MyJICT
Malaysian Journal of Information and Communication Technology
EISSN 0127-7812

Vol. 10, Issue 2, pp. 105-125 | DECEMBER 2025
DOI: https://doi.org/10.53840/myjict10-2-222

A NEW APPROACH OF AUTHORSHIP ATTRIBUTION BASED ON NORMALIZED WORD BIGRAM TRANSITION PROBABILITY: APPLICATION ON THE QURAN AND HADITH

Halim Sayoud

Laboratory of Speech Communication and Signal Processing, USTHB University, Algiers

halim.sayoud@gmail.com

Article history:
Submission date: 6 Oct 2025
Received in revised form: 13 Dec 2025
Acceptance date: 23 Dec 2025
Available online: 24 Dec 2025

Keywords:
Natural Language Processing, Text Mining, Computational Linguistics, Stylometry, New Features in Authorship Attribution, Original Probabilistic Features, Quran, Hadith.

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interest:
The author(s) have declared that no competing interests exist.

Cite as:
Sayoud, H. (2025). A New Original Approach of Authorship Attribution based on Normalized Word Bigram Transition Probability: Application on the Quran and Hadith. *Malaysian Journal of Information and Communication Technology* (MyJICT), 10(2), 105-125. https://doi.org/10.53840/myjict10-2-222

ABSTRACT

After several studies on the intrinsic characteristics of the authors' style, we noticed a thorough link between the author's style and the use of specific word bigrams: some bigrams were found to have transition probabilities that are specific to only one author. Thus, we propose the use of a new set of features based on the normalized Forward and Backward Probabilities. We called it Word Bigram Transition Probability (i.e., WBTP). The approach described in this paper is proposed for the first time to the knowledge of the author. It can be used in any task of authorship attribution with the same or different topics. Three evaluation experiments of authorship attribution are conducted: Evaluation on SIMSTYL corpus, on HAT corpus, and on a subset of the Guardian corpus. Furthermore, a specific application of author discrimination between two ancient religious books (i.e., the Quran, and Hadith) has been carried out using this new approach. A comparison of this new set of features with some state-of-the-art features has been made. Results showed a high accuracy of authorship attribution. Furthermore, the results have shown that this new feature is less sensitive to the topic and can then be used with document belonging to different topics. Concerning the application of Author discrimination between the Quran and Hadith, the results show that the score of discrimination is 100%, confirming once again that the two authors are different, and that the holy Quran (believed to be a Divine revelation) could not have been invented by the Prophet.

Introduction

An interesting factor, which is common to the whole humanity, is that people do not possess the same characteristics, or let us simply say, they are different. From this real fact, many researchers have proposed intelligent systems and scientific approaches to recognize human beings thanks to their physiological and/or behavioral features. Such systems rely on the uniqueness of some biological or behavioral characteristics of human beings, which enable for individuals to be recognized using automated algorithms (Drozdowski, 2020). One can quote the field of biometrics, for instance, where the fingerprints, iris, speech and many other features were widely proposed in the literature and even used in practice (Sayoud, 2011).

On the other hand, several non-biometric features had been investigated by some researchers of pattern recognition by providing interesting performances (Sayoud, 2011), such as key stroke dynamic, for instance (Li, 2022).

In the same order of ideas, the writing style is very specific to authors and has the advantage to be the unique characteristic that can remain available for thousands of years (i.e. texts can be preserved and memorized several thousands of years, as it is the case with ancient scriptures on parchments and stones). For example, even though the documents related to the Quran (known to be the Divine book of God) and the Hadith (known to be the statements of the Prophet) are dated from the 7th century (i.e. 609 - 632 CE), several ancient parchments that are dated to that period of time (thanks to radiocarbon analysis) show a clear image of the sacred text (Sayoud, 2018). This fact makes the author identification task possible, and consequently leads to a very motivating research domain to investigate (Sayoud, 2021).

By definition, the analysis of the writing style, also called stylometry, is a research domain of Natural Language Processing (NLP) that is used to identify the real author of a piece of text (Uddagiri, 2023), with many applications in practice, such as forensic linguistics (Alduais, 2023), religious disputes (Sayoud, 2012) or plagiarism detection (Yeshilbashian, 2022) for instance.

In this context, many features were proposed and many classification schemes were introduced. More particularly, several researchers used words, or more generally, vocabulary-based features, in authorship attribution, where one can quote the works of Mendenhall using sentence length counts and word length counts. One can also quote the vocabulary richness functions (Yule, 1944) or simply the use of function words (Kestemont, 2014; Sayoud, 2022), or even the use of the most frequent words as described by Burrows (Patton, 2021).

Other researchers tried using word bigrams and more generally word n-grams, as described in (Ouamour, 2013). Although word bigrams were employed, this type of feature appears sensitive to the text topic, for instance in some medical reports, one can often find the following word bigram "*Sore Throat*", which is very specific to the health topic but probably not to the author style.

In 2001 Khmelev investigated the probability transition between characters (Khmelev, 2001) by using a Markov chain model (Kang, 2018), however, his research work was not expanded to other features – Unfortunately, Khmelev died dramatically in 2004 (Memorial, 2020).

Hence, the use of normalized word bigram transition probability was not used before the present work, at least, to the knowledge of the author when he began to write this paper.

That is, in this new research work, we propose the use of a statistical feature based on the transition probabilities between successive words. Furthermore, we describe the use of this new probabilistic feature on three experiments of Authorship attribution (AA): in the first experiment, an evaluation of the proposed approach is conducted on the simulated text corpus SIMSTYL, in the second experiment, the new proposed features are evaluated on the HAT corpus with 100 authors, and in the third experiment, an experimental evaluation is made on a cross-topic subset of the Guardian corpus. Finally, a specific application of author discrimination (Sayoud, 2012) between two ancient religious books (i.e., the Quran, and Hadith) has been carried using this new approach.

The manuscript is organized as follows: in the second section, we give the description and details of our new approach; in the third section, we explain the probability computation procedure; in the fourth section, we explain the probability normalization procedure. In the fifth section, we present the different experiments of authorship attribution. In the sixth section we describe the application of authorship analysis on the Quran and Hadith and in the seventh section, we end our manuscript by a conclusion and discussion.

Description

Let us assume that we have a large text document T with a set of different words $W_{i\,(i=1..N)}$, where each word appears in the document with a specific probability denoted by p(W_i).

Similarly, les us suppose that the text T contains M different word bigrams $B_{j\,(j=1..M)}$, where each bigram B_j can be represented as a couple of 2 successive words, called prefix (P_j) and suffix (S_j), as follows: $B_j = [P_j, S_j]$.

In this context, we will denote the bigram probability by p(B_j), the prefix probability by p(P_j) and the suffix probability by p(S_j).

Note: In this theory, we assume that the text document is sufficiently large to provide significant bigram probabilities.

In this new approach, we define two probabilistic ratios: the Forward Probability and the Backward Probability.

The Forward Probability (denoted by FWPj) is defined by:

$$FWP_j = \frac{\mathrm{p}(B_j)}{\mathrm{p}(P_j)} = \frac{\mathrm{p}([P_j,S_j])}{\mathrm{p}(P_j)} \qquad (1)$$

Differently, the Backward Probability (denoted by BWPj) is defined by:

$$BWP_j = \frac{\mathrm{p}(B_j)}{\mathrm{p}(S_j)} = \frac{\mathrm{p}([P_j,S_j])}{\mathrm{p}(S_j)} \qquad (2)$$

These ratios have the particularity to be bounded between 0 and 1 and could be considered as conditional probabilities, namely: $\mathrm{p}(B_j \,/\, P_j)$ for the first one and $\mathrm{p}(B_j \,/\, S_j)$ for the second one.

For concreteness, suppose that we have a simple text in the following form:

"The old man was very happy. Do you know who was that old man? The man is a foreigner, but I was very happy and very proud to meet with him."

In the previous paragraph, taken as example, there are two frequent (i.e. repeated) word bigrams, namely: "*old man*" and "*very happy*".

Let us compute the corresponding frequencies of the first bigram (i.e. "*old man*"), with m representing the bigram frequency, n_1 representing the bigram prefix frequency and n_2 representing the bigram suffix frequency.

We get:

m=*Freq*(*old man*) = 2/N, n_1=*Freq*(*old*) = 2/N, n_2=*Freq*(*man*) = 3/N

where N represents the total number of words (N=31 in this example)

and *Freq* denotes the frequency operator.

Note that the terms prefix and suffix denote the first word and second word of the bigram, respectively.

That is, if we compute the relative frequency given by the ratio of the bigram frequency to the prefix frequency, in this example, and if we assume that the probability is equal to the frequency, we get the following *Forward Probability* (FWP):

$$FWP_{\textit{old man}} = \frac{p(B_{old\,man})}{p(P_{old})} = m/n_1 = \frac{2/N}{2/N} = 1. \tag{3}$$

On the other hand, if we compute the relative frequency given by the ratio of the bigram frequency to the suffix frequency, in the same example, and if we assume that the probability is equal to the frequency, we get the following *Backward Probability* (BWP):

$$BWP_{\textit{old man}} = \frac{p(B_{old\,man})}{p(S_{man})} = m/n_2 = \frac{2/N}{3/N} = \frac{2}{3} = 0.66. \tag{4}$$

Hence, we can summarize these results as follows:

$$FWP_{\textit{old man}} = 1 \tag{5}$$

$$BWP_{\textit{old man}} = 0.66 \tag{6}$$

Similarly for the second bigram (i.e. "*very happy*"), by computing the corresponding frequencies of this bigram, with m representing the bigram frequency, n_1 representing the bigram prefix frequency and n_2 representing the bigram suffix frequency. We get:

m=*Freq*(*very happy*) = 2/N, n_1=*Freq*(*very*) = 3/N, n_2=*Freq*(*happy*) = 2/N.

As previously, the computation of the forward and backward probabilities gives the following values:

$$FWP_{\textit{very happy}} = 0.66 \tag{7}$$

$$BWP_{\textit{very happy}} = 1 \tag{8}$$

As we can see, even though the two previous types of bigrams do have the same frequency (i.e. 2/N), they do not possess the same FWP and BWP probabilities. This discriminative aspect, not only can represent a specific feature for the author style, but it could even add further characterization on how every part of the bigram is used.

Probability Computation Procedure

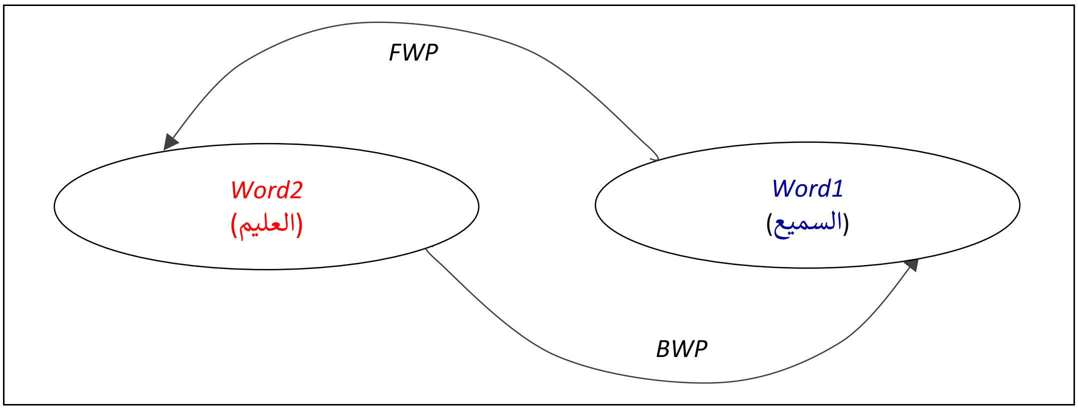
In the following, we describe the required steps to compute the Forward and Backward probabilities FWP and BWP of the different word bigrams.

Let us recall that, in our study, the term Word Bigram represents a couple of successive words as follows: [*Word1 Word2*], where *Word1* denotes the prefix (1st word) and *Word2* denotes the suffix (2nd word).

Now, let us take an example:

Suppose we are interested in the following Arabic bigram (السميع العليم), which is written from the right to the left and which is referred to by [*Word1 Word2*]. See Figure 1.

Figure 1: Graphical description of the FWP and BWP probabilities



The different steps of the procedure are:

- Compute the occurrence of *Word1*, denoted by n_1.
- Compute the occurrence of *Word2*, denoted by n_2.
- Compute the occurrence of the bigram [*Word1 Word2*], denoted by m
- Finally, compute the forward and backward probabilities thanks to formulas 1 and 2.

Probability Normalization Procedure (PNP)

In the context of transition probability (section 2), the sum of the *Forward Probability* and *Backward Probability* is not equal to one. In order to make the sum of the different probabilities equal to one, for a purpose of normalization, we have proposed the normalized forward and backward probabilities as follows:

$$NFWP = \frac{FWP}{FWP+BWP} \tag{9}$$

and

$$NBWP = \frac{BWP}{FWP+BWP} \tag{10}$$

where *NFWP* denotes the normalized *Forward Probability* and *NBWP* denotes the normalized *Backward Probability.*

In this way, we get:

$$NFWP + NBWP = \frac{FWP}{FWP+BWP} + \frac{BWP}{FWP+BWP} = \frac{FWP+BWP}{FWP+BWP} = 1. \tag{11}$$

This fact can be easily verified for the first example of section 2:

$$NFWP = \frac{FWP}{FWP+BWP} = \frac{1}{1+0.66} = 0.6024.$$

$$NBWP = \frac{BWP}{FWP+BWP} = \frac{0.66}{1+0.66} = 0.3976.$$

Then, by computing the sum, we get:

$$NFWP + NBWP = 0.6024 + 0.3976 = 1.$$

In this context, we notice that only one normalized transition probability is required (i.e. *NFWP* or *NBWP*), since the dual form is deducible from Equation 11.

Experiments of Authorship Attribution

We recall that in this research work a new set of features is proposed and employed - we called it WBTP (i.e. Word Bigram Transition Probability). In practice, we use the concatenation of the NFWP with the NBWP vectors to form the WBTP vector.

We evaluate the efficiency of this new probabilistic feature on three experiments of authorship attribution: in the first experiment, an evaluation on the simulated text corpus SIMSTYL is conducted, in the second experiment we make an authorship attribution on the HAT corpus, with 100 different authors, and in the third one, we evaluate the approach on a subset of the cross-topic Guardian corpus.

First Experiment: Evaluation on the Simulated Text Corpus SIMSTYL

The simulated text corpus (SIMSTYL) is composed of 2 different text groups, where each group is characterized by a specific different style. In a single group there are two different texts of the same style (Sayoud, 2022b).

This small corpus is represented in the following table and can be freely downloaded and used for a purpose of reproducibility.

Table 1: The simulated SIMSTYL corpus

	Simulated Author style 1		**Simulated Author style 2**	
	Text 1A	**Text 1B**	**Text 2A**	**Text 2B**
Text	aaa red xyz red ggg ggg red hat drk gyy red hat red hat zzz red kkk hat dee red hat frr red hat hhh red hat ppp yyy red	ccc red ryz red gug gpg red hat erk gzy red hat red hat zoz red kok hat dce red hat irr red hat huh red hat upp zyy red	qbc qef xuz zef glg gxg red hat irk ghy red hat red hat azz agc dqf frr dze red hat fri red hat pqp red hat hhh rrr rte	ubc tef iuz olg gig des red hat ipk ghf red hat red hat zkz anc dnf for see red hat fir red hat sqp red hat phh xbc xef
Document size	30 words	30 words	30 words	30 words
Occurrence of the word bigram “red hat”	6	6	6	6
NFWP and NBWP of the word bigram “red hat”	0.41 - 0.59	0.41 - 0.59	0.50 - 0.50	0.50 - 0.50

Two types of features are evaluated during this experiment of AA (Authorship Attribution), namely: word bigrams and WBTP.

The distance matrix of author style classification on SIMSTYL, using the WBTP with Manhattan distance, is given as follows. The used distance is equal to the Manhattan distance divided by its maximum value.

Table 2: Distance matrix of author style classification on SIMSTYL, using the WBTP with Manhattan distance

Text / **Text**	**1A**	**1B**	**2A**	**2B**
1A	0	0.99	1	1
1B	0.99	0	1	1
2A	1	1	0	0.99
2B	1	1	0.99	0

The distance matrix of author style classification on SIMSTYL, using word Bigrams with Manhattan distance, is given as follows. The used distance is equal to the Manhattan distance divided by its maximum value.

Table 3: Distance matrix of author style classification on SIMSTYL, using Word Bigrams with Manhattan distance

Text Text	1A	1B	2A	2B
1A	0	1	1	1
1B	1	0	1	1
2A	1	1	0	1
2B	1	1	1	0

The distance matrix of author style classification on SIMSTYL, using the WBTP with Spearman distance, is given as follows.

Table 4: Distance matrix of author style classification on SIMSTYL, using the WBTP with Spearman distance

Text Text	1A	1B	2A	2B
1A	0.000	1.259	1.260	1.260
1B	1.259	0.000	1.260	1.260
2A	1.260	1.260	0.000	1.260
2B	1.260	1.260	1.260	0.000

The distance matrix of author style classification on SIMSTYL, using word Bigrams with Spearman distance, is given as follows.

Table 5: Distance matrix of author style classification on SIMSTYL, using Word Bigrams with Spearman distance

Text Text	1A	1B	2A	2B
1A	0.000	1.216	1.216	1.216
1B	1.216	0.000	1.216	1.216
2A	1.216	1.216	0.000	1.216
2B	1.216	1.216	1.216	0.000

The performances of author style classification with the different experimental protocols are given as follows.

Table 6: Score of correct author style classification for all experiments on SIMSTYL

Feature	Manhattan distance	Spearman distance
Word Bigrams	50%	50%
WBTP (*proposed feature*)	100%	75%

By observing these results, one can say that conventional word bigrams were not able to discriminate between the two author styles, since the frequencies of their word bigrams are similar.

However, the WBTP feature shows a clear difference in styles between the two author styles (i.e., style 1 and style 2), where one can see a better classification in the distance matrices, leading to a better classification accuracy.

Second Experiment: Evaluation on the HAT corpus

This corpus is composed of 100 groups of Arabic texts that are extracted from 100 different Arabic books (Sayoud, 2021b). The books are written by 100 different authors and each group contains 3 different texts that are written by the same author, which means that each group belongs to only one author. The texts have a medium/short size: the average text length is about 1100 words per document and there are 3 documents per author, which corresponds to 300 documents in the total corpus.

Experimental protocol on the HAT corpus

In this experiment, we vary the number N of authors from 2 up to 100 authors and notice the performances of AA (Authorship Attribution) versus the number of authors.

Since every author has 3 different texts, in every iteration we try to identify one document from the M considered documents, where:

$$M=3N-1 \tag{12}$$

For the computation of the transition probabilities, all word bigrams are taken into account in most cases.

The identification approach is based on the Nearest Neighbour Classification technique using specific distances: Cosine distance, Manhattan distance or Spearman distance. The Validation method is based on the LOO (Leave One Out) cross validation technique.

As described previously, the used feature vector WBTP is obtained by concatenating the NFWP and NBWP vectors.

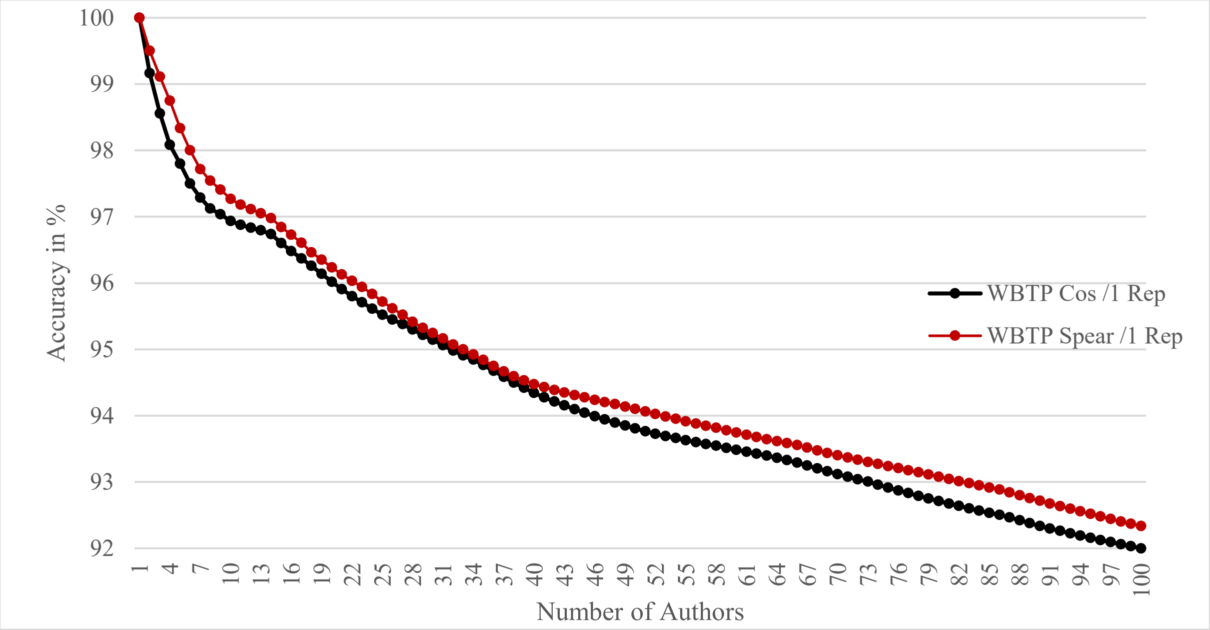
In Figure 2, we have presented the results of authorship attribution on the HAT corpus by varying the number of authors from 2 up to 100 authors, by using two distance measures (cosine and Spearmen distances), and by using the WBTP features. For the cross validation process, the whole data is decomposed into several folders of ***k*** authors, denoted by F_{ki} (i=1..N_k, where N_k is the number of folders with ***k*** authors). Then, in every folder, an LOO cross validation technique is applied during the identification process.

Finally, for every group of F_{ki} folders, the mean of the different accuracies, in a folder, are computed:

$$A_k = \underset{i}{mean}(A_{ki}) \tag{13}$$

As one can see in Figure 2, when the number of authors k increases from 2 authors to 100 authors, the accuracy decreases continuously. Although, Spearman distance is a bit better than cosine distance, the difference is inappreciable.

Figure 2: Authorship Attribution Accuracy on HAT corpus, using the WBTP.



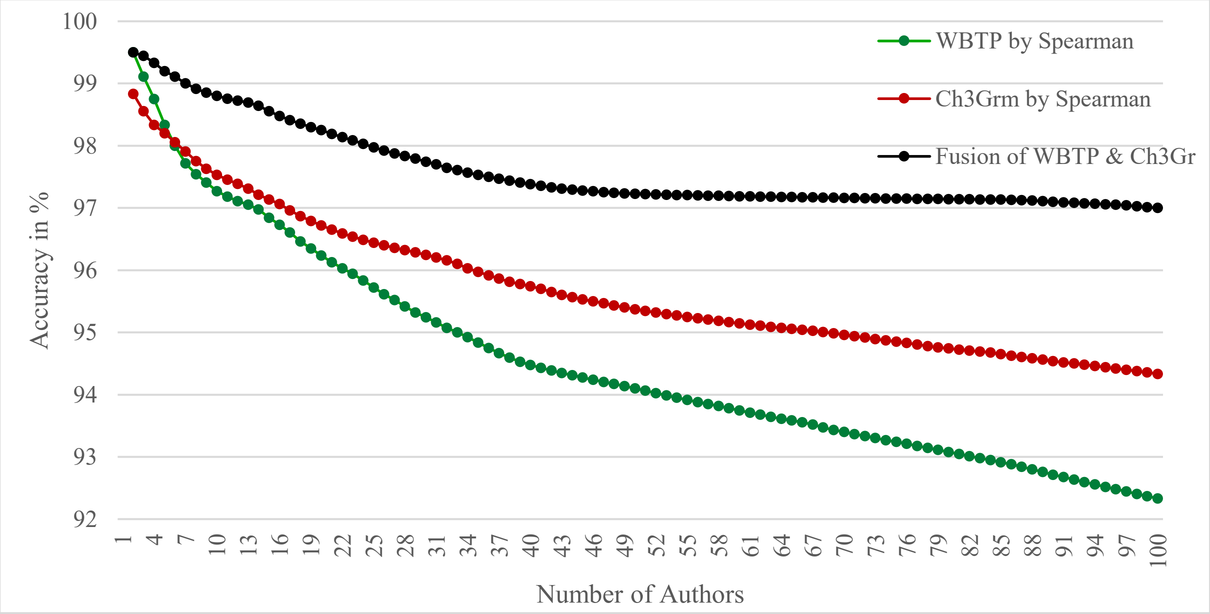
In Figure 2, we see that the WBTP is interesting in AA, since the corresponding accuracy is over 0.92 with 100 authors. Another important point to note is that the performances of AA decreases when the number of authors increases.

Comparison with state-of-the-art features

In order to evaluate the new set of probabilistic features a comparison with word bigrams and character trigrams is applied.

Figure 3 shows a comparison between WBTP and character trigrams (Ch3Grm). One notices that character trigrams perform better than WBTP when the number of authors is large, however for less than 7 authors, the WBTP is better. We also notice that the fusion between the two features gives a much higher score of identification: for instance, a score of 97% with 100 authors.

Figure 3: Authorship Attribution Accuracy on the HAT corpus, using WBTP, character trigrams (Ch3Grm) and the fusion between them.

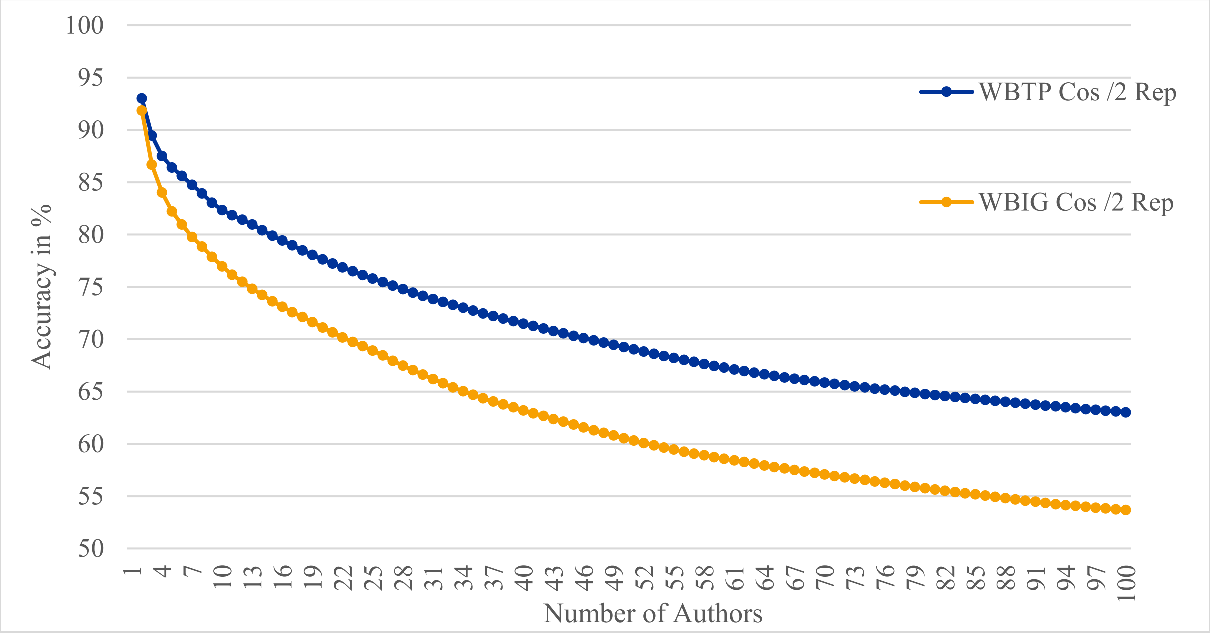


Another interesting point to note from Figure 3, is that the AA accuracy, for a single feature, decreases when the number of authors increases. However, in the case of the fusion, the accuracy tends to be constant (about 0.97) from 50 authors up to 100 authors. This fact makes the fusion more robust than the other features taken alone.

Comparison with Word Bigrams

Figure 4 shows a comparison between word bigrams (WBIG) and WBTP, where we kept only the bigrams that occur at least twice. One notices that the WBTP feature performs better than word bigrams. For example, with 100 authors, the difference in accuracy is more than 0.09. Also, as previously, one can see in Figure 4 that when the number of authors increases from 2 authors to 100 authors, the accuracy continuously decreases.

Figure 4: Authorship Attribution Accuracy on the HAT corpus, using WBTP and Word Bigrams (WBIG) with cosine distance. Note that in this particular experiment, the algorithm keeps only bigrams that are repeated at least twice.



Third Experiment: Evaluation on a cross-topic subset of the Guardian corpus

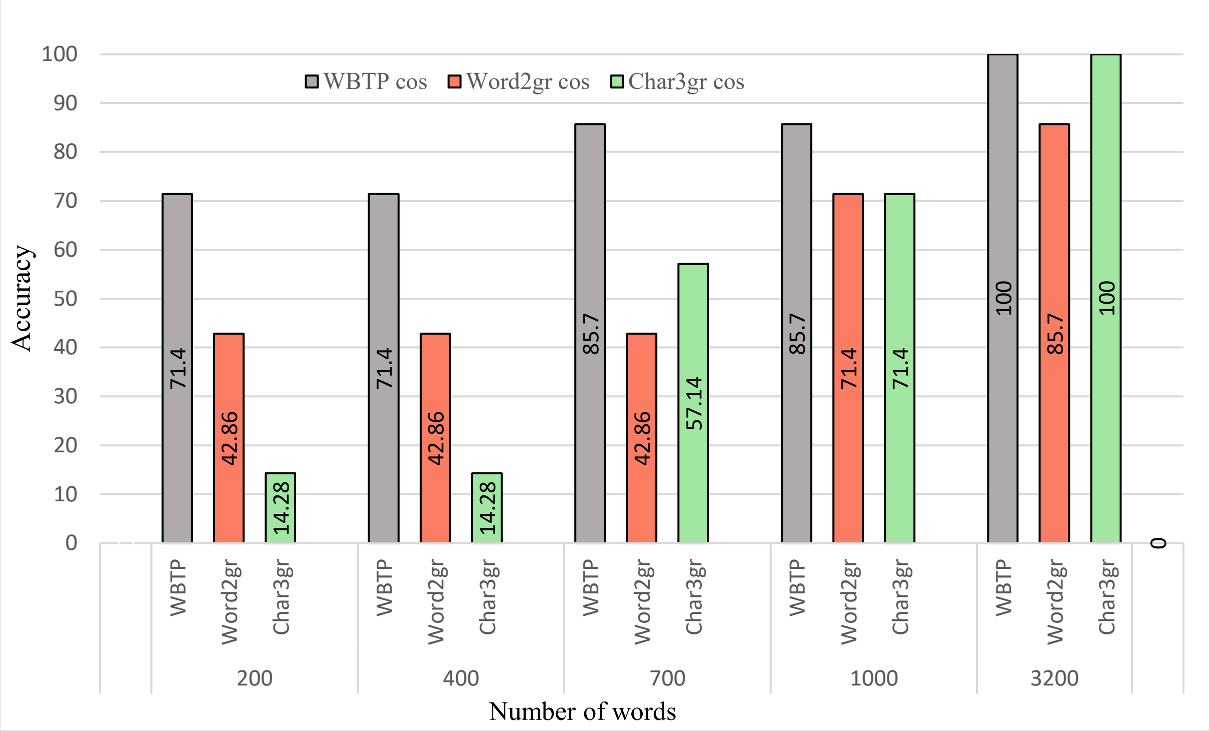
In this experiment, we used a subset of the Guardian corpus (Stamatatos, 2013) (Sidorov, 2019). This corpus is composed of texts published in The Guardian daily newspaper written by 13 authors in one genre on four topics. It contains opinion articles (comments) about World, U.K., Culture, and Politics, where we kept only the following topics: Politics and Society. Furthermore, we kept only the authors who possess enough data, which resulted in a set of 7 authors only. Thus, in this subset, each author has 2 documents related to 2 different topics: Politics and Society.

To perform cross-topic experiments of authorship attribution, we set up our data as follows: the politics documents are used for the training, while the society ones are considered as testing data.

Each politics document has a fixed size of 4500 word per document (i.e., training data), but in the society folder (i.e., testing data), there are 5 different document sizes for each author: a document of 4500 word, a document of 1000 words, a document of 700 words, a document of 400 words and a document of 200 words.

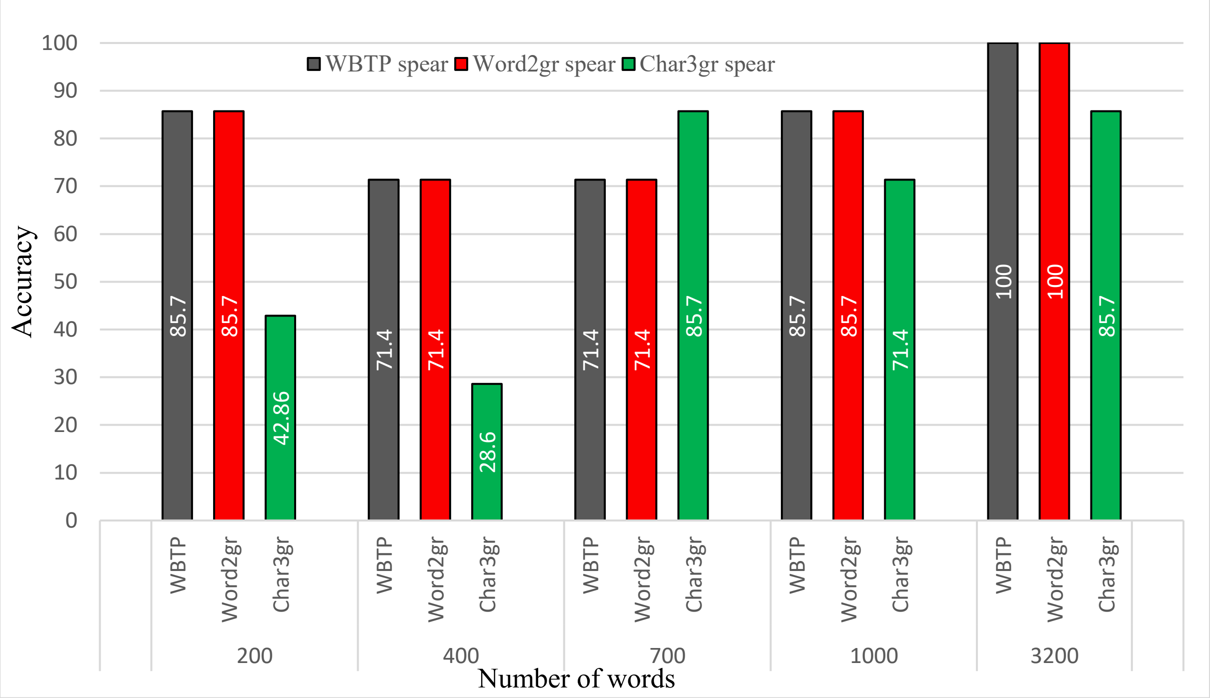
Three types of features are used: WBTP, word bigram and character trigram. Furthermore, three metric types are employed for the classification by using the nearest neighbour technique: Cosine distance, Spearman distance and Manhattan distance.

Figure 5: Authorship Attribution Accuracy on the Guardian subset using cosine distance.



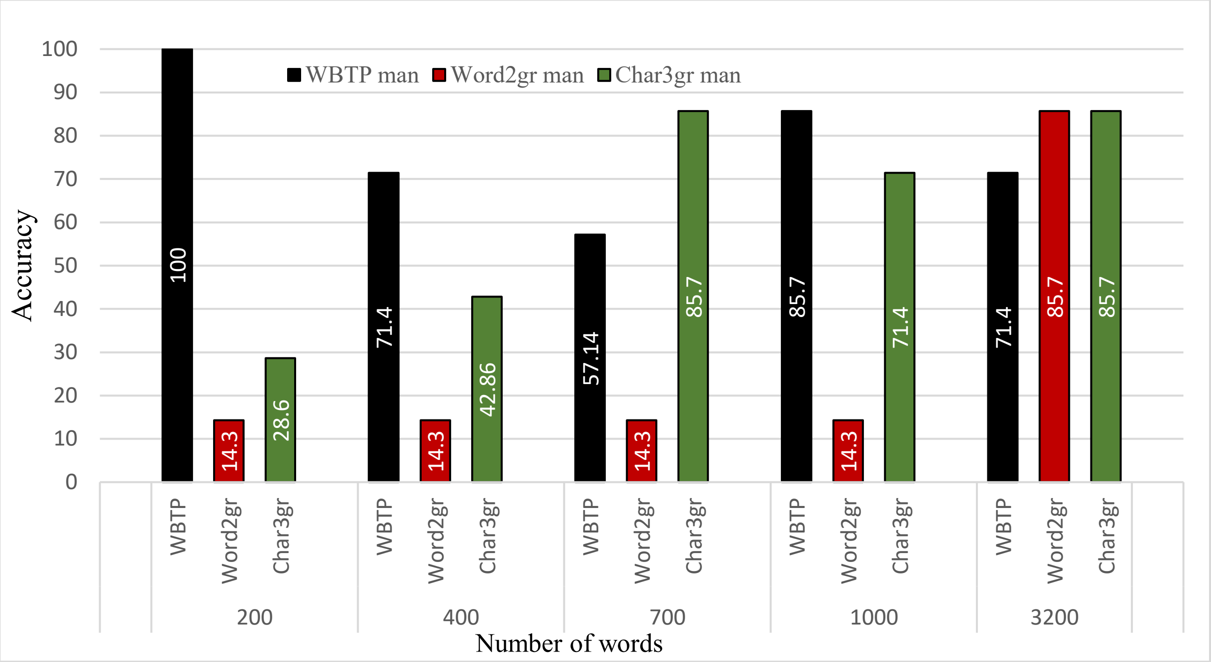
By using the cosine distance and according to Figure 5, the best feature in the Guardian subset (with cross-topic AA) is the proposed WBTP.

Figure 6: Authorship Attribution Accuracy on the Guardian subset using Spearman distance.



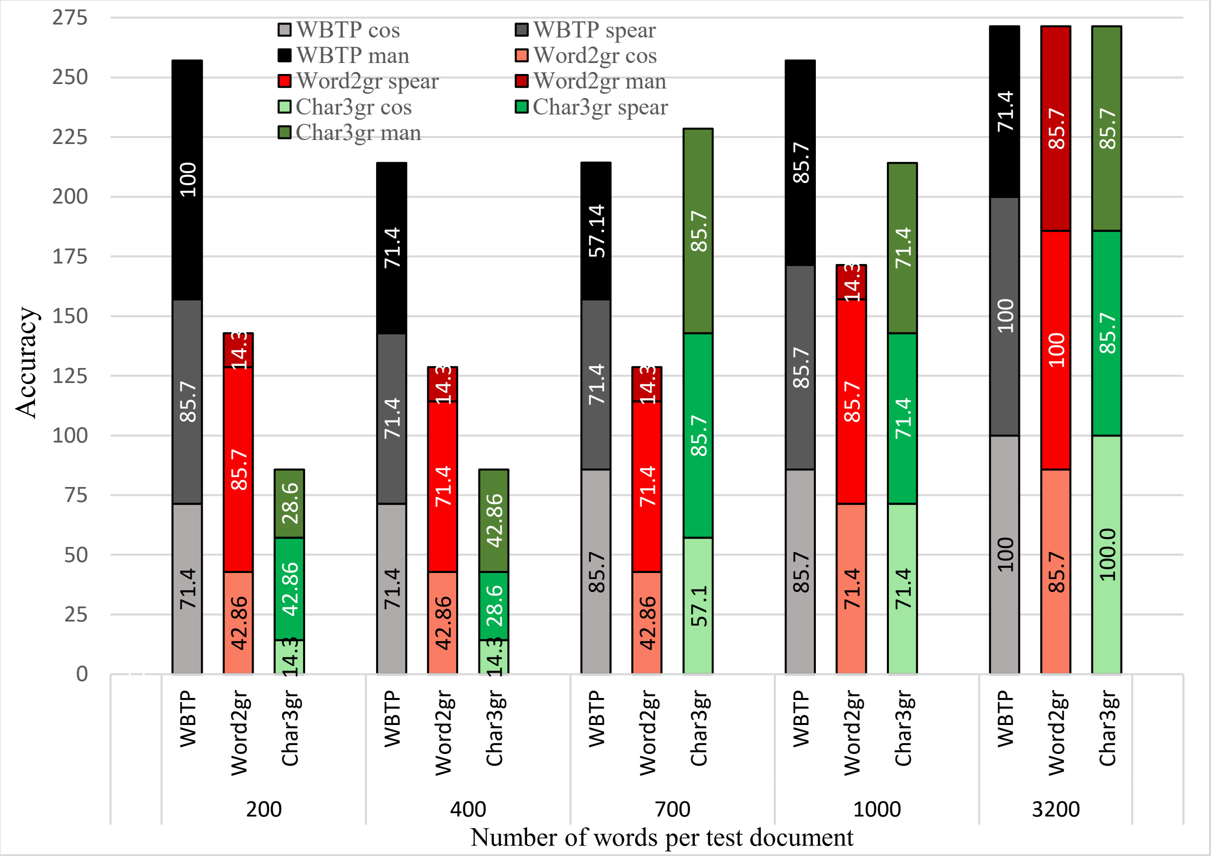
By using the Spearman distance and according to Figure 6, the best features in the Guardian subset (with cross-topic AA) are the WBTP and Word bigram.

Figure 7: Authorship Attribution Accuracy on the Guardian subset using Manhattan distance.



By using the Manhattan distance and according to Figure 7, the best features in the Guardian subset (with cross-topic AA) are the WBTP and Character trigram. However, for very short documents (200 words and 400 words per document), the best feature appears to be by far the proposed WBTP.

Figure 8: Authorship Attribution Accuracy on the Guardian subset with all distances.



In Figure 8, we can see a cumulative representation of the accuracy for the different features and with the different distances. It is clear in this figure that the proposed WBTP feature represents the best feature among the used ones, especially for short documents, where there is a great difference in performances between this one and the other features. These results not only show that the WBTP feature is interesting in cross-topic AA, but also that it can be very efficient with short documents.

Application of Authorship Analysis on the Quran and Hadith

Experiments of Author Identification on the Quran and Hadith

Herein, the main task consists in performing an author discrimination on two ancient Arabic books: Quran and Hadith (Sayoud, 2012; Sayoud, 2015; Sayoud, 2021; Sayoud, 2022). The experiments of author identification on the Quran and Hadith are made under the following protocol:

- There are theoretically 2 authors: the Author of the Quran and the Author of the Hadith.
- There are 37 text segments of the same size, namely: 29 text segments from the Quran and 8 text segments from the Hadith.
- The WBTP is computed and used as feature.
- Two types of distances are employed: Cosine distance and Spearman distance.
- Two classification techniques are used: the nearest neighbour technique and the centroid technique.
- The LOO cross-validation technique is employed.

The different results of identification on the Quran and Hadith segments are displayed in Table 7, where one can see that for every experiment the score of correct attribution in 100%. These clear results show that the Authors of the two investigated books are different and confirm the previous results on the topic (Sayoud, 2012).

Table 7: Authorship Attribution Accuracy on the Quran and Hadith using the WBTP.

	Accuracy in %			
Classification technique	Centroid based technique		Nearest neighbour technique	
Type of distance	Cosine distance	Spearman distance	Cosine distance	Spearman distance
AA Accuracy	100%	100%	100%	100%

Experiments of Clustering on the Quran and Hadith

The experiments of clustering are conducted on the Quran and Hadith, with a set of 37 text segments of the same size, namely: 29 text segments from the Quran and 8 text segments from the Hadith. We used 5 different clustering approaches, namely: Hierarchical clustering, PCA based clustering, FCM based clustering, GMM based clustering and Sammon mapping.

Hierarchical clustering

Hierarchical clustering is a method of cluster analysis, which seeks to build a hierarchy of clusters. It has been widely used in NLP as one quote the works of (Lupea, 2021).

In general, there are two types: Agglomerative clustering, which is a bottom-up algorithm, and divisive clustering, which is a top-down algorithm. In our case, we used the first clustering type with two types of distances: Cosine distance and Spearman distance.

The resulting linkage of the different documents is called "Dendrogram". It represents the different possible clusters in a graphical way. By observing the dendrogram, it will be possible to estimate the actual number of clusters and the corresponding documents for each cluster, since all similar documents should be linked together into a single cluster. Finally, every isolated big cluster should represent an author style and then one author.

The following Figures, 9 and 10, represent the hierarchical clustering obtained with the WBTP feature, by using the cosine distance and Spearman distance respectively. By observing the different dendrograms of those figures, we can observe two separate clusters, one cluster in red on the left and another one in blue on the right. These results clearly show that the two investigated documents (i.e., Quran and Hadith) have two different author styles, and then consequently their corresponding Authors should probably be different.

Figure 9: Hierarchical clustering by cosine distance using word bigram transition probabilities

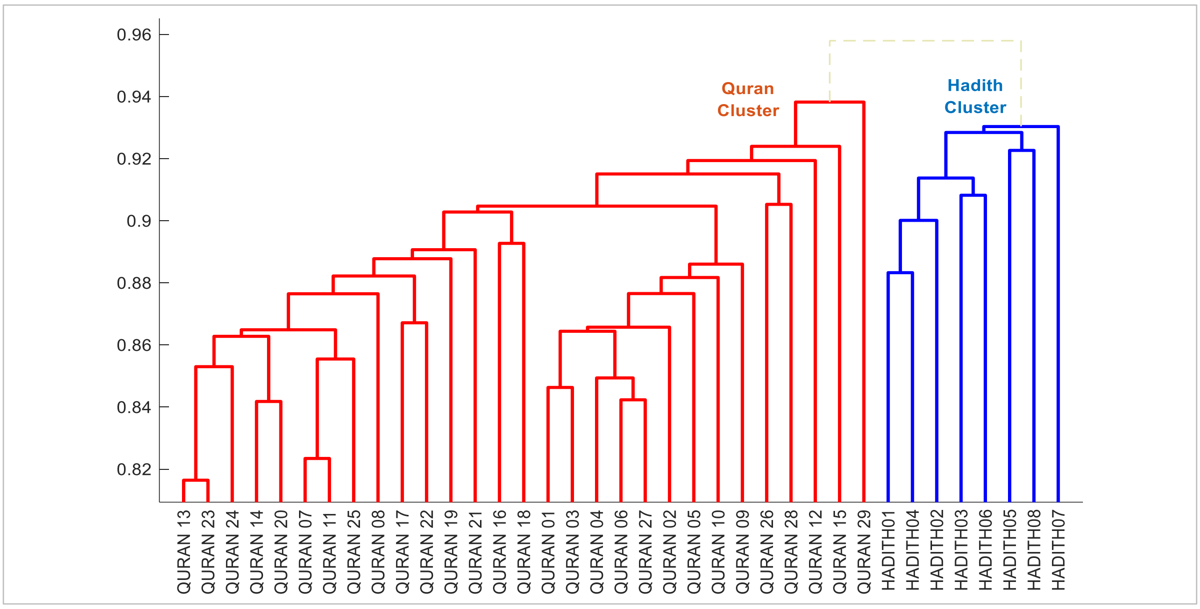
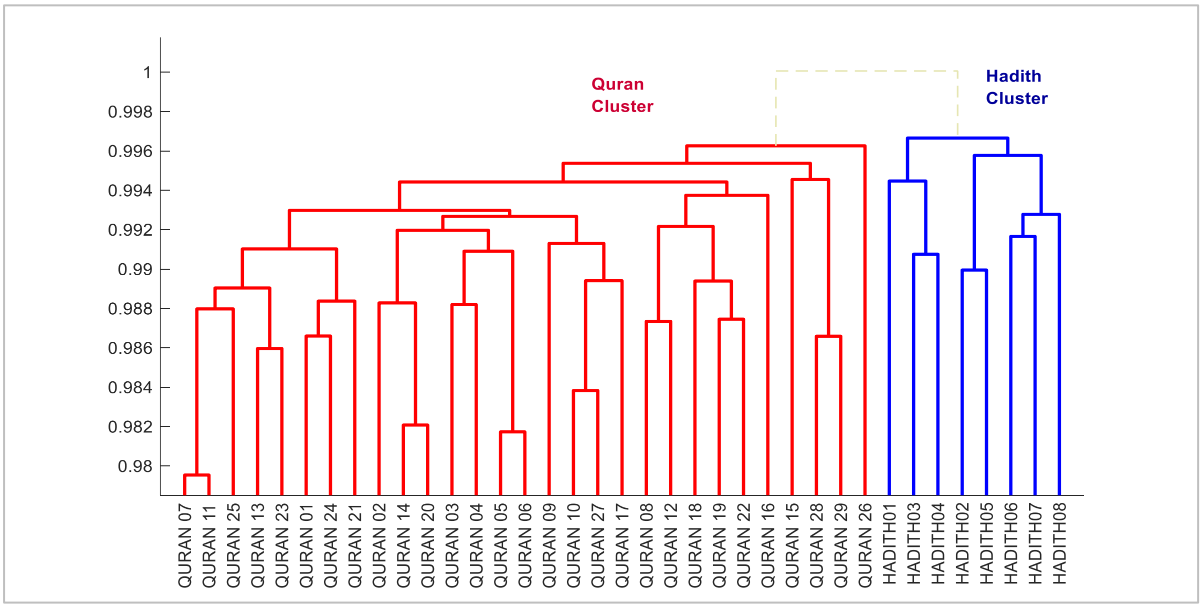


Figure 10: Hierarchical clustering by Spearman distance using word bigram transition probabilities



PCA based clustering

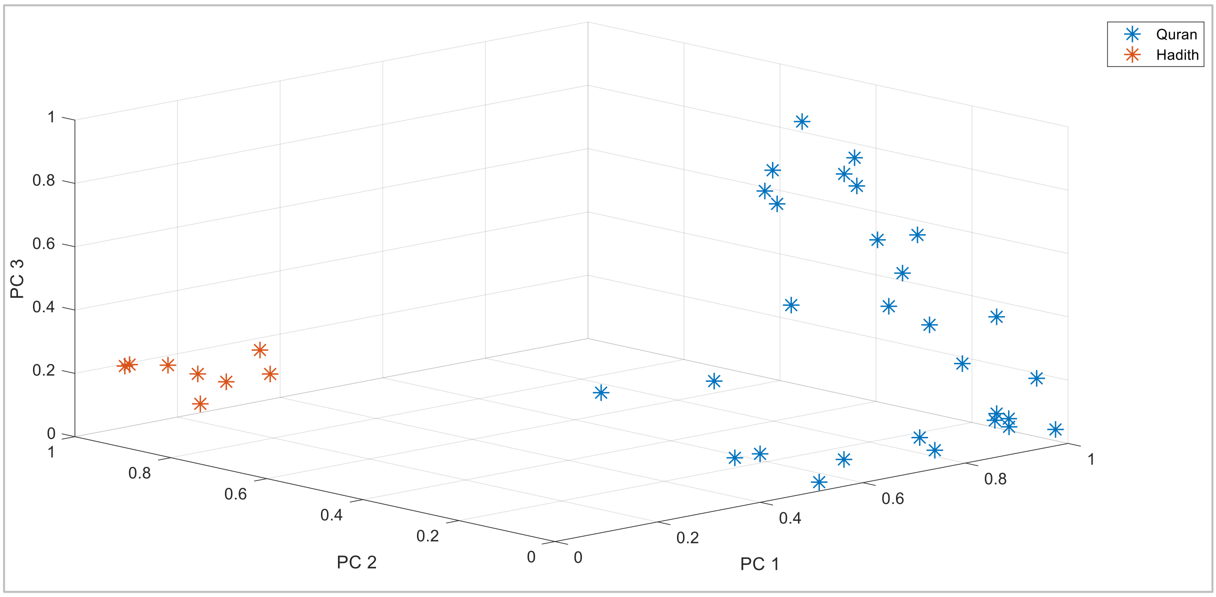
PCA or Principal component analysis provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the hidden, simplified dynamics that often underlie it (Shlens, 2003).

PCA is mathematically defined as a linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on.

Other variants of the PCA algorithm do exist, such as the FPCAC (Sottile, 2021). In this variant, the algorithm looks for clusters of functions according to the direction of largest variance, outlined by the functional PCA scores, assigning events to the best cluster based on a proper distance measure (Sottile, 2021).

The use of PCA is advised in complex data analysis, when the most important features are not known in advance. And by reducing the dimensionality to a lower consistent one, the visual data analysis becomes usually easier and clearer.

Figure 11: PCA representation of the different text segments of the Quran and Hadith using word bigram transition probabilities.



In Figure 11, corresponding to a 3 dimensions PCA, one can see that there are 2 separated groups of documents, a blue one on the right, grouping all Quran segments, and a red one on the left, grouping all Hadith segments. This fact suggests that the 2 authors’ styles are different.

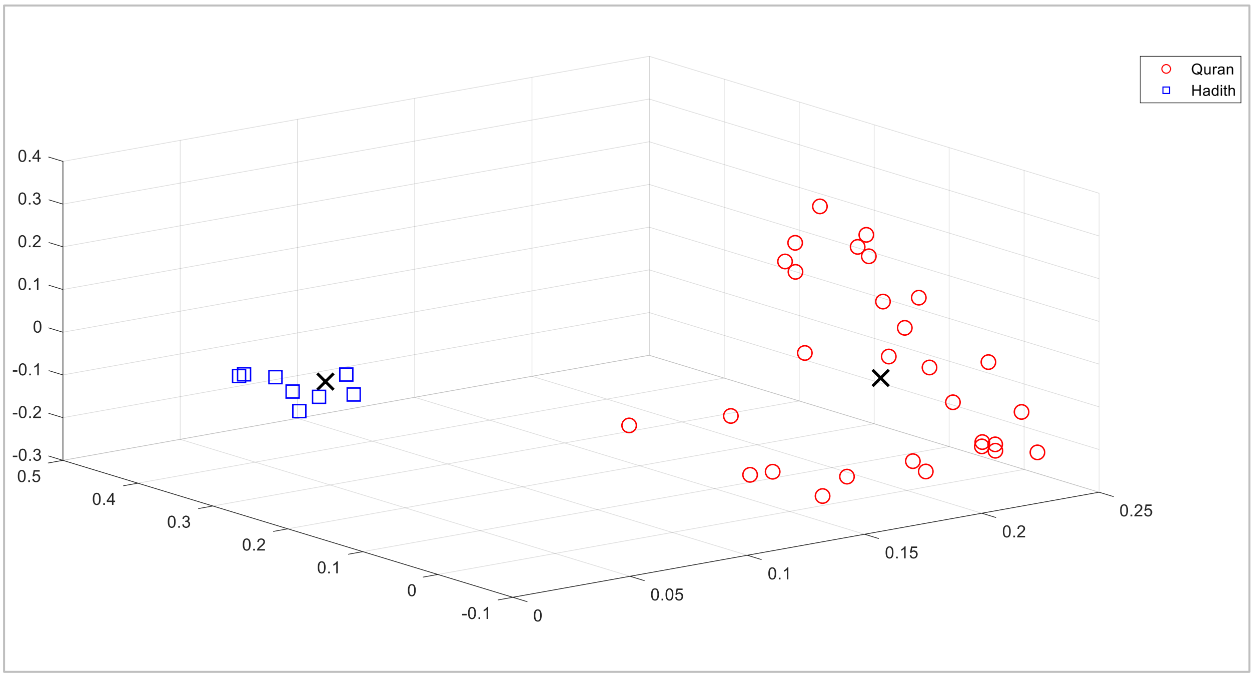
FCM clustering

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not hard but fuzzy in the same sense as fuzzy logic (Suganya, 2012).

In fuzzy clustering, every point has a degree of belonging to clusters, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. With Fuzzy C-Means (FCM), the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster. In 2D or 3D, Fuzzy C-mean can provide an interesting graphical representation of the different samples and the corresponding clusters to which they should belong.

The Fuzzy C-Means clustering has provided the following 3D representation (see Figure 12), where we can observe two separated clusters with two separated centroids, one cluster in red surrounding the right centroid and another one in blue surrounding the left centroid. This result shows that the two investigated documents (i.e. Quran and Hadith) have two different author styles.

Figure 12: FCM clustering after PCA reduction (with the corresponding centroids: "x" symbols) of the Quran and Hadith segments.



GMM based clustering

Mixtures of distributions have provided a mathematical-based approach to many random phenomena (McLachlan, 2003). Due to their efficiency, finite mixture models have received increasing attention over the years, from a practical and theoretical point of view. The GMM (Gaussian Mixture Model) based clustering is an unsupervised learning that finds the unknown parameters of marginal GMM distribution and responsibilities for each data (Jovanović, 2021).

In case of multivariate data, the multi-variate normal components are recommended because of their wide applicability and computational convenience. They can be fitted iteratively by maximum likelihood and the expectation maximization method.

In a normal mixture model-based approach, one assumes that the data are from a mixture of a specified number g of multivariate normal densities in some specific proportions pi$_1$, …, pi$_g$, that is, each data item is taken to be at realization of the mixture probability density function (see equation 14).

$$f(u;\Psi) = \sum_{i=1}^{g} \pi_i \phi(y;\ \mu_i, \Sigma_i) \quad (14)$$

where $\phi(y;\ \mu_i, \Sigma_i)$ denotes the p-variate normal density probability function with mean μ_i, and covariance Σ_i.

Here the vector Ψ of unknown parameters consists of the mixing proportions π_i, the means μ_i and the covariance matrices Σ_i.

When the mixture model is fitted, a clustering of the data into g clusters can be obtained in terms of the posterior probabilities of component membership for the data. The assignment of the data into g clusters is performed by assigning each data point to the component to which it has the highest posterior probability of belonging (McLnchlan 2001).

The GMM based clustering is performed after PCA reduction into the 2 most important components. That is, two types of visualizations are provided: a 2D representation (with those two components) and a 3D representation including the probability density function as third component (see Figures 13 and 14).

In both figures, we notice that the different text samples have been clustered into 2 main groups: Quran cluster, on the left side, gathering all the Quran texts and a Hadith cluster, on the right side, gathering all Hadith texts.

In the 2D representation, the Gaussian mixtures are represented by different ellipsoids surrounding the two clusters, while in the 3D representation, the Gaussians are more visible since they are represented in form of 3D Gaussians surrounding the different clusters.

While, the first representation is sharper, the two representations are similar in terms of clustering information: so, we easily notice that all Quran texts are closely grouped together and all Hadith ones are closely grouped together too. This fact confirms, once again, that the two author styles of the 2 books are different.

Figure 13: GMM clustering after PCA reduction of the Quran and Hadith segments, with the corresponding probability density functions in 3D.

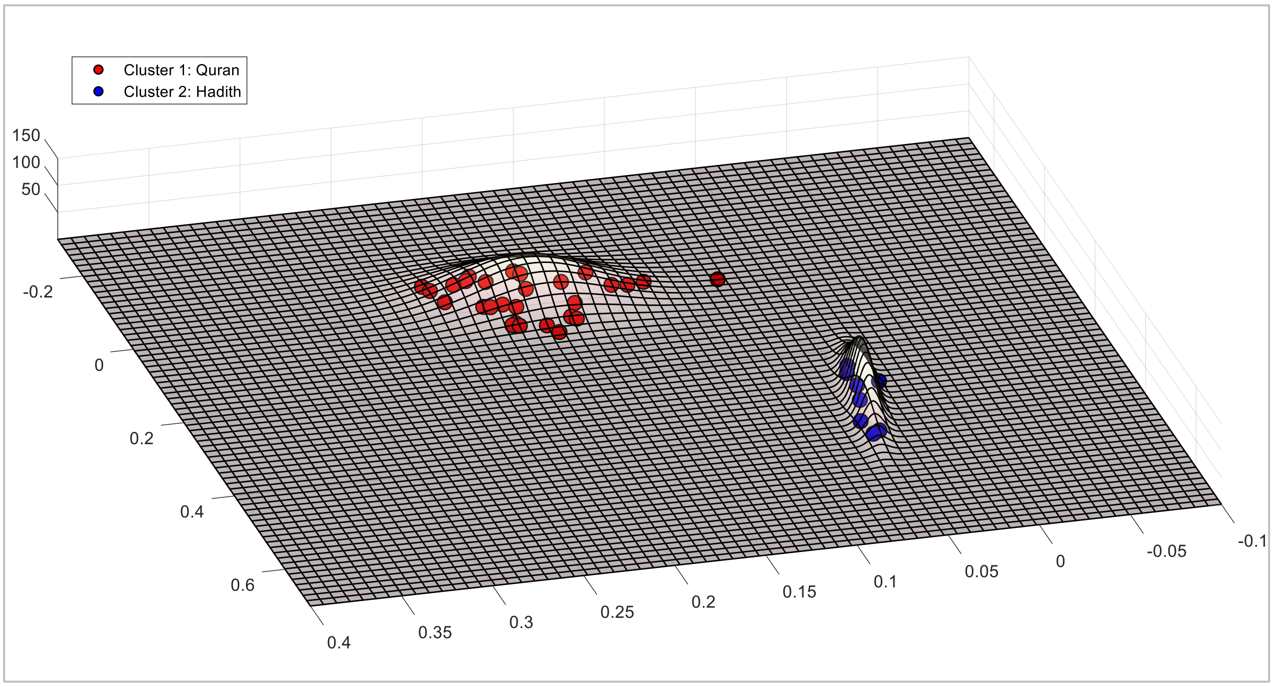
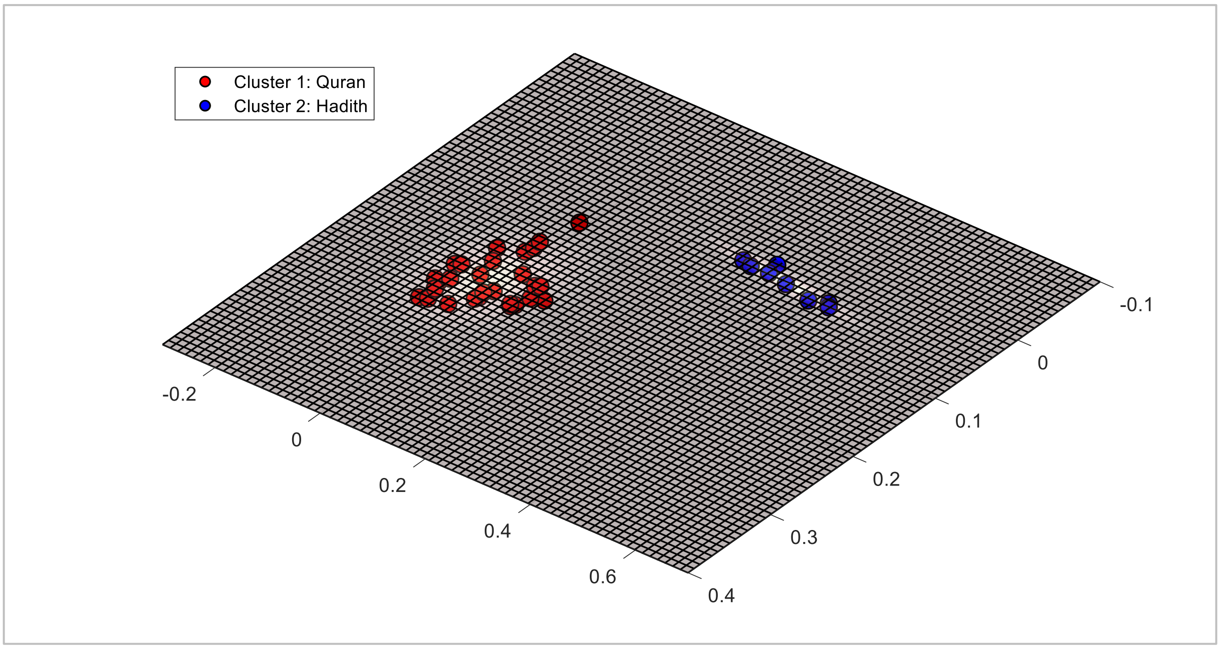


Figure 14: GMM clustering after PCA reduction of the Quran and Hadith segments in 2D.



Sammon mapping based clustering

Sammon mapping is an algorithm that maps a high-dimensional space to a low-dimensional space by trying to keep the structure of inter-point distances in the high-dimensional space in the low-dimension projection (Kim, 2003). This technique is interesting in exploratory data analysis. The method was proposed by John W. Sammon in 1969, and is considered a non-linear approach since the mapping cannot be represented as a linear combination of the original variables such as in PCA.

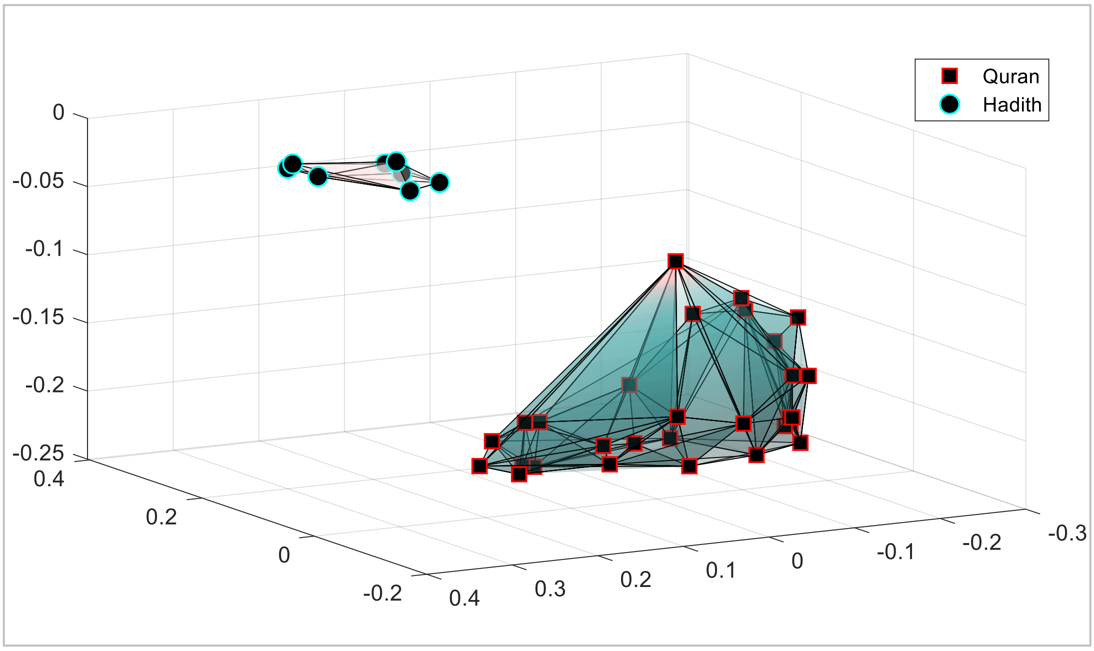
The Sammon projection is an algorithm that maps a high-dimensional space to a space of lower dimensionality, to project the population into 2 or 3 dimensions for observing the relation among populations (Snášel, 2022).

Theoretically, by denoting the distance between the ith and jth elements in the original space by d^*_{ij}, and the distance between their projections by d_{ij}. Sammon's mapping aims to minimize the following error function, which is often called Sammon's error:

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \qquad (15)$$

In 3 dimensions, the Sammon-based graphical representation is quite interesting, since it makes a sharp separation of the different elements by bringing closer all the similar ones. In Figure 15, the 37 text segments are represented in 3D using Sammon mapping.

Figure 15: Sammon mapping clustering after PCA reduction of the Quran and Hadith segments.



The resulting visual representation shows 2 main clusters: one on the bottom right grouping all the Quran segments and another one on the top left grouping all the Hadith segments. Furthermore, those two sets of texts are covered by an interpolated surface between samples of a same type for a visual comfort.

In this figure, it appears that the Quran Author style cluster is well separated from the Hadith one, confirming the clear difference in author styles between the two books.

Conclusion and Discussion

We have proposed a new pertinent set of features based on the normalized Word Bigram Transition Probability, which are well adapted to represent the author style. The described approach is proposed for the first time, at least to the knowledge of the author, and can be used in any task of authorship attribution provided that the document size of the training data is sufficiently large, in order to accurately compute the transition probabilities, even though the experimental results have shown that it can also be applied on short texts.

In this research work, three experiments were performed: the first experiment represents an evaluation on the simulated text corpus SIMSTYL, the second experiment concerns the authorship attribution on the HAT corpus with 100 authors, in Arabic language, the third experiment is an evaluation on a cross-topic subset of the Guardian corpus in English language.

Furthermore, an application of authorship discrimination between two ancient religious books (i.e., Quran, and Hadith) has been conducted by using the new proposed approach.

The experimental comparison of this new set of features with some state-of-the-art features, in several datasets, has shown that the WBTP presents high performances in author profiling and in authorship discrimination. Furthermore, a fusion between the WBTP and character trigrams was proposed too, which interestingly showed a higher accuracy (i.e., the best accuracy ever obtained on the HAT corpus).

As for the cross-topic corpus, the results have shown that this new set of features is less sensitive to the topic and can then be used with documents belonging to different topics. This particular point is quite interesting, since most of existing features of AA require to have documents related to the same topic, which represents a difficult condition to fulfil in practice.

The results of this investigation have shown that the WBTP is interesting for three reasons:

- It presents high performances in authorship attribution;
- It has been used successfully in cross-topic authorship attribution;
- It has been used efficiently with short text documents (i.e., about 200 words).

Moreover, it could be employed, in association, with other conventional features in a form of fusion, for applications requiring high degree of accuracy, as noticed in the fifth section.

Regarding the application of author discrimination between the two religious books (i.e., Quran and Hadith), several experiments have been conducted, namely: Author discrimination using nearest neighbour distance, author discrimination using centroid based distance, hierarchical clustering, PCA based clustering, FCM clustering, GMM based clustering and Sammon mapping.

The Author discrimination experiments have shown that the score of discrimination is 100%, which confirms that the two Author styles are completely different. Similarly, the results of clustering experiments have shown that there are two distinct clusters representing the two author styles, which consequently shows that the holy Quran (*believed to be a Divine revelation*) could not have been invented by the Prophet.

Although several experiments of Author discrimination were already performed on those two ancient books by using different features, different classifiers and different experimental protocols, one wanted to see if this new set of features could lead to the same conclusion or not.

Coming back to the new approach of authorship attribution, which is based on Word Bigram Transition Probability, one can say that this new set of features is pertinent, robust (against the topic change) and appears to be a promising feature in AA.

Acknowledgement

The author of this paper would like to thank, first and foremost, our Creator for granting us the opportunity to complete this research. Gratitude is also extended to all those who contributed with their advice and comments throughout this work.

References

Alduais, A., Al-Khulaidi, M. A., Allegretta, S., & Abdulkhalek, M. M. (2023). Forensic linguistics: A scientometric review. *Cogent Arts & Humanities*, 10(1), 2214387.

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. (2020). Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2), 89-103.

Jovanović, A., & Perić, Z. (2021). Two-dimensional GMM-based clustering in the presence of quantization noise. *Facta Universitatis. Series: Automatic Control and Robotics*, 20(2), 099-110.

Kang, M., Ahn, J., & Lee, K. (2018). Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications*, 94, 218-227.

Khmelev D. V. and F. J. Tweedie, "Using markov chains for identification of writers, *Literary and Linguistic Computing*, vol. 16, no. 3, pp. 299-307, 2001.

Kim Y., B. Moon. New Usage of Sammon's Mapping for Genetic Visualization. Conference: Genetic and Evolutionary Computation - GECCO 2003, *Genetic and Evolutionary Computation Conference*, Chicago, IL, USA, July 12-16, 2003, pp 1136-1147

Kestemont, M. (2014). Function words in authorship attribution. From black magic to theory?. In Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), pp. 59-66.

Li, J., Chang, H. C., & Stamp, M. (2022). Free-text keystroke dynamics for user authentication. *In Cybersecurity for Artificial Intelligence* (pp. 357-380). Springer, Cham.

Lupea, M., Briciu, A., & Bostenaru, E. (2021). Emotion-based hierarchical clustering of romanian poetry. *Studies in Informatics and Control*, 30(1), 109-118.

McLachlan, G. J., Peel, D., & Bean, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3-4), 379-388.

Memorial website of D. Khmelev, Website consulted in August 2020. http://people.oregonstate.edu/~kelberta/dima/

Ouamour S., 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 – 2013, Beijing, China, October 10-12, 2013. http://www.cyberc.org/cyberc2013/.

Patton, J., & Can, F. (2021). A Detailed Stylometric Investigation of the İnce Memed Tetralogy. Technical Report #MiamiU-CSA-04-001, Miami University.

Sayoud H. Biometrics: An Overview on New Technologies and Ethic Problems. *International Journal of Technoethics*, IGI Global, Vol. 2, No. 1, January 2011.

Sayoud H. Author discrimination between the Holy Quran and Prophet's statements. Literary and Linguistic Computing 2012, *Literary and Linguistic Computing*, Vol. 27, No. 4, 2012, pp 427-444.

Sayoud H., 2015. Segmental analysis based authorship discrimination between the holy Quran and prophet's statements. *Digital Studies journal*, Canada, Volume 6, Issue 1, 2015, Congress 2015. https://www.digitalstudies.org/article/id/7268/.

Sayoud H. Statistical Analysis of the Birmingham Quran Folios and Comparison with the Sanaa Manuscripts. *HDSKD journal*, Vol. 4, No. 1, pp. 101-126, December 2018. ISSN 2437-069X.

Sayoud H., H. Hadjadj. (2021). Authorship Identification of Seven Arabic Religious Books -A Fusion Approach, *HDSKD journal*, Vol. 6, No. 1, pp. 137-157, December 2021. ISSN 2437-069X. DOI 10.5281/zenodo.6353805.

Sayoud H., S. Ouamour (2021b). HAT - A new Corpus for Experimental Stylometric Evaluation in Arabic. 12th International Conference of Experimental Linguistics, 11 - 13 October 2021, Athens, Greece. Proceeding of EXLING'2021: https://exlingsociety.com/past-proceedings/#2021.

Sayoud, H. (2022). Stylometric Comparison between the Quran and Hadith based on Successive Function Words: Could the Quran be written by the Prophet? *International Journal on Islamic Applications in Computer Science and Technology*, Vol. 10, Issue 2, June 2022, 01- 06.

Sayoud H. (2022b). SIMSTYL – The Simulated Text Corpus. 2022. Dataset and description available in http://scholarpage.org/SimStyl.pdf.

Shlens J. (2003). A tutorial on principal component analysis: Derivation, discussion and singular value decomposition. 2003. https://www.cs.cmu.edu/~elaw/papers/pca.pdf.

Sidorov, G. (2019). Syntactic n-grams in computational linguistics (pp. 125-125). *Cham, Switzerland*: Springer International Publishing.

Snášel, V., Kong, L., & Pan, J. S. (2022). Visualization of Population Convergence Results by Sammon Mapping in Multi-objective Optimization. *In Advances in Intelligent Systems and Computing*, pp. 295-304. Springer, Singapore.

Sottile, G., Francipane, A., Adelfio, G., & Noto, L. V. (2021). A PCA-based clustering algorithm for the identification of stratiform and convective precipitation at the event scale: An application to the sub-hourly precipitation of Sicily, Italy. *Stochastic Environmental Research and Risk Assessment*, 1-15.

Stamatatos E. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2):421–439, 2013.

Suganya, R., & Shanthi, R. (2012). Fuzzy c-means algorithm-a review. *International Journal of Scientific and Research Publications*, 2(11), 1.

Uddagiri, C., & Shanmuga Sundari, M. (2023). Authorship Identification Through Stylometry Analysis Using Text Processing and Machine Learning Algorithms. In Proceedings of Fourth International Conference on Computer and Communication Technologies: IC3T 2022 (pp. 573-581). Singapore: Springer Nature Singapore.

Yeshilbashian, Y. M., Asatryan, A. A., & Ghukasyan, T. G. (2022). Plagiarism Detection in Armenian Texts Using Intrinsic Stylometric Analysis. *Programming and Computer Software*, 48(7), 435-444.

Yule G. U., The Statistical Study of Literary Vocabulary. *Cambridge: Cambridge University Press*, 1944.