and SMO based Support Vector Machines.
- The second series of experiments performs a hierarchical clustering on the 37 text segments, in order to see how many possible clusters really exist and if the hypothesis of a unique author is possible.

After observing all the experimental results and since the two books appear to have the same genre and theme, it would be reasonable to deduce the following conclusions:

- The two books should have different authors (or at least two different author styles);
- All the Quran texts are grouped together in one cluster and all the Hadith texts are grouped together in another distinct cluster.

This implies that the two books (Quran and Hadith) are written by 2 different authors or at least with 2 different styles.

References

F. Mosteller and D.L. Wallace, "Applied Bayesian and Classical Inference: The Case of the Federalist Papers", Springer vol.13, no 10 ,pp.1-15, 1964.

D. I. Holmes, "The evolution of stylometry in humanities scholarship", Literary and linguistic computing, vol. 13, no. 3, pp. 111-117, 1998.

H. Van Halteren, "Linguistic profiling for author recognition and verification", Proc. of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 199 -205, 2004.

O. De Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics", ACM Sigmod Record, vol. 30, no. 4, pp. 55-64, 2001..

D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, and L. Ye. "Author identification on the large scale". In Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA), 2005.

E. Stamatatos, "A survey of modern authorship attribution methods", Journal of American Society for information Science and Technology, vol. 60, no. 3, pp. 238-556, 2009.

H. Sayoud, "Investigation of Author Discrimination between two Holy Islamic Books". IET (ex-IEE) Teknologia Journal. Vol. 1, Issue. 1, July 2010, pp. X-XII.

H. Sayoud, "Author Discrimination between the Holy Quran and Prophet's Statements". LLC journal, Literary and Linguistic Computing Journal, Oxford-University Press. Accepted and published on-line in 2012. Citation reference: doi: 10.1093/llc/fqs014, vol. 7, No. 4, 2012, pp 427-444.

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

K. Shaker, "Investigating Features and Techniques for Arabic Authorship Attribution". Submitted for the degree of Doctor Of Philosophy On compilation of research in the Department Of Computer Science School of Mathematics and Computer Science Heriot-Watt University, March 2012.

H. Sayoud, "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, 2015.

E.Stamatatos, W.Daelemans, B.Verhoeven, M.Potthast, B.Stein, P. Juola, M.Sanchez-Perez, and A.Barron-Cedeno, "Overview of the author identification task at PAN 2014". Analysis, 13:31, 2014.

Y.Kim, "Convolutional neural networks for sentence classification". International Conference on Empirical Methods in Natural Language Processing (EMNLP), October 25-29, 2014, Qatar.

S. Ouamour, H. Sayoud , “Authorship attribution of ancient texts written by ten Arabic travelers using character N-Grams”. in Proceedings of International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–5, 2013.

Y. Seroussi, I. Zukerman, and F. Bohnert, “Authorship Attribution with Topic Models”, Assoc. Comput. Linguist.,vol. 40, no. 2, pp. 269–310, 2014.

I. Ibrahim. “A brief illustrated guide to understanding Islam”. Library of Congress, Catalog Card, Texas, USA, 1996, pp. 97-754. Published by Darussalam, Publishers and Distributors.

T. M. Hashmi. Fundamentals of Hadith Interpretation – an English translation of ‘‘Mabadi Tadabbur-i- Hadith’’ by. Lahore: Al-Mawrid. www .monthly-renaissance.com/DownloadContainer.aspx? id¼71 (last accessed in 2012).

S.S. Keerthi, S.K. Shevade, C. Bhattacharyya “Murthy, Improvements to Platt’s SMO Algorithm for SVM Classifier Design”. Neural Computation 13, pp. 637–649, 2001.

A. Jain, "Biometric Identification," .Communications of the ACM, vol. 43, pp.91-98, 2000.

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

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