


Please cite the Published Version

Mohamed, Emad and Sarwar, Raheem  (2022) Linguistic features evaluation for hadith authenticity through automatic machine learning. *Digital Scholarship in the Humanities*, 37 (3). pp. 830-843. ISSN 2055-7671

DOI: <https://doi.org/10.1093/llc/fqab092>

Publisher: Oxford University Press (OUP)

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/635194/>

Usage rights:  In Copyright

Additional Information: This is a pre-copyedited, author-produced version of an article accepted for publication in *Digital Scholarship in the Humanities* following peer review. The version of record Emad Mohamed, Raheem Sarwar, Linguistic features evaluation for hadith authenticity through automatic machine learning, *Digital Scholarship in the Humanities*, Volume 37, Issue 3, September 2022, Pages 830–843 is available online at: <https://doi.org/10.1093/llc/fqab092>

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Linguistic Features Evaluation For Hadith Authenticity Through Automatic Machine Learning

Emad Mohamed* and Raheem Sarwar

Research Group in Computational Linguistics, University of Wolverhampton,
United Kingdom

Abstract

There has not been any research that provides an evaluation of the linguistic features extracted from the matn (text) of a Hadith. Moreover, none of the fairly large corpora are publicly available as a benchmark corpus for Hadith authenticity, and there is a need to build a “gold standard” corpus for good practices in Hadith authentication. We write a scraper in Python programming language and collect a corpus of 3651 authentic prophetic traditions and 3593 fake ones. We process the corpora with morphological segmentation and perform extensive experimental studies using a variety of machine learning algorithms, mainly through Automatic Machine Learning, to distinguish between these two categories. With a feature set including words, morphological segments, characters, top N words, top N segments, function words and several vocabulary richness features, we analyse the results in terms of both prediction and interpretability to explain which features are more characteristic of each class. Many experiments have produced good results and the highest accuracy (i.e., 78.28%) is achieved using word n-grams as features using the Multinomial Naive Bayes classifier. Our extensive experimental studies conclude that, at least for Digital Humanities, feature engineering may still be desirable due to the high interpretability of the features. The corpus and software (scripts) will be made publicly available to other researchers in an effort to promote progress and replicability.

1 Introduction

Hadith is of paramount importance in Islamic studies as well as in Muslim daily life. “A Hadith is a report of what the Prophet Muhammad (ca. 570–632) said, did, or tacitly approved. Each Hadith consists of a chain of transmitters, called the *isnad*, and the actual text that was transmitted, called the *matn* (Lucas 2008).” In traditional Hadith studies, these reports are defined as *sahih*, *hasan*, or *da’if*. *Sahih*, or authentic, is a report that has been continually transmitted by narrators who are not known for any mendacity. A *hasan* Hadith may have some weakness, but its *matn* (text) is plausible, and it deserves to be put in to practice. A *da’if* Hadith, or a weak one, is one that does not meet these conditions (Juynboll 2007) (p. 27). A further Hadith category is the *mawdo’*, which refers to fake Hadith invented by some people and attributed to the Prophet for religious, political or other purposes (Al-Albani 1992).

*Corresponding Author

Scholars of Hadith have developed a complex and fine-tuned system for Hadith classification, and there are differences within these classifications. We adopt a simpler approach by considering only those that are listed as Sahih (authentic) and Mawdo' (fabricated), taking only the two extremes of the Hadith authenticity spectrum and ignoring those in the middle.

Ever since the beginning of the Muslim civilisation, the narration and reporting of Hadith has been a thorny issue, as there are many Hadiths in circulation that are known to be fake or inauthentic. Scholars of Islam have authored many treatise listing the authentic and inauthentic Hadith, although there are usually differences across these commentaries depending on the chain of narrators and the text of the Hadith itself. To give an example of a *very weak* Hadith:

يأتي على الناس زمان هم فيه ذئاب، فمن لم يكن ذئبا أكلته الذئاب

There will come a time when people are wolves, and those who are not wolves will be eaten by wolves¹.

In his analysis of this Hadith, Al-Albani (1992) rejects it as either very weak or fake based on the fact that it was solely reported by Ziyad ibn Abi Ziyad Al-Jassas, who was a known liar. Declaring a Hadith fake or weak based on the credibility of the narrators is the most common method. Brown (2008) convincingly claims that *matn criticism* was also part of the arsenal of hadith scholars, but they had to show indifference to the content in opposition to their rationalist opponents who did not care about the *isnad*. Al-Albani's collection of weak and fake Hadiths contains 7162 prophetic traditions, only 753 times does he discuss the *matn* (textual content) of a Hadith, not all of which are related to his judgment on the authenticity of the Hadith.

Most of the existing studies on Hadith authenticity rely on the analysis of the chain of reporters. In addition, there has not been any research that offers evaluation of the linguistic features extracted from the *matn* (text) of Hadiths, i.e., there is no way of knowing the effectiveness of features in defining the boundary between authentic and inauthentic Hadith. Moreover, none of the fairly large corpora is *publicly available*² as a benchmark corpus for Hadith authenticity and there is a need to build a "gold standard" corpus for good practices in Hadith authentication.

While the computational investigation and analysis of the chain of reporters is a worthwhile task, we focus on the textual content of the Hadith rather than its chain of reporters. In this paper, we use machine learning to explore (1) to what extent machine learning can be used to predict the authenticity of Hadith based on its *matn*, and (2) whether there are linguistic differences between authentic and inauthentic Hadiths. For the purpose of this research, we intend to focus on the following research questions:

1. Given a corpus of authentic and fake Hadiths, can machine learning be used to distinguish between the two types based on *matn*?
2. What, if any, are the linguistic features that set authentic Hadith apart from fake ones?

Our focus is thus on the two categories identified by Hadith scholars as certain: either certainly the prophet's tradition, or certainly not the prophet's tradition. We will not consider the ambivalent cases where the majority of scholars differ since these will not help explain the possible differences between inauthentic and authentic Hadiths. The outcome of the machine learning experiments in this research are models that can assign a specific Hadith to either the authentic or the fake category. If such models are accurate enough, we can then use them to determine whether a specific weak/suspicious Hadith is either fake or authentic. The current study can be seen as the next logical step, and is different from all previous studies in the following respects:

¹Unless otherwise indicated, all the translations provided in this article are by the authors.

²Source and Hadith number are provided or the corpus used can be downloaded

- The focus of the current study is the distinction between authentic and fake Hadith using linguistic features extracted from *matn* of the Hadiths and the evaluation of these features using automatic machine learning (AutoML).
- We provide an analysis of the most discriminative linguistic features, which helps explain why a certain Hadith is authentic.
- We have collected a large corpus of authentic and fake Hadith, and a third of disputed Hadith to conduct extensive experimental studies on Hadith authenticity. We provide clear evaluation. The corpus and software (scripts) will be made publicly available to other researchers in an effort to promote progress and replicability.

The rest of the paper is organised as follows. Section 2 presents the literature review. Section 3 illustrates the data and methods. Section 4 presents our results and discussion. Section 5 contains the concluding remarks.

2 Literature Review

Several studies focus on authorship attribution of the religious texts such as Hadith and Quran (Sayoud 2012; Sayoud 2015; Hadjadj and Sayoud 2016; Sayoud and Hadjadj 2017; Sayoud 2018). However, in this paper we tackle the question of distinguishing between fake and authentic prophetic traditions. Existing notable Hadith authenticity methods can be organised into machine learning and rule-based methods (Binbeshr, Kamsin, and Mohammed 2021) as described in the following subsections. A summary of the literature review is provided in Table 1.

2.1 Machine Learning-based Methods

In this part of the paper, we review existing notable machine learning-based studies on Hadith authenticity.

Abdelaal and Youness (2019) classify Hadith into four categories: Sahih, Hasan, Da'if and Mawdoo', according to the reliability and memory of the narrators. They applied two classifiers, Decision Tree (DT) and Naïve Bayes (NB) and report that NB outperforms the DT classifier. However the source and the size of the corpus used is not reported by the authors and the corpus is not publicly available. Similarly, Ghanem, Mouloudi, and Mouchid (2016) report that the order of narrators is critical for Hadith authentication. They represent each narrator as a term and use Support Vector Machine (SVM) with Learning Vector Quantization (LVQ) to consider the order of narrators. This method was validated on a small corpus containing 160 Hadiths. However, the authors have not reported the source of the corpus and it is not publicly available. Aldhlan et al. (2012) also propose to classify Hadith according to the validity of its Isnad. They use the DT classification model with a mechanism of handling missing Isnad attributes to classify Hadiths into four categories: Sahih, Hasan, Weak, and Mawdoo. The method has been validated on a small corpus of 999 Hadiths. The corpus they used is not publicly available as a benchmark corpus for Hadith authentication.

Shatnawi, Abuein, and Darwish (2011) extract Hadith text from web pages and ascertain the degrees of correctness by looking it up in the Sheikh Al- Albani Hadith collection (correct series and weak series) and their degrees of correctness according to Sheikh Al-Albani studies. They built a positional index of the Sheikh Al-Albani Hadith collection (SAHDB) and extract Hadith terms as queries from the passed web page, then execute these queries against SAHDB index. The corpus retrieved from webpages as part of these experiments is not publicly available as a

benchmark corpus for Hadith authentication. Moreover, they don't offer any feature evaluations and the corpus used.

Hassaine, Safi, and Jaoua (2016) create a binary relation for each category of Hadith (authentic and non-authentic) where the Hadiths correspond to the objection of the relation, and the words correspond to its attributes. They then obtained keywords for each category, using hyper rectangular decomposition to order in terms of importance and feeding extracted keywords through a logistic regression model to perform classification. They evaluated their method on 1,600 Hadiths. The accuracy of their method decreases as the number of opinions per Hadith decreases and they don't offer any feature evaluations. Moreover, the corpus used in their study is not publicly available.

Elewa (2018) approached the problem of some questionable Hadiths in the books of Baukhari and Muslim. He used a model analysing word length, lexical richness (Type/Token Ratio), and lexical density to determine which Hadiths are authentic. Elewa's study falls under corpus linguistics. Although an important study, Elewa (2018) is limited in terms of the feature set it employs and does not offer any evaluation. There is no way of knowing how effective the three measures he uses are in demarcating the boundary between authentic and inauthentic Hadith. Also, his study's main focus is a deep and close analysis of a limited number (40 Hadiths) of publicly available Hadiths in each of the two books examined.

2.2 Rule-based Methods

Azmi and AIOfaidly (2014) use a simple rule-based method to classify Hadith veracity into Sahih, Hassan, and Weak. They assign a weight for each narrator in the Hadith chain. This weight depends on narrator's trustworthiness, generation, and deficiencies. Hence, Hadith authenticity is determined by calculating the normalized sum of all narrators' weights (NSNW) in the Isnad. This method was evaluated on 2,932 Hadiths. The corpus used is not publicly available as a benchmark corpus for Hadith authentication.

Bilal and Mohsen (2012) present a cloud-based system that relies on narrators' information to extract facts from the user query and classifies Hadiths. The corpus used is not publicly available as a benchmark corpus for Hadith authentication. Similarly, Ghazizadeh et al. (2008) present a fuzzy system that determines the validity rate of a Hadith based on Hadith narrators and continuity. It authenticates Hadiths into five categories: Unknown, Mawdoo', Da'if, Hasan, and sahih. The corpus used is not publicly available as a benchmark corpus for Hadith authentication.

Two rule-based studies have been published which present web extensions. These that read the content of the specific Arabic website and screen the text to verify whether it contains any Hadith text by comparing with the authentic source. If a comparison produces a complete match, then the text is marked with a green to highlight the text as authentic. If some words/letters are missing, then the text is marked red to indicate that the Hadith is unauthentic. The corpus used is not publicly available as a benchmark corpus for Hadith authentication. (Kabir, Tayan, et al. 2019; Kabir, Hasan, et al. 2018).

Type	Paper	Method	Retrieved from	# Hadiths
ML-Based	1. (Abdelaal and Youness 2019)	Naive Bayes	N/A	N/A
	2. (Hassaine, Safi, and Jaoua 2016)	Logistic Regression	Shamela.ws, hdith.com	1600
	3. (Ghanem, Mouloudi, and Mourchid 2016)	Support Vector Machines	N/A	160
	4. (Aldhlan et al. 2012)	Decision Trees	Sahih Al-Bukhari, Jami'u Al-Termithi, and Silsilat Al-Hadiths Al-Dae'ifah w'Al-Mawdhu'ah	999
	5. (Shatnawi, Abuein, and Darwish 2011)	Positional Index and Web Data Extraction	Al-Selseleh AISahihah and Al-Selseleh ALDa'eefah (Al-albani)	
Rule-Based	6. (Kabir, Tayan, et al. 2019)	Similarity Search	Sahih Al-Bukhari and Sahih Muslim Book	N/A
	7. (Kabir, Hasan, et al. 2018)	Similarity Search	Sahih Al-Bukhari and Sahih Muslim Book	N/A
	8. (Azmi and AIOfaidly 2014)	Normalized Sum of Narrators' Weight	Sahih Al-Bukhari and Sunan Tirmizi Books	2,932
	9. (Bilal and Mohsen 2012)	SaaS System	N/A	N/A
	10. (Ghazizadeh et al. 2008)	Fuzzy System	Usul Al-Kafi (Volume 1) Book	N/A

Table 1: Summary of literature review

3 Data & Methods

In order to answer the two research questions set out in the introduction, we adopt the following method, using the data outlined below.

3.1 Data

There are several books dedicated to answering the question of whether a certain Hadith is authentic or inauthentic. For the purpose of this study, we derive our data from the following sources:

1. For authentic Hadith, we use *The Series of Authentic Hadith and some Benefits Associated with them*, which was compiled from other sources, by Nasir Al-Din Al-Albani (1914-1999), a major figure in Hadith studies. We have extracted 3651 Hadith from this book, which is the vast majority of the content. The Hadiths in the book are numbered from 1 to 4065, but manual examination revealed that there are gaps in the numbering. For example, Hadith 4033 immediately follows Hadith 4006. We have not examined every Hadith manually, so we are not certain how many have been missed. Our method of extraction favoured precision over recall as we extracted only those in quotes that are associated with a number according to the style of the book. The total number of traditions in the corpus is 7244, with the corpus being unintentionally almost balanced.
2. Although Al-Albani also authored a book on non-authentic Hadith, his book is a compilation of both proven inauthentic and disputed Hadith. For this reason, we have turned to other sources of inauthentic Hadith, namely: (1) *The Book of Fake Hadith* by Al-Saghani, from which we have extracted 2279 Hadiths, (2) *The Book of Fake Hadiths* by Ibn Al-Jawzi, from which we extracted 1279 Hadiths and (3) *The Book of Fake Traditions* by Ali Qurra Daghi, which provided 661 Hadiths. The total of the Hadiths extracted from the three books is 3593, after removing duplicates and very similar traditions. During this process is a Hadith was duplicated as part of longer Hadith we maintained the longer version.

3.2 Methodology

We perform Hadith authenticity in four steps: (i) data cleaning and preparation; (ii) preprocessing; (iii) features extraction; and (iv) machine learning. These steps are explained in the following subsections.

3.2.1 Data Cleaning and Preparation

The *authentic Hadith sub-corpus* was straightforward to obtain. We extracted the Hadiths from an EPUB book published by Shamela, a website dedicated to making Islamic books available for wider readership. Since EPUB is in a compressed XHTML format, the process of obtaining the text of the Hadiths involved minor processing using regular expressions and html processing tools.

The *Fake Hadith sub-corpus* is a different matter. Although it was still downloaded from Shamela, the corpus came from three books with different formattings. In addition to the regular extraction and cleaning, the biggest challenge was how to detect and remove repeated items. Repeated items come in two forms:

- When a Hadith is narrated in two or more different sources with only minor lexical differences, in which case we keep only one of them. For example, the following Hadith

الأَعْمَالُ بِالنِّيَّةِ، وَلِكُلِّ أَمْرٍ مَا نَوَى، فَمَنْ كَانَتْ هِجْرَتُهُ إِلَى اللَّهِ وَرَسُولِهِ فَهَاجَرَ إِلَى اللَّهِ وَرَسُولِهِ، وَمَنْ كَانَتْ هِجْرَتُهُ لِدُنْيَا يُصِيبُهَا أَوْ امْرَأَةٍ يَتَزَوَّجُهَا، فَهَاجَرَ إِلَى مَا هَاجَرَ إِلَيْهِ.

Deeds are judged by the intention. He who emigrates to Allah and His messenger, then his emigration is for Allah and His messenger, and he emigrates for worldly benefits or for a woman to marry, then his emigration is for that to which he emigrated.

has many versions, of which we only keep one version. These eliminations happens automatically with minor manual corrections.

- When the same words are used, but the ordering is different, we only keep one version.
- When a Hadith is a part of another Hadith, we retain the longest version.

Our main rationale for discarding repeated items is that we do not want one of the versions to fall in the test set while a near-match, a super-text, or a sub-text is part of the training sets in the experiments below. This would cast doubt on the validity of the results ³.

As part of this process, we remove all punctuation from the Hadith corpus. While punctuation is essential for understanding text and authorship attribution, Hadith pre-dates this practice, and many Hadiths could be punctuated in numerous ways. In this case, punctuation reflects the understanding of the Hadith scholar (or the typist) and not necessarily the intentions of the original utterer. We remove punctuation in both the fake and authentic sub-corpora.

3.2.2 Preprocessing

The main form of processing we perform in this paper is morphological segmentation. Arabic is a morphologically rich language in which a white space-delimited unit, the orthographic word, may not map exactly to a linguistic word. Many prefixes and suffixes constitute part of the written word, and these affixes can be mapped to full words in many other languages, as they perform syntactic roles, such as conjunctions and prepositions. For this reason, morphological segmentation is usually carried out to separate these and, as a consequence, mitigate the problem of scarcity in linguistic resources. For example, the Arabic word *fsnstxmdhA* فسئستخدما is made up of the conjunction *f* (then), the future particle *s* (will), the verb *nstxmd* (we use), and the third person singular feminine object pronoun *hA* (it, her, them). The verb *nstxmd* itself is made up of two units: the first person plural imperfect prefix *n* (we) and *stxmd*, the imperfect verb stem (use).

Take for example the Arabic word *fsnstxmdhA*, as shown in Figure 1.

This morphological complexity leads to a high type token ratio, where there are many unique words, many of which are not actually unique but rather are morphological variations. Morphological segmentation helps reduce the number of hapax legomena, and renders the text more amenable to computational analysis. For morphological segmentation, we use the Arabic-SOS package, which is specialised in Classical Arabic and is known to produce the best results in the religious genre (Mohamed and Sayyed 2019). In order to establish whether the tool is good enough for our purposes, we manually examined a randomly selected set of 100 Hadiths (50 fake, 50 authentic). The set contained 2014 words, of which 16 were incorrectly segmented, putting the segmenter’s accuracy at 99.21%. This is a very high accuracy.

³While every effort has been made to maintain this separation, one cannot rule out the possibility of minor insignificant leaks. For this reason, we will make the data available if and when this article is accepted for publication. This should make it easy for other researchers to examine and improve on the current results.

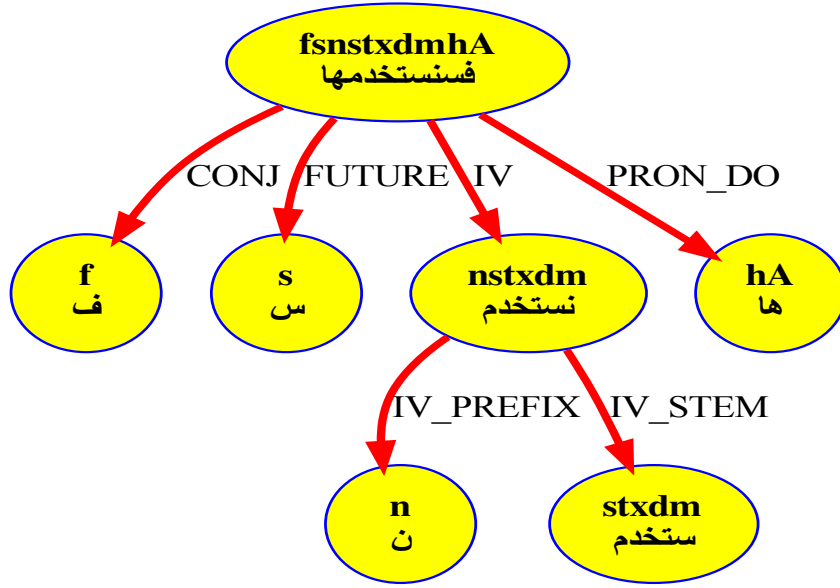


Figure 1: The morphological analysis of an Arabic word

3.2.3 Features Extraction

Hadith authenticity can be performed in two steps. In the first step we extract features from Hadiths. In the second step, we apply a machine learning algorithm to predict the authenticity of the Hadith. In this paper we extract the following features from each Hadith. These features are selected based on previous studies that compared authorship attribution methods and found in this regard these are successful features, and NLP tools available for Arabic (Grieve 2007; Eder 2015; Savoy 2015; Reborra et al. 2019). Segmentation features are considered here due to the morphological complexity of the Arabic script.

- *Word ngrams.* An ngram is a unit of dividing text into words. For example, in the sentence: *The Prophet did not say many of the things attributed to him*, unigrams are: ['the', 'prophet', 'did', 'not', 'say', 'many', 'of', 'the', 'things', 'attributed', 'to', 'him'], while in bigrams every two consecutive words make one unit: [('the', 'prophet'), ('prophet', 'did'), ('did', 'not'), ('not', 'say'), ('say', 'many'), ('many', 'of'), ('of', 'the'), ('the', 'things'), ('things', 'attributed'), ('attributed', 'to'), ('to', 'him')]. Trigrams are every three words together. In this experiment, a Hadith can be represented as a combination of various ngram ranges. We use combinations of word ngrams to find out which combination is the best discriminator (Nutanong et al. 2016).
- *Segment ngrams.* This is very similar to word ngrams except that segment will be used instead of word.
- *Character ngrams.* Instead of treating a Hadith as a combination of words, we can treat it as a combination of characters (Sarwar, Li, et al. 2018). While this may seem counter-intuitive, it is useful in capturing similarities and unique traits. It may be the case that what distinguishes speaker A from Speaker B is that Speaker A uses a specific letter, or group of

letters, more often. This also helps with morphological complexity, and has proved to be a very powerful technique in text classification in general.

- *Top N words.* It can be seen from the literature on authorship attribution that it is the most frequent words, commonly function words, that are capable of distinguishing between different authors (Shaker and Corne 2010). The reason for this is that when using all of the words in the corpus, the theme of the document may have an impact and this may not necessarily relate to the author’s style. Using the top N words is thus more expressive in terms of attribution even though it may not be as accurate for the purpose of text classification. In the literature, N stands for a limited number, and in most cases, we use the top 100 to the top 500 words.
- *Top N segments.* This is very similar to the top N words except that we use segments instead of words. For example, the word *fsykfykkhm* above will be treated as six different units: f, s, y, kfy, k, and hm. This is better able to capture function words as many function words in Arabic are actually bound morphemes. This way, function words are better represented, which makes detecting author style more straightforward.
- *Function Words.* Function words are those words that are not lexical, the lexical being nouns, verbs, adjectives and their derivatives (Sarwar and Nutanong 2016). These include pronouns, prepositions, demonstratives, conjunctions and similar words. We have collected 87 such words. It has to be noted that using these as features requires segmentation as many function words are bound morphemes.
- *Vocabulary Richness.* Vocabulary richness features can be used for authorship prediction as well as feature contribution explanation (Sarwar, Porthaveepong, et al. 2020). In the context of the current paper, we make use of the following vocabulary richness features:
 - Type Token Ratio. The *TTR* is the number of unique words in a document divided by the total number of words. This is also known as *Lexical Richness*, and is an indication of how vast the author’s vocabulary is. A larger *TTR* indicates a richer vocabulary repertoire. One issue with this measure is that it is sensitive to the length of the document. For this reason, we divide the resulting number by the length of the document as a way of normalisation (Sarwar, Yu, et al. 2018).

$$\frac{\text{unique words}}{\text{total number of words}}$$

- Average Word Length in Letters. This is how many letters a word has on average, and is computed by dividing the length of the document, in letters, by the number of words in a document. This could be a unique authorial feature as some authors use longer words than others. Another way to approach this is to compute the distribution of words lengths: the ratio of 1-letter words, the ratio of 2-letter words, etc (Sarwar, Li, et al. 2018).
- Average Word Length in Segments. Instead of counting how many letters a word has, we count how many segments it has. For example, the word *hjrth* is made up of three segments: *hjr+t+h*. This is a marker for morphological complexity. Words with more segments are more morphologically complex. This measure is computed by dividing the whole number of segments in a document by the number of words in that document.

- Document Length. This is computed by analysing how many words there are in each Hadith (Sarwar, Urailetrprasert, et al. 2020).
- Lexical Density. This is computed by dividing the number of function words in the document by the total number of words. The idea behind *LD* is that most of the meaning in a sentence is carried by lexical items (nouns, adjectives, verbs and adverbs). Function words (i.e. non-lexical items) are connectors of the meanings, and these connectors can help differentiate between authors. To obtain the list of function words in Arabic, we listed all pronouns, demonstratives, prepositions, and particles. We then used segmentation to count the frequencies of these function words in the document, since many of these are bound morphemes. For example, the preposition *b* in Arabic hardly ever stands alone and is always attached to the words as a clitic. *bAlktAb* (Eng. with/by the book) thus has a preposition and a noun (Sarwar and Nutanong 2016).

3.2.4 Machine Learning Algorithms

For researchers to design a machine learning solution, they have to go through many steps involving feature engineering (extracting features from the text), preprocessing, and algorithm selection. This is an arduous process that takes both time and effort, and is not guaranteed to yield the best model. A researcher may decide to use Support Vector Machines while Extreme Gradient Boosting could actually be the best approach to the problem.

We do not opt for a specific algorithm in this paper, instead we use automatic machine learning (AutoML) to learn the best parameters and algorithms. We use the Tree-based Pipeline Optimization Tool (TPOT) for selecting the best pipelines for the experiments in this paper. TPOT uses Genetic Algorithms to select the best classifiers and tune the hyperparameters. TPOT does not only evaluate a single classifier, but examines both individual and stacked classifiers. Stacking is when one applies consecutive classifiers, as well as dimensionality reduction techniques, to the data (Le, Fu, and Moore 2020). All the experiments involved the Scikit-Learn Machine Learning Library (Pedregosa et al. 2011).

In all of the experiments below, we divide the data at random into 80% for training and 20% for testing, and we keep the same division across all of the experiments. The training set is then tuned in a five-fold cross validation manner. The best algorithms and hyper-parameters are then used to predict results from the test set. We then evaluate the performance of the best pipeline (preprocessing, classifier or stack of classifiers and hyper-parameters), on the test set to obtain the numbers listed here in the results table ⁴.

Accuracy is our evaluation metric of choice throughout this paper. Accuracy, computed as the number of correct predictions divided by the number of all instances, is a suitable measure when the corpus is (almost) balanced, which is the case in these experiments.

4 Results & Discussion

In this section we provide answers to the two research questions posed at the beginning of this article. First, we answer the question of whether machine learning can predict fake vs authen-

⁴When we started doing this work, we decided to use Extreme Gradient Boosting (XGBoost) and Deep Neural Networks for classification. We conducted several experiments using manual optimisation. When we started using AutoML, we found out that the models suggested by TPOT were always superior to the manually optimised ones, especially those that involved stacking.

tic prophetic traditions. We then move to the question of what features mark fake vs authentic traditions.

4.1 Predicting Fake vs Authentic Traditions

Table 2 lists the results of the experiments we have carried out in this paper. Using whole words as features reports an accuracy level of 77.1% which is significantly higher than the random chance accuracy (i.e., 50%). This clearly indicates that the word distribution is different across both Hadith categories, i.e. the words and their probabilities are distinctive. Higher order word ngrams, when we use bigrams and trigrams, produces even better results, but the difference between using two words and three words does not seem to result in a significant change in performance. In all three word-based classification models, the best performer was the Multinomial Naive Bayes classifier. The word features were in terms of word counts, as words were used as features with the feature values being the number of occurrences for each word (or higher order ngram) in each document. In all of the word-based text classification experiments presented, the Multinomial Naive Bayes algorithm produced the best results.

Features	Values	Algorithm	Accuracy
Words	1-grams	MNB	77.1
	1+2 grams	MNB	78.2
	1+2+3 grams	MNB	78.28
Segments	1 grams	MNB	74.33
	1+2 grams	MNB	77.16
	1+2+3 grams	MNB	77.43
Char	1 grams	RF	65.38
	1+2 grams	RF	71.86
	1+2+3 grams	LR	74.28
	1+2+3+4 grams	MNB	76.14
Top Words	100	MNB (MLP)	67.86
	200	MLP (BNB)	69.59
	300	ET (MNB (GB))	69.66
Top Segments	100	ET (PCA)	66.8
	200	BNB (RF (RF))	70.46
	300	ET	71.29
Function Words	-	ET	65.7
Vocabulary Richness	-	ET (PCA)	58.49

Table 2: The results of classification. MNB = Multinomial naive Bayes, RF = Random Forests, LR = Logistic Regression, BNB = Bernoulli Naive Bayes, ET = Extra Trees, GB = Gradient Boosting, MLP = Multi-layer Perceptron, PCA = Principal Components Analysis.

Using segmentation produces results that are slightly worse than those produced by whole words, with the combination of 1+2+3 ngrams producing an accuracy of 77.43% using the Multinomial Naive Bayes algorithm. This may indicate that segmentation is not absolutely necessary for producing good results and that it may not be required to perform this extra step.

While word and segment features proved useful for document classification, they may not be the best at capturing author style as they are theme-dependent. One way around this is the use of character ngrams. The results show that using the distributions of single letters alone is much better than chance at telling the difference between authentic and fake Hadith, with an

accuracy of 65.38%, which means that the character distribution across the two categories is significant. Going from 1 character to 1 + 2 improves the accuracy by six absolute points (71.86%). The best accuracy is achieved by a combination of unigrams, bigrams, trigrams, and qudrigrams using the MultiNominal Naive Bayes classifier. This indicates that characters are very useful in discriminating between the two categories, with the added advantage that they are not exactly lexical.

Top N words can also be used for classification and they are mostly used for authorship attribution. When we use the top 100 words, we obtain an accuracy of 67.86%, which means this could be a viable solution. The accuracy gets better as we increase the number of words, with 200 top words yielding an accuracy of 69.59% and 300 words achieving 69.66%. The best results were achieved by stacking Muttilayer perceptrons, Multinomial Naive Bayes and Gradient Boosting.

Top segments fare even better than top words with the best results achieved by the 300 top segments and the Extra Trees classifier achieving 71.29%. In both top N words and top N segments experiments, we see stacking algorithms are the winning solution in five of the six experiments reported. Stacking is used in machine learning to combine several classifiers that are different in terms of their weaknesses and strengths. A third classifier is then used to learn when each of the base classifiers should be used. The number of base classifiers does not have to be twos, and each base classifier could be a pipeline of features and different processing steps (“Stacked generalization” 1992).

Function words perform much better than chance at 65.7%, on par with character unigrams, and they do so using the Extra Trees algorithm. They are only followed by linguistic features as the worst predictors overall as the experiment demonstrated an accuracy of 58.5%.

The poor performance with the most interpretable experiments, function words and linguistic features, shows that the most telling features, from a computational perspective, do not have to match human expectations. While these features may not do very well in the prediction task, they may shed some explanatory light on the phenomena under discussion.

4.2 Explaining the Differences between Fake and Authentic Prophetic Traditions

So far, we have only considered the problem of prediction. In machine learning, there is always the distinction between interpretable machine learning, where you can peek into the process and determine what factors are responsible for the prediction and distinction, and uninterpretable machine learning algorithms that give better results most of the time, but are mostly like black boxes. In this section, we focus on providing some interpretation as we discuss the features (independent variables) that are responsible for the classification. Given that we are not using a single specific algorithm, we will examine feature importance using a model-agnostic algorithm: *Permutation Feature Importance*⁵. PFI works by shuffling the values of a specific feature then measuring the impact of this shuffling on the classifier’s performance. If the performance suffers as a result of this shuffling, then this is an indication that the feature is important. When repeated enough times, with all the features in the corpus, this gives us an estimate of how much each feature contributes to the model. The output of this feature importance calculation is very important for many real-world scenarios, but for our purposes, it seeks to answer the why question: *Why was Hadith X classified as fake while Hadith Y classified as authentic?* In the following, we will examine the most important features in the word-based experiments, the character-based experiments, the segment-based experiments, and the linguistic features experiment. We will only discuss the best

⁵https://scikit-learn.org/stable/modules/permutation_importance.html

performing experiment in each set. We will highlight the *Linguistic Features* experiment since it is the highest performing of all experiments.

4.2.1 Word

The experiment based on document classification, in which words are used as features scores the best accuracy with unigrams achieving 77.1% and 1 to 3 grams achieving 78.28%. This can be attributed to the fact that many of these words do not repeat much, either because they're more specific to one category, or because of their morphology. When we use the *odds ratio* (OR), we find, for example, that the word *qblkm* [English: *before you*] has an OR of 0.063, which means it is 16 times more likely to be found in authentic Hadiths while the word *AlTryq* [English: *the way*] has an OR of 0.0675, which means it is 15 times more likely to occur in authentic Hadiths. In the same vein, the word *Ulama'* [English: *scholars*] is 22.5 times more likely to be found in fake Hadith, similar to the word *rjb*, which is 19.4 times more likely to occur in fake Hadiths.

4.2.2 Segment-based Experiments

Table 3 presents the most distinctive segments for fake/authentic classification. While the table may be self-explanatory, we can notice that the most distinctive segment for authentic Hadith, camels, comes from the Arabian environment, and is repeated as a collective noun and as a singular noun. The word has an OR of 0.036, meaning it is 27.8 times more likely to occur in authentic Hadith than in fake Hadiths. We also notice that the highest OR segments list in the authentic corpus contains two verbs, both in the imperative form, which in turn proves the interactive and oral nature of authentic Hadith. For fake Hadiths verbs are very low on the list. The most distinctive word from fake Hadith is scholars, with an OR of 27.65 (27.65 times more likely to occur in fake Hadiths). The role of scholars in Islam is logically a later development. We also find words related to sufism including wool, ascetism, and love.

The Historical Dictionary of Sufism defines Sufism as "commonly used to describe various aspects of the Islamic mystical tradition and its institutions. Some scholars have argued that it derives from the Arabic term *sūf* (wool), suggesting that the earliest Sufis were ascetic types known for wearing rough woolen garments." The same source also explains (under the entry LOVE) that Sufis commonly use LOVE to describe the relationship between God and the Sufi seeker, punning on the dual use of the Arabic root *hb*, "suggesting that one related root meaning, "to produce seed," gives an insight into the "seminal" significance of love which takes root in the heart" (Renard 2015).

4.2.3 Function Words

The function words as features experiment yielded an accuracy of 65.7%, but it can still be useful in explaining feature contribution: which function words contribute to the decision of whether a certain Hadith is fake or authentic? Based on feature permutation, the top 5 function words are given in Table 4. The function words with more feature permutation importance have greater probability to be included in the authentic Hadiths than the non-authentic ones.

4.2.4 Vocabulary Richness Features

Linguistic features did not prove to be particularly good at predicting which Hadiths are authentic and which are fake, with an accuracy of 58%, which is better than chance, but not as good as the other methods. These results were obtained using a pipeline of the Robust Scaler and the MLP

Segment	English	OR	Segment	English	OR
إبل	Camels	0.036	علماء	scholars	27.65
خطايا	Sins	0.0459	رجب	Rajab	19.41
ذاك	That	0.0633	صوف	wool	15.31
خبث	Rust/malice	0.0675	محب	love/lover	14.29
يومئذ	That day	0.0724	إخلاص	Ikhlas	13.82
دجال	Antichrist	0.0777	عربي	Arab	11.22
اذهب	Go	0.078	زهد	asceticism	11.22
بايع	Declare allegiance	0.078	حلاوة	sweetness/halva	11.22
بعير	camel	0.0842	حمام	pigeons	10.23
قرن	generation/century	0.0842	فاسق	dissolute	10.19

Table 3: Most distinctive segments

Word	English	ProbFake	ProbAuthentic
كم	you (pl)	0.016	0.0245
كان	was	0.011	0.024
و	and	0.153	0.158
ما	what/not	0.021	0.026

Table 4: Top 5 function words

classifier. Linguistic features can still be used to offer some explanation as to which features are more characteristic of which class of Hadith. Figure 2 shows a ranking of the most important features according to the Multi-layer Perceptron Classifier.

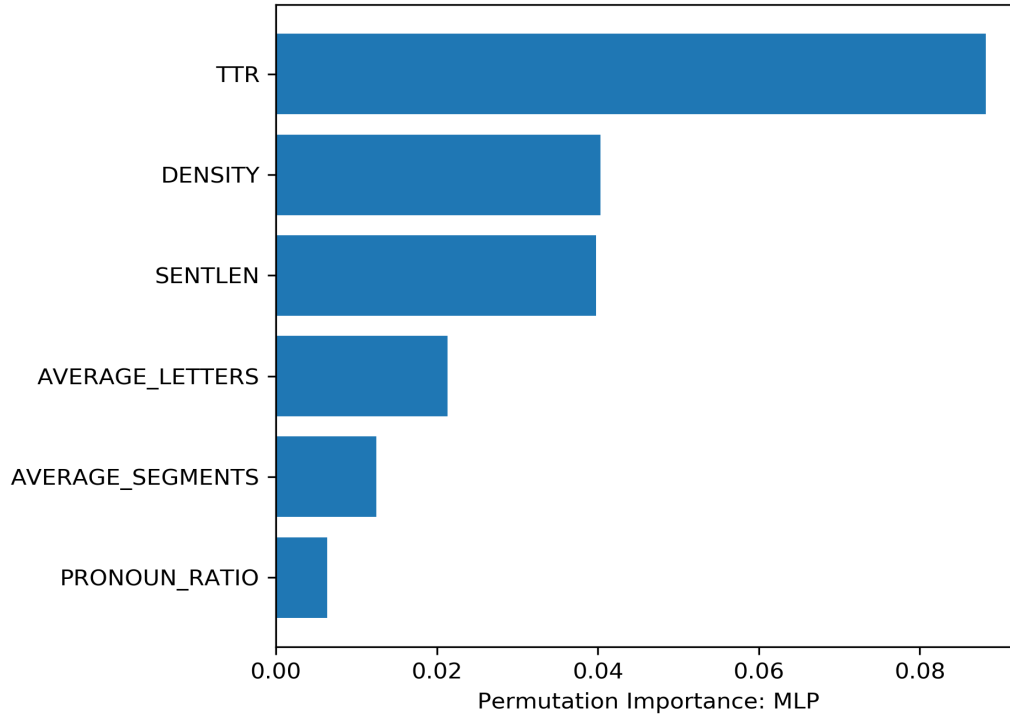


Figure 2: Feature importance for linguistic features using the MLP Classifier

The Type Token Ratio is the most important feature according to this model. We can see that there is a difference between the TTR for authentic and fake Hadiths. While authentic Hadiths have a mean TTR of 0.090616 (SD = 0.077420, median = 0.071429), fake Hadiths have a mean TTR of 0.112078 (SD = 0.092806, median = 0.090000). Higher TTR is characteristic of written language as the higher TTR, the richer the vocabulary, and the less the repetition. "[I]f a speech sample contains 20 words and they are all different we obtain the 'ideal' TTR: 20/20 = 1.00. On the other hand the sample in which the same word is repeated 20 times yields a figure of 1/20 = 0.05"⁶

Sentence Length ranks third in terms of feature importance, and when we compare LD for fake and authentic Hadith, we find that the average LD for fake Hadiths is 15.276525 (SD = 14.604292) compared to 18.963268 (SD = 20.53) for authentic Hadith. Authentic Hadiths are thus longer on average. This is consistent with the formality hypothesis since written language is more packed. Paltridge (2007: 13-19) explains that the written language is structurally more complex but the spoken language is more spread out. While we have not studied the Hadiths structurally, we have seen that fake Hadiths are more spread out than authentic ones, i.e. they have more words.

The average number of letters per word comes next in terms of feature importance. When we examine word lengths in the corpus, we find that the mean length of words in the fake Hadith is 4.13 letters (STD = 0.633666, MEDIAN = 4) compared to 3.258 in the authentic ones (STD = 0.398806, MEDIAN = 3.25). This shows that there is a clear difference between word lengths in the two categories. One possible interpretation, in line with the formality hypothesis, is that Hadith is oral in nature, and orality is more informal, which leads to words being shorter. Fake Hadith, on the other hand, were intentionally written to deceive and extra efforts may have been put into them. In their investigation of formal vs informal texts, Abu Sheikha and Inkpin found the average word length was the most important feature to distinguish informal from formal language with longer words characterising formal texts (Sheikha and Inkpen 2012).

While the average number of segments does not seem to be significant for distinguishing between fake and authentic Hadiths, this is probably because it strongly correlates with the average number of letters in a word. The Person correlation coefficient between the two is 0.79, which is a very high positive correlation. There is still a difference between the average number of segments in both cases with the average for fake Hadiths being 1.72 [STD = 0.285, MEDIAN = 1.6875] compared to 1.51 [STD = 0.185, MEDIAN = 1.5] for authentic Hadiths, which is again in line with the formality hypothesis.

Lexical Density differs from the other features in that while higher lexical density is more aligned with written language, LD for fake Hadith is 0.5 on average compared to 0.6 for authentic Hadith. This is an exception.

The overall differences between fake Hadiths and authentic Hadiths might be then due to the distinction between written and spoken language (or formal and informal discourse). There is more nominalisation in written discourse than in spoken discourse. Nominalisation refers to preferring the use of nouns to the use of verbs. This is related to structural complexity as "[w]ritten texts also typically include longer noun groups than spoken texts. This leads to a situation where the information in the text is more tightly packed into fewer words and less spread out than in spoken texts." This does not necessarily mean that fake Hadith were written. It shows that they display characteristics of written discourse and are thus more likely to have been more pre-planned or pre-designed while authentic Hadiths are more spontaneous and real time. In fact, Biber uses the terms *interactional* versus *edited* texts (Biber 1986), which could be more accurate for Hadith

⁶Brian Richards (1987). Type/Token Ratios: what do they really tell us?. *Journal of Child Language*, 14, pp 201-209 doi:10.1017/S0305000900012885

investigation.

5 Conclusion

We have presented an automatic system that can distinguish between fake and authentic prophetic traditions using automatic machine learning. With a feature set including words, morphological segments, characters, top N words, top N segments, function words and several vocabulary richness features, we analyse the results in terms of both prediction and interpretability to explain which features are more characteristic of which class. While many experiments have produced good results, the best model used hand-crafted explicit linguistic features, which indicates that, especially for the Digital Humanities, feature engineering may still be desirable, at least for their high interpretability. Our findings indicate that word ngrams are the most discriminating features to differentiate between authentic and fake Hadiths. Specifically, word ngrams, where the value of n ranges from 1 to 3, produces the best results (78.28%) using Multinomial Naive Bayes (MNB) as a classification method.

While we have used a large number of standard features, we acknowledge that this study is not without limitations: first, the corpus used in this study only has morphological segmentation annotation, but is not annotated with part of speech tags. The reason for this is that we have tested the quality of some available POS taggers for Arabic and we have found that their accuracy is poor. We plan to address this in a future study where we modify existing POS taggers to process classical Arabic and measure its effect on authorship attribution. Second, we have used hand-crafted feature extraction; however, we also plan to explore more automated methods of feature extraction, especially those that use neural language models and deep learning classification.

References

- Abdelaal, Hammam M and Hassan A Youness (2019). "Hadith Classification using Machine Learning Techniques According to its Reliability." In: *SCIENCE AND TECHNOLOGY* 22.3-4, pp. 259–271.
- Al-Albani, Nasser (1992). *The book of weak and inauthentic ḥadīth*. Riyadh, Saudi Arabic: Dar Al-Ma'arif.
- Aldhlan, Kawther A et al. (2012). "Novel mechanism to improve hadith classifier performance." In: *2012 international conference on advanced computer science applications and technologies (ACSAT)*. IEEE, pp. 512–517.
- Azmi, Aqil M and Amjad M AlOfaidly (2014). "A novel method to automatically pass hukm on hadith." In: *5th Int. Conf. on Arabic Language Processing (CITALA'14)*, pp. 26–27.
- Biber, Douglas (1986). "Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings." In: *Language* 62.2, pp. 384–414.
- Bilal, K and S Mohsen (2012). "A Cloud-based Distributed Expert System for Classification of Ahadith." In: *10th International Conference on Frontiers of Information Technology, Islamabad, IEEE*.
- Binbeshr, Farid, Amirrudin Kamsin, and Manal Mohammed (2021). "A Systematic Review on Hadith Authentication and Classification Methods." In: *Transactions on Asian and Low-Resource Language Information Processing* 20.2, pp. 1–17.
- Brown, Jonathan A. C. (2008). "How We Know Early Hadīth Critics Did Matn Criticism and Why It's so Hard to Find." In: *Islamic Law and Society* 15.2, pp. 143–184.

- Eder, Maciej (2015). "Does size matter? Authorship attribution, small samples, big problem." In: *Digit. Scholarsh. Humanit.* 30.2, pp. 167–182.
- Elewa, Abdelhamid (2018). "Authorship verification of disputed Hadiths in Sahih al-Bukhari and Muslim." In: *Digital Scholarship in the Humanities* 34.2, pp. 261–276.
- Ghanem, Mohamed, Abdelaaziz Mouloudi, and Mohammed Mourchid (2016). "Classification of hadiths using LVQ based on VSM considering words order." In: *International Journal of Computer Applications* 148.4.
- Ghazizadeh, Mehdi et al. (2008). "Fuzzy expert system in determining Hadith 1 validity." In: *advances in computer and information sciences and engineering*. Springer, pp. 354–359.
- Grieve, Jack (2007). "Quantitative Authorship Attribution: An Evaluation of Techniques." In: *Literary and Linguistic Computing* 22.3, pp. 251–270.
- Hadjadj, Hassina and Halim Sayoud (2016). "Towards an authorship analysis of two religious documents." In: *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*. IEEE, pp. 369–373.
- Hassaine, Abdelaali, Zeineb Safi, and Ali Jaoua (2016). "Authenticity detection as a binary text categorization problem: Application to Hadith authentication." In: *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*. IEEE, pp. 1–7.
- Juynboll, G. H. A (2007). *Encyclopedia of canonical ḥadīth*. Leiden, The Netherlands: Brill.
- Kabir, Muhammad Nomani, Md Munirul Hasan, et al. (2018). "Development of a web-extension for authentication of online Hadith texts." In: *International Journal of Engineering & Technology* 7.2.5, pp. 19–22.
- Kabir, Muhammad Nomani, Omar Tayan, et al. (2019). "On the development of a web extension for text authentication on Google Chrome." In: *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, pp. 1–5.
- Le, Trang T, Weixuan Fu, and Jason H Moore (2020). "Scaling tree-based automated machine learning to biomedical big data with a feature set selector." In: *Bioinformatics* 36.1, pp. 250–256.
- Lucas, Scott C. (2008). "Major Topics of the Hadith." In: *Religion Compass* 2.2, pp. 226–239.
- Mohamed, Emad and Zeeshan Sayyed (2019). "Arabic-SOS: Segmentation, Stemming, and Orthography Standardization for Classical and pre-Modern Standard Arabic." In: *DATECH2019*.
- Nutanong, Sarana et al. (2016). "A scalable framework for stylometric analysis query processing." In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, pp. 1125–1130.
- Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12, pp. 2825–2830.
- Rebora, Simone et al. (2019). "Robert Musil, a war journal, and stylometry: Tackling the issue of short texts in authorship attribution." In: *Digit. Scholarsh. Humanit.* 34.3, pp. 582–605.
- Renard, John (2015). *Historical dictionary of Sufism*. Rowman & Littlefield.
- Sarwar, Raheem, Qing Li, et al. (2018). "A scalable framework for cross-lingual authorship identification." In: *Information Sciences* 465, pp. 323–339.
- Sarwar, Raheem and Sarana Nutanong (2016). "The Key Factors and Their Influence in Authorship Attribution." In: *Res. Comput. Sci.* 110, pp. 139–150.
- Sarwar, Raheem, Thanasarn Porthaveepong, et al. (2020). "StyloThai: A scalable framework for stylometric authorship identification of thai documents." In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19.3, pp. 1–15.
- Sarwar, Raheem, Norawit Urailetrprasert, et al. (2020). "CAG: Stylometric Authorship Attribution of Multi-Author Documents Using a Co-Authorship Graph." In: *IEEE Access* 8, pp. 18374–18393.

- Sarwar, Raheem, Chenyun Yu, et al. (2018). “An effective and scalable framework for authorship attribution query processing.” In: *IEEE Access* 6, pp. 50030–50048.
- Savoy, Jacques (2015). “Comparative evaluation of term selection functions for authorship attribution.” In: *Digit. Scholarsh. Humanit.* 30.2, pp. 246–261.
- Sayoud, Halim (2012). “Author discrimination between the Holy Quran and Prophet’s statements.” In: *Literary and Linguistic Computing* 27.4, pp. 427–444.
- (2015). “Segmental analysis-based authorship discrimination between the holy quran and prophet’s statements.” In: *Digital Studies/Le champ numérique* 6.1.
- (2018). “Visual Analytics Based Authorship Discrimination Using Gaussian Mixture Models and Self Organising Maps: Application on Quran and Hadith.” In: pp. 158–164.
- Sayoud, Halim and Hassina Hadjadj (2017). “Fusion based authorship attribution-application of comparison between the Quran and Hadith.” In: *International Conference on Arabic Language Processing*. Springer, pp. 191–200.
- Shaker, K. and D. Corne (2010). “Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis.” In: *2010 UK Workshop on Computational Intelligence (UKCI)*, pp. 1–6.
- Shatnawi, Mohammed Q, Qusai Q Abuein, and Omar Darwish (2011). “Verification hadith correctness in islamic web pages using information retrieval techniques.” In: *Information & Communication Systems*, p. 164.
- Sheikha, Fadi Abu and Diana Inkpen (2012). “Learning to Classify Documents According to Formal and Informal Style.” In: *Linguistic Issues in Language Technology* 8.
- “Stacked generalization” (1992). In: *Neural Networks* 5.2, pp. 241–259.