IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 October 2022, accepted 21 October 2022, date of publication 26 October 2022, date of current version 3 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217457

RESEARCH ARTICLE

Fabricated Hadith Detection: A Novel Matn-Based Approach With Transformer Language Models

KAMEL GAANOUN[1] **AND MOHAMMED ALSUHAIBANI**[2]
[1]Association for Business Intelligence (AMID), Beni Mellal 23000, Morocco
[2]Department of Computer Science, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia

Corresponding author: Mohammed Alsuhaibani (m.suhibani@qu.edu.sa)

The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

ABSTRACT Muslims rely primarily on the Quran and the Hadiths in all their spiritual life and consider them as sacred sources. If the Quran is God's word, then the Hadiths are God's instructions in the words of the Prophet Muhammad. Since Hadiths are transmitted through multiple narrators, they have been extensively studied to ensure their authenticity. The purpose of this study is to detect fabricated Hadiths, or Mawdu, which is the type of Hadith most rejected by Muslim scholars. The study utilises the central text and content of Hadith, Matn, rather than solely focusing on Hadith chain of narrators, Sanad. In order to accomplish this, we create and release the first dataset dedicated to Mawdu Hadiths, called MAHADDAT. Furthermore, we set up a Mawdu Hadith (MH) detection system based on a transformer language model, BERT, achieving a 92.47% $\mathrm{F1}_{MH}$ score.

INDEX TERMS Arabic NLP, BERT, fabricated Hadiths, Hadtih authentication, Mawdu, transformers.

I. INTRODUCTION

The development of automated classification systems has been necessitated recently due to the rapid increase in data. The majority of the data is presented as text. Hence the need for text classification which involves grouping texts into one or more pre-established categories or classes. Such a process may organize, arrange, and categorize virtually any sort of text, including files, documents, and text from the internet. For instance, news can be classed by its authenticity [1], articles by topic [2], service requests by urgency [3], and social media status by sentiment [4], to name a few examples.

In reality, the work in the area of text classification is mostly applied to English language texts using Machine Learning (ML) algorithms, with relatively much fewer works in other natural languages [5]. Even though Arabic is among the top fifth spoken languages in the world with more than 20 countries having it as the official language and more than 400 million native and non-native speakers [6], there are fewer attempts of classifying Arabic texts. It is even considerably less when it comes to religious-related Arabic texts. One of the two pillars of Islam, together with the holy Quran (i.e. the sacred scripture of Islam), is Hadith (i.e. the sayings of Prophet Muhammad Peace Be Upon Him (PBUH)) and both are the principal references for more than 1.5 billion Muslims around the world [7].

Arabic word Hadith literally means the discourse of a person. A recount of the teachings, acts, and sayings of the Prophet Muhammad PBUH is what Hadith means in the context of religion [8], [9]. *Sanad* (chain of narrators back to the Prophet) and *Matn* (the actual content and central text) are the names of the two Hadith's primary components. They combine to form the fundamental elements of each Hadith. To organise Hadiths according to their topics (a.k.a Hadith classification), scholars began classifying Islamic literature in antiquity [10]. Scholars have also paid a great deal of thought to deciding the authenticity of Hadith. Consequently, rules and procedures were devised for achieving that goal of determining the authenticity degree of Hadith (a.k.a. Hadith authentication) as Sahih (accurate or correct), Hasan (good), Daif (weak) or Mawdu (Fabricated) [11]. Mawdu Hadiths (MHs) are considered to be the worst among non-authentic Hadiths, because they are made-up, manipulated or fabricated Hadiths that are falsely attributed to the Prophet PBUH [12].

From ML and Natural Language Processing (NLP) perspectives, Hadith classification and authentication can be

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

viewed as text classification practices, since they are both concentrating on classifying Arabic text into some predefined classes. However, in Hadith science, the classification of Hadith according to the topics known as *Tasneef Al Hadith* whereas *Takhreeg Al Hadith* is the reference for the authentication of Hadith and assigning its authenticity degree. As such, the work presented in this paper will be solely on Hadith authentication with a particular focus on detecting MHs.

Hadith scholars consult both the Sanad and Matn to establish the reliability of a certain Hadith. Because the narrators in the Sanad must not be disconnected, researchers examine each narrator's status to determine whether or not they are consistently trustworthy and connected [13]. Hadith scholars also examine Hadith's Matn to see if it agrees or disagrees with the grammar of Arabic, authentic Hadiths, or what is being said in the Quran. Because there is a possibility that Matn uses unsuitable language or expressions that do not align with Muslims beliefs, authentic Hadiths or the discourse of the Prophet Muhammad PBUH [14], [15], [16] which in that case falls under the definition of MHs.

The last few years have witnessed some efforts in the literature in regards to automatically authenticating Hadiths [15], [16], [17], [18], [19], [20]. Such efforts solely focus on the Sanad in order to assign the Hadith's degree of authenticity. Moreover, almost all the available studies focus on authenticating whether a Hadith is Sahih or not, none were solely focusing on studying the nature of the MHs nor building the authentication process around it, which, we argue, is a very vital way to look into the problem.

Indeed, the Sanad is there to explain the validity of Hadith. Nonetheless, nowadays, the majority of Muslims reference Hadiths without stating their Sanad, in contrast to the early Islamic era [14]. Additionally, with more people having access to the internet and social media nowadays, the problem has grown because the number of fabricated Hadiths is steadily increasing. Therefore, it is critically important to study the problem of Hadith authentication considering the Matn, which is the gap that is being addressed in this paper.

ML algorithms have been the foremost and sole technique in all the previously presented studies related to Hadith authentication as highlighted in the related work section next. However, most recently, ML and NLP have seen unprecedented breakthroughs with the introduction of Transformer Language Models (TLMs). Such TLMs include Bidirectional Encoder Representations from Transformers (BERT) [21], Generative Pre-trained Transformer (GPT-3) [22], XLNet [23] and Robustly Optimized BERT (RoBERTa) [24]. TLMs report the state-of-the-art results in a large number of ML and NLP tasks [25], [26], [27], [28].

Still, there is yet to be any study utilising TLMs to deal with Hadith authentication. Therefore, the work presented in this paper takes the advantage of utilising all available Arabic TLMs to study Hadith authentication for the first time.

We propose an approach using various Arabic TLMs to deal with Hadith authentication using the central text of Hadiths, the Matn. The aim is to automatically detect MHs using the Matn. In particular, we have utilised AraBERTv2 [29], Arabic-BERT [30], QARIB [31], CAMeLBERT MSA [32], CAMeLBERT CA [32], mBERT base [21] and XLM-RoBERTa base [33] centring to identify the MHs. The contribution of the paper can be summarised as follows:

- A Matn-based MH detection system taking into account studying and understanding the central text and content of Hadith, rather than solely focusing on the chain of narrators.
- Exploiting all the available Arabic TLMs to automatically authenticate Hadiths. To the best of our knowledge, this is the very first work considering TLMs for such a task.
- Proposing two new datasets, called NAH Plus and MAHADDAT, with broad discussion of the creation phases, analysis and statistics prior to the thorough experiments. Both datasets are released and publicly available.
- A main focus to detect MHs and studying their natures among all Hadiths rather than only concentrating on classifying Hadith as Sahih or Daif.
- A comprehensive comparison study between numerous classical ML algorithms and Arabic TLMs in Hadith authentication performance.

The proposed model together with the thorough experiments reveals that employing Arabic TLMs for Hadith authentication and detecting MHs is fully justified by reporting the state-of-the-art results with several metrics and evaluation methods.

The rest of the paper is organized as follows: section II presents the related work, section III mainly describes the data creation, collection and processing steps, section IV is dedicated to the classification methodology, and section V introduces the experiments and results. The results are discussed in section VI. Finally we present our conclusions along with some future directions in section VII.

II. RELATED WORK

Hadiths, together with the Holy Quran, serve as Muslims' primary source of law, hence authenticating Hadiths is essential. It is as well equally crucial to classify Hadiths into groups or topics to make them simpler to search for and identify. The problems of Hadith classification and authentication can be resolved using a variety of NLP techniques. Nevertheless, relatively not many have looked into this in the literature. Although this paper's work focuses primarily on Hadith authentication and detecting MHs, we sought to take a broader approach and review some prior research in Hadith classification.

One of the early works to computationalise the Hadith classification process is backdated when Jbara et al. [34] presented an approach for classifying Hadiths' topics into 13 classes (books) of Sahih Al-Bukhari. Such classes include *faith*, *knowledge*, *praying*, *hajj* (*pilgrimage*), *eclipse*, *almsgiving*, *fasting*, and *medicine*. Similarly, Alkhatib [35] study the effectiveness of categorising Hadiths into 8 different classes (books) using ML classifiers. Later, with a much focus on extracted quote Hadiths from four different books, Al-Kabi et al. [8] train and compare three ML classifiers to predict the four classes. This work was afterwards extended by Al-Kabi et al. [36] with enlarging the Hadiths dataset. Besides, Afianto et al. [37] presented an approach to categorises Hadiths into three predefined categories: *suggestion*, *prohibition*, and *information*. Having Arabic Sahih Al-Bukhari's translated to the Indonesian language, two classification models with backpropagation Neural Network (NN) were proposed by Bakar et al. [38]. Rostam and Malim [39] followed an alternative mode and suggested a technique that uses text categorization to classify particular classes by figuring out how the resources relate to one another. The authors combined various resources comprising Quran and Hadith. Mediamer [40] shifted the focus to the impact of feature extraction and preprocessing towards Hadith classification.

It is obvious that the Matn is used in all of the aforementioned studies on Hadith classification because it functions rather like topic modelling. However, as we will demonstrate in the following paragraphs, such a thing is conspicuously absent from Hadith authentication.

Over a decade ago, Zahedi et al. [41] presented a fuzzy expert system with an ambition to authenticate Hadiths with its rate of validity. The system initiates with domain experts' inputs for developing a knowledge base with some essential rules taking into account the narrators' names, particularly the Sanad, as a main focus for the rating process. Analogously, a combination of expert system and ML techniques was employed by Aldhlan et al. [42] as a new classification method to authenticate 999 Hadiths to their validity degree (e.g. *Sahih* or *Hasan*). In particular, a tree structure model with a Decision Tree (DT) [43] classifier along with selected attributes of the instances extracted from Hadith books. Rather than relying on building or training a model, Shatnawi et al. [44] presented a technique for extracting hadith phrases from web pages and using a positional index created from a database of Hadiths to authenticate Hadiths as *Sahih* or *Daif*.

Moreover, Najiyah et al. [17] opted for building an expert table of Hadith comprising various characteristics and codes based on consultation with domain experts. The intention was to authenticate and classify Hadiths into *Sahih*, *Mawdu* or *Daif* according to such characteristics and codes which will be then used for creating a decision tree and a rule degree of hadith. Similarly motivated, Abdelaal and Youness [7] introduced a ML-based algorithm to authenticate Hadiths based on the characteristics of the narrator such as reliability and memory. Furthermore, Ghanem et al. [19] represents Hadith as vectors in the Vector Space Model (VSM) [45] and Term Frequency-Inverse Document Frequency (TF-IDF) as term weighting indicating its importance in order to classify 160 Hadiths into an authenticity grade.

As opposed to exploiting expert systems or ML techniques, a simple method was proposed by Azmi and AlOfaidly [46]. The scheme essentially automates the process by formulating the rules used by Hadith scholars to authenticate and rate the validity of 2800 Hadiths from Sunan Al-Tirmizi based on the Sanad. Similarly, based on four main criteria concerning only Sanad, namely the reliability and preservation of the narrators, the flaw in the chain of transmission, and connected chain, Ibrahim et al. [13] offered a theoretical authentication framework that would determine if a Hadith is Sahih or not. Taking a different tack at authenticating Hadiths using the Sanad, Balgasem and Zakaria [47] addresses the problem by recognising the Arabic names in the chain of narrators using Part-of-Speech (POS) and Named Entity Recognition (NER).

Although Hadith scholars consult both the Sanad and Matn to establish the authenticity of a certain Hadith [11], all the above-mentioned studies were restricted to using Sanad. The importance of using Matn for Hadith aligns with Hadith scholars examining Hadith's Matn to see if it agrees or disagrees with other authentic Hadiths or what the Quran says [16]. Matn on occasions uses unsuitable language or expressions that do not align with Muslims beliefs or the discourse of the Prophet Muhammad PBUH. To the best of our knowledge, only two previous work has merely focused on Matn for Hadith authentication which will be discussed in the following paragraph.

Firstly, Hassaine et al. [20] explored the possibility of a Hadiths authentication process based solely on the Matn. This was accomplished by maintaining a binary relation (for each class, authentic and non-authentic) approach. Precisely, the proposed work begins with manually extracting keywords of each Hadith, authentic and non-authentic, using hyper rectangular decomposition, and these extracted keywords are then fed into ML algorithms for authentication. Secondly, comprehensive experiments for the evaluation of Hadith authenticity with various ML and deep learning classifiers were lately conducted by Tarmom et al. [16]. For example, Support Vector Machine (SVM) [48], Naïve Bayes (NB) [49] and DT classifiers and Long Short-Term Memory (LSTM) [50], Convolutional Neural Network (CNN) [51] and CNN-LSTM deep learning classifiers. Both Sanad and Matn were utilized in the proposed experiments.

Most recently, numerous AI, ML, and NLP tasks have seen unprecedented and extraordinary results with the help of TLMs. However, no attempts have yet been made to explore the usefulness of exploiting TLMs in Arabic for Hadith authentication or classification. The single attempt for all we know to deal with Hadith using TLMs came lately by Emha et al. [52] dealing with Indonesian Hadiths translated from Arabic. In particular, a semi-supervised BERT with an additional feed-forward neural network was proposed to classify the Indonesian Hadiths. The feed-forward network especially operates on the narrators for Hadith for the execution of NER, for Indonesian Hadith texts in particular. The experiments show that the proposed model utilizing BERT with NER was exceedingly effective.

TABLE 1. NAH and LK corpora content description.

Corpus	LK	NAH
Number of Hadiths	34,088	7,363
Covered Books	Sahih Bukhari, Sahih Muslim, Sunan Abu Dawood, Sunan Al-Nasai Sunan Altarmithi, Sunan Ibn Maja	Miayat hadith daeif wamawdu muntashira bayn alkhutaba walwoaz, Abatil walManakir waSahih walMashahir, Allali almasnua fi alahadith almawdua, Alahadith aldaifa fi kitab riad alsaalihin, Aljadu alhathith fi bayan ma lays bihadith, Alfawayid almajmua fi Ahadith almawdua
Covered Degrees	Mainly Authentic	Mainly Non Authentic
Content	Book, Chapter_Number, Chapter_English, Chapter_Arabic, Section_Number, Section_English, Section_Arabic, Hadith_number, English_Hadith, English_Isnad, English_Matn, Arabic_Hadith, Arabic_Isnad, Arabic_Matn, Arabic_Comment, English_Grade, Arabic_Grade	Book, Hadith_Number, Full Hadith, Isnad, Matn, Author's Comments, Degree, Authenticity, Topic

To overcome the various limitations that exist in the literature concerning Hadith authentication, we propose the first thorough study exploiting Arabic TLMs for Hadith authentication and MHs detection using the central content of Hadiths, the Matn.

III. DATA

A. AVAILABLE CORPORA

In the systematic review proposed by Binbeshr [11], it was concluded that Sahih Al-Bukhari is the most widely employed corpus in Hadith studies. It was also stated that almost all the datasets used in Hadiths literature are not publicly available. Although the number of books devoted to Hadith narration is considerable, we rarely dispose of a structured digital version ready to process. While most of the efforts have been directed at Sahih Al-Bukhari, the existing works do not give access to the used subsets of the book. A second disadvantage is that Sahih Al-Bukhari focuses solely on Sahih Hadiths (SHs), which does not entirely meet the objective of our study. It is only newly that Non-Authentic Hadiths (NAH) [53] corpus was created and made public, becoming the first corpus dedicated to non-authentic Hadiths. In addition, Leeds and King Saud University (LK) corpus [54] was also lately published. The latter gathers the Hadiths from the six most well-known books of Hadiths concerning SH; known as *Al-Sihah Al-Sittah*, or "The Authentic Six". Although not all of the Hadiths in these books are authentic, their name derives from the fact that most of them are considered authentic.

Afresh Hadith corpus covering 9 books of Hadith was also made public [55], which includes the same books of the LK corpus as well as the contents of *Musnad Ahmad Ibn Hanbal*, *Malik Muwatta*, and *Sunan Al Darimi*. It contains more Hadiths than LK, however, it does not distinguish between the Matn and the Sanad of Hadith. Indeed, LK and NAH corpora explicitly distinguish between Sanad and Matn to facilitate the work on them. For these reasons, we opt to base our work on these two corpora by adapting them to our problem, as described in the next section. We demonstrate the details of LK and NAH corpora in Table 1.

B. CORPUS CREATION PROCESS

To mimic the actual preponderance of MHs, we used the LK and NAH corpora as starting point to obtain an unbalanced final corpus for Mawdu and authentic Hadiths. For this purpose, we apply the processing steps described in the following subsections. The whole corpus creation process is demonstrated in Fig. 1.

1) CLEANING PHASE

As a means of guaranteeing an optimal quality of input for the models, we clean the variables concerning the Matn and the degree of authenticity of the Hadiths as follows:

LK corpus cleaning relies on the *Arabic_Matn*, *English_Grade*, and *Arabic_Grade* fields, and apply cleaning decisions below:

- Using Dorar[1] to check Hadiths when Arabic and English grades differ → 2 Hadiths
- Removing Hadiths with empty Arabic Matn → 826 Hadiths
- Removing Hadiths with empty English and Arabic grades → 380 Hadiths
- Removing Hadiths with no English grade and an Arabic grade that we were not able to classify as authentic or not → 32 Hadiths
- Keep the following grades: Sahih - Authentic, Daif - Weak, Hassan - Good, Hassan - Sahih, Mawdu - Fabricated, Munkar → removed 708 Hadiths

For the NAH corpus we use Matn and Degree fields for the subsequent cleaning steps:

- Removal of Hadiths without Matn → 1,246 Hadiths
- For Hadiths without degree information, we scrap the degree from Hdith website,[2] see next section III-B2 → 3,352 Hadiths
- Removal of authentic Hadiths → 359 Hadiths
- Grouping Hadiths with degrees meaning Mawdu and leave the remaining as Daif. Degrees classified as

[1] https://dorar.net/

[2] http://hdith.com/

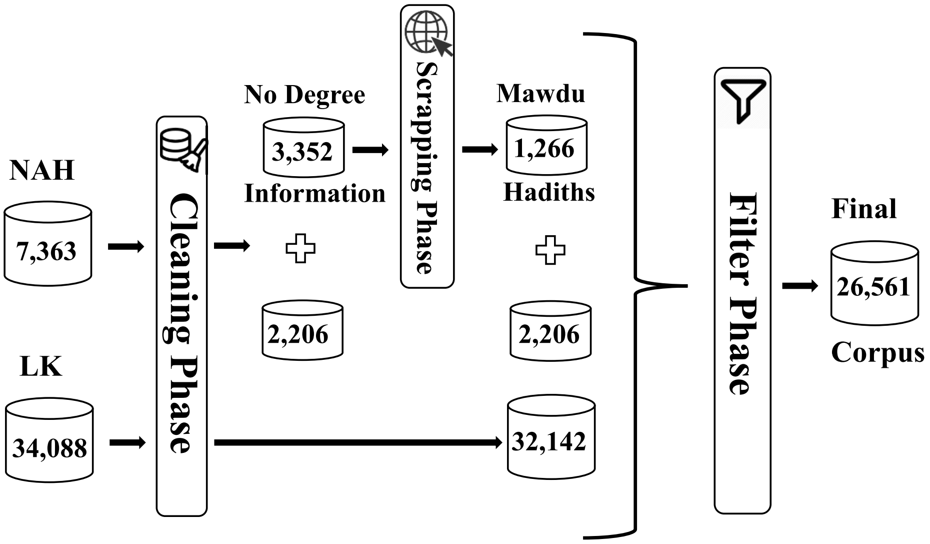


FIGURE 1. Corpus creation process.

Mawdu are : مفترى ملحون, باطل, موضوع, كذب (false, fabricated, a lie, slanderous).

2) SCRAPPING PHASE

In light of the very limited number of MHs available, we decided to further add additional MHs. To accomplish this, we scraped the two well known websites specialized in Hadith indexing, Dorar and Hdith. Compared to Dorar, Hdith has a better search engine that can identify Hadiths even when they are misspelled. Hdith was used to recover the authenticity degrees of 3,552 Hadiths that lacked this information in the NAH corpus. While Dorar was utilized to scrape additional MHs not present in the NAH corpus.[3] This method recognized 900 Hadiths out of the 3,552 Hadiths as Mawdu from Hdith. Dorar has a section titled "Widespread but unauthentic Hadiths", from which we scraped all 1,315 Hadiths, including 366 MHs after filtration.

After the cleaning and scrapping phases, we obtain an enhanced version of the NAH corpus, which we call NAH Plus, which is now publicly available.[4] NAH Plus contains only non-authentic Hadiths, with available Matn in addition to the Hadiths for which we have recovered the degrees of authenticity. This corpus contains a total of 3,660 non-authentic Hadiths.

3) FILTER PHASE

We filter the NAH Plus dataset to extract the MHs, and the same process was applied to the LK corpus, which contains 29 MHs. In addition, we add the 366 MHs scraped from Dorar and delete any duplicates. As a result, we create a new corpus dedicated to MHs, with a total of 2,452 Hadiths. We call this dataset **MA**wdu **HAD**ith **DAT**aset or **MAHADDAT** meaning "Never narrated" in Arabic. To the best of our knowledge, this is the first dataset dedicated to MHs. We have released and made the dataset publicly available to encourage more research on this field.[5]

For the final dataset, we mix MAHADDAT with a filtered version of LK corpus retaining only authentic SHs, and we remove the duplicated Hadiths. In addition, Hadiths with a Matn indicating their authenticity degree were identified and removed from the LK corpus. In order to avoid any indication for the classification models, we removed Hadiths containing the words صحيح authentic or حسن good.

In addition to the previous steps, we applied an extra step for the LK corpus. In fact, the creators of this corpus state in their paper the following: "*in our corpus we incorporated the Prophet in the Matn instead of Isnad*". The Isnad/Sanad usually includes the chain of successive narrators that ends with an expression related to the Prophet, such as قال رسول الله صلى الله عليه وسلم "The prophet Peace Be Upon Him said.", أن رسول الله صلى الله عليه وسلم قال "That the Messenger of God, may God bless him and grant him peace, said", عن النبي صلى الله عليه وسلم أنه قال "On the authority of the Prophet, may God bless him and grant him peace, that he said". It is worth mentioning that including these expressions in the Sanad creates a bias for the models, since these expressions are not part of the Matn. Moreover, since the NAH corpus does not include these expressions, the models will have an indication towards the Hadiths Sahih. We therefore decided to eliminate these expressions in order to eliminate any potential bias. That said, the results without this cleaning step were also obtained (and presented in the supplementary file submitted alongside this paper), where we

[3] Python selenium 4.1.5 library was used for the scraping phase.

[4] https://github.com/kamelgaanoun/mhdetection/tree/main/nah_plus

[5] https://github.com/kamelgaanoun/mhdetection/tree/main/mahaddat

TABLE 2. Description of the final corpus (MAHADDAT combined with the enhanced LK).

	Number of Hadiths	Number of characters			Number of words			
		Including diacritics	Excluding diacritics	Excluding diacritics and punctuation	Total	Average	Median	Maximum
Sahih	24,109	7,937,516	5,025,378	4,969,969	1,016,116	∼42	29	1,698
Mawdu	2,452	576,345	542,933	528,824	109,836	∼45	16	5,608
Total	26,561	8,513,862	5,568,312	5,498,794	1,125,952	∼42	28	5,608

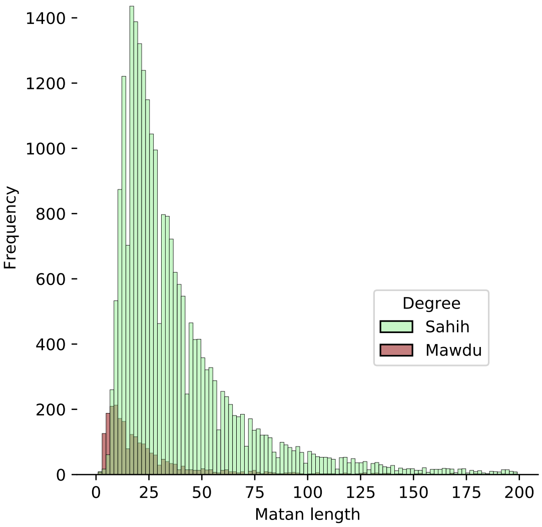


FIGURE 2. Matn length distribution (with extreme length omitted).

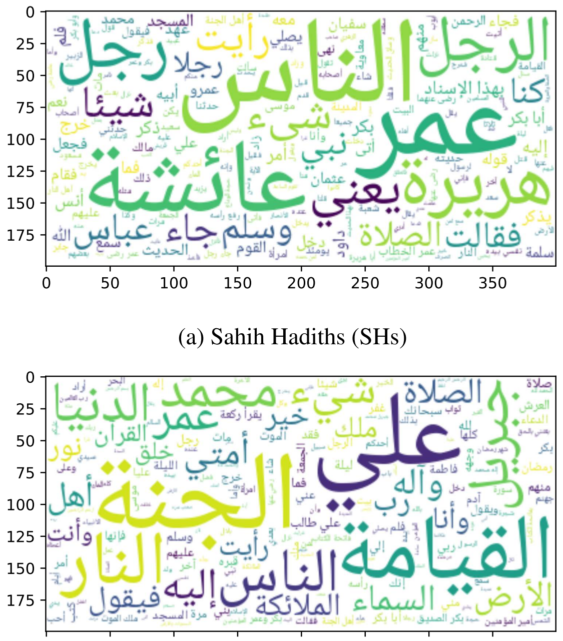
observe better results than our final scores due to the bias raised above.

C. DESCRIPTION OF USED CORPUS

There are 26,561 Hadiths in the final corpus, of which 24,109 are Sahih and 2,452 are Mawdu, representing 9.23% of the total. We describe the details of this dataset in Table 2. While the average number of words is quite similar for Sahih and MHs, 42 and 45 respectively, the median number is not. In fact, if we omit the extremes, the MHs have much lower words than the SHs (see Fig. 2). Parallel to this, for longer Hadiths, Mawdu ones can be three times longer than SHs, attaining a maximum of 5,608 words.

We also analyzed the most frequent words for both types of Hadiths, Sahih and Mawdu. For this purpose, we first eliminated the Arabic-specific stopwords derived from the NLTK 3.7 library.[6] In addition, we built a list of stopwords specific to Hadiths, including words like عز وجل : Almighty, الله: God, فقال : and he said, سمعت : I heard, النبي : The prophet. The complete list of the 515 additional stopwords is released and publicly available.[7]

We visualize the result of this analysis with the help of the wordcloud in Fig. 3a for the SHs and Fig. 3b for the MHs.



(a) Sahih Hadiths (SHs)

(b) Mawdu Hadiths (MHs)

FIGURE 3. Corpus wordcloud.

For the SHs, we notice rather the recurrence of names of companions who were very intimate with the Prophet PBUH and who are known by the important number of Hadiths that they reported. We see for example the name of عائشة Aisha, one of the wives of the prophet PBUH and أبو هريرة Abu Hurairah, a companion who rarely separated from the prophet.

While for the MHs, we note the recurrence of words relating encouragement towards good deeds and rewards hoped for by the faithful, such as الصلاة Prayer, الجنة Paradise, and also words serving to frighten towards النار the hell or القيامة the day of resurrection for example. In addition we also note a fairly high frequency for the word علي which can mean "on me" or the first name of a companion among the four rightly guided caliphs, and who is also the first Imam followed by the Shia group. In order to estimate the prevalence of the word علي related to the companion whose full name is علي ابن ابي طالب Ali Ibn Abi Talib and since the MHs do not have diacritics, we have relied on the context in which

[6] https://www.nltk.org/

[7] https://github.com/kamelgaanoun/mhdetection/tree/main/stopwords

this word is quoted. Indeed, we define a window of 10 words before and after the word علي whenever it is encountered. Then, we consider it to be the companion when one of the following rules is verified:

- The delimited window includes one of the following words: طالب, أمير, رضى, الإمام, الحسين which are globally qualifiers or words strongly related to the companion
- The word علي is directly preceded by one of the following words: أين, مع, إلى, يا (Hey, To, With, Where is) which are prepositions that cannot occur before the word علي meaning "on me"

Based on this method, we concluded that at least 47% of instances of the word علي concern the companion in the MHs and only 20% in the SHs.

IV. CLASSIFICATION METHODOLOGY

A. ML MODELS

1) TEXT REPRESENTATION

As ML algorithms cannot process texts directly, a preliminary step called text representation or vectorization is required. As part of this step, each document is represented by a vector, whose components are, for example, its words, so that the learning algorithms can exploit them [56]. As a result, a collection of texts can be represented by a matrix whose rows are the terms that appear at least once and whose columns are the documents. This matrix generally contains weights assigned to each word, depending on the method used to calculate these weights, we get different matrices. These weights correspond to the contribution of each word to the semantics of its document. A commonly used approach in this field is TF-IDF. This method weights each word based on its frequency in all documents, while giving advantage to rare words. Thus for a word w and a document d, the TF-IDF is calculated according to the following equation:

$$TF-IDF(w,d) = frequency(w,d) \times \log \frac{D}{D_w} \qquad (1)$$

with:

- frequency(w, d): Number of occurrences of w in d
- D: Total number of documents
- D_w: Number of documents containing the word w

The expression in (1) may differ slightly from one implementation to another. We use the one from the Pyhton Sklearn library[8] which is written as follows:

$$TF-IDF(w,d) = frequency(w,d) \times (\log \frac{D}{D_w} + 1) \qquad (2)$$

In addition, we also experiment with another variation of this method using a logarithmic transformation of frequency(w, d) in order to reduce the importance of terms with high frequency [57], and we note this variation as LogTF-IDF.

The large number of words in the corpus can lead to large matrices, affecting the complexity and accuracy of the models. By using a dimension reduction method, we can keep the most important features while restricting the number of features. The method used here is the Singular Value Decomposition (SVD) [58].

2) EXPERIMENTED ML MODELS

We present below an overview of the different ML models used in our experiments:

Random Forest (RF) [59] is a decision tree-based ensemble learning technique. Multiple decision trees are created using data sets that have been split from the original data. During each stage of the decision tree, a subset of variables is randomly selected. The model then selects the mode for all predictions in each decision tree.

Logistic Regression (LR) [60] is based on the concept of linear regression but adapted to the case where the explained variable takes discrete values. It also has the particularity of predicting not the value of the variable itself, but rather the probability of occurrence of an event. For the case of a variable with two values, we refer to it as binary logistic regression and the outcome will be a probability of occurrence of the event bounded between 0 and 1. Logistic regression is modeled according to the following equation:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (3)$$

where:

- p is the probability of event occurrence and $1 - p$ is the probability of failure.
- β_0 to β_n are the regression coefficients.
- X_1 to X_n are the independent variables.

Naïve Bayes (NB) is a probabilistic classifier based on the Bayes theorem [61]. Assuming all explanatory variables are considered independently, this algorithm relies on a strong assumption. The term *naive* comes from the fact that we assume this independence of the variables. In our binary classification, for example, NB will assume that the words in a document *appear* independently of each other.

Support Vector Machine (SVM) will find a hyperplane or boundary between the two classes of data (for a binary classification problem) that will maximize the margin between the two classes. There are many planes that can separate the two classes, but only one plane can maximize the margin or distance between the classes.

Gradient Boosting (GB) [62] as the name suggests, it utilizes two main concepts, Gradient and Boosting. Boosting is an iterative method consisting in reinforcing successive models at each iteration by giving more weight to cases with high values with respect to the loss function, these cases are called difficult cases. Boosting is a kind of method allowing the model to learn from its previous errors. Gradient is

[8] https://scikit-learn.org/

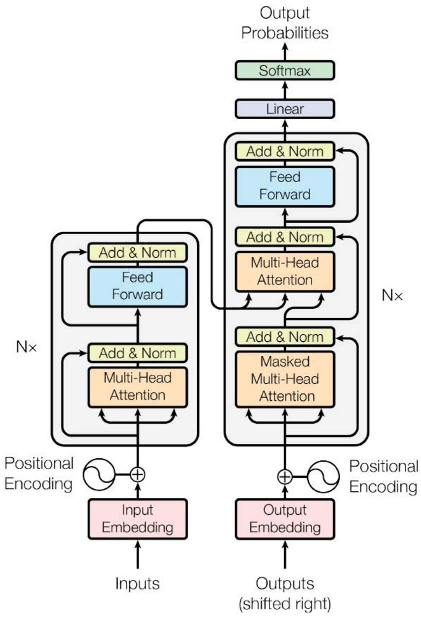


FIGURE 4. Transformer model (as described in the original paper [65]).

the optimization method that allows to minimize the loss function.

Xtreme Gradient Boosting (XGBoost) [63] is a particular implementation of the GB algorithm with more advanced approximation methods, such as the use of second order gradients, as well as a better generalization using the L1 and L2 regularization methods.

Light Gradient Boosting Machine (LGBM) [64] is another decision tree-based algorithm that is faster and uses less memory than XGBoost. Additionally, it splits the decision tree differently from XGBoost. As a matter of fact, LGBM splits the tree leaf-wise, unlike XGBoost, which splits the tree level-wise.

B. BERT MODELS

1) DEFINITION

BERT is based on Transformers, which consists of two distinct blocks: an encoder that reads the input text and a decoder that predicts the task. In BERT, only the encoder block is involved since the goal is to generate a language model with the main objective of creating an attention mechanism that learns the contextual relationships between words (or subwords) in a text. The architecture of this encoder block is composed of several attention layers, and the number of the latter differs according to the version of BERT. Indeed, BERT comes in two architectural variations, BERT base and BERT large, with the first containing 12 attention layers and the latter having 24. Fig. 4 illustrates this architecture.

For the training purpose, BERT relies on two training tasks, namely the Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, 15% of the words are randomly masked with the [MASK] token, followed by replacement of 10% of these tokens by random words, and 10% by the original word. The model aims to predict the masked tokens iteratively until convergence is reached. As for NSP, it uses pairs of sentences as inputs, the model predicts if the second sentence is the next sentence in the original document. The inputs consist of 50% pairs where the second sentence is the actual next sentence, and 50% pairs where a random sentence is used.

TABLE 3. Selected BERT models configurations.

Model	Dataset Size	No. of Tokens	No. of Parameters	Arabic Variants
AraBERTv2	77GB	8.6B	135M (base) 369M (large)	MSA
Arabic-BERT	95GB	8.2B	110M (base) 336M (large)	MSA DA
ARBERT	61GB	6.2	163M	MSA
QARIB	-	14B	135M	MSA DA
CAMeLBERT_MSA	107GB	12.6B	109M	MSA
CAMeLBERT_CA	6GB	847M	109M	CA
mBERT_base	-	1.5B	167M	MSA
XLM-RoBERTa_base	2.5TB	295B	278M	MSA DA

The aforementioned process is called pre-training phase. During this stage, BERT acquires knowledge related to the examined language, constructing a language model capable of understanding the relationships between the words of the language in question. It is a major advance in NLP in that, based on the knowledge acquired earlier by BERT, it is possible to apply transfer learning in a second phase called fine-tuning using a limited volume of data in a specific domain.

2) AVAILABLE BERT MODELS

When BERT was initially designed only for English, the first multilingual model based on the BERT architecture, called mBERT, soon followed. It used Wikipedia from the 104 most commonly represented languages for its training. Following that, Facebook AI researchers released XLM-RoBERTa, a second multilingual model based on CommonCrawl [66]. While both models partially support the Arabic language, they were limited concerning Arabic-related downstream tasks, and the need for Arabic-specific models became increasingly persistent. This need for a specialized model trained uniquely on the Arabic language, and able to achieve better performances gave birth to several successful models such as AraBERT [29], ArabicBERT [30], ARBERT [67], QARIB [31], CAMeLBERT [32]. As the Hadiths are written in Classical Arabic (CA) and Modern Standard Arabic (MSA), we only mention the BERT models that apply to these two variants and ignore models for Dialectal Arabic (DA).

As illustrated in Table 3, CAMeLBERT_CA [68] is the only model trained on a CA corpus, with the rest being mainly based on MSA. Aside from the multilingual models, all the MSA models use similar sources, unlike CAMeLBERT_CA, which uses Open Islamicate Texts Initiative Corpus (OpenITI) [68], with more than 11,000 Islamic books and 7.5M pages.

V. EXPERIMENTS AND RESULTS

A. EVALUATION

In order to make this work comparable and encourage more research in this field, in addition to making our dataset publicly available, we present our results using various metrics commonly employed for supervised classification. Along with the standard metrics of accuracy, precision, and recall, we also use Area Under ROC Curve (AUC) and F1-measure that are more appropriate to the problem of unbalanced data. We introduce below each of the used metrics.

Accuracy describes the effectiveness of a model in correctly predicting both positive and negative individuals in a symmetrical way. It measures the rate of correct predictions for all individuals, and it is generally presented in the form of the following ratio:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

With:

- TP: the individuals that the model was able to predict as positive (individuals concerned by the studied event)
- TN: the individuals that the model was able to predict as negative (individuals not concerned by the studied event)
- FP: the individuals that the model wrongly predicted as positive
- FN: the individuals that the model wrongly predicted as negative

For the case of our study, TP are the Hadiths correctly predicted as Mawdu, while TN are the Hadiths correctly predicted as Sahih, and this definition holds for the rest of metrics using Positive and Negative terms.

One of the limitations of this metric is that it is only meaningful for datasets with equal distribution of classes.

Precision helps answer the question, ‘‘What proportion of MH predictions, were actually correct?’’, and is calculated as the ratio between TP to the total number of individuals predicted as Mawdu as in (5):

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall allows to answer the question, ‘‘What proportion of true MH results were correctly identified?’’, and is calculated as the ratio between TP to the total number of MH as in (6):

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

F1-measure evaluates the ability of a classification model to efficiently predict positive individuals (MH in our case), by making a trade-off between precision and recall. It is particularly used for tasks dealing with unbalanced data. The F1-score summarizes the precision and recall values in a unique metric as expressed in (7):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

We use the F1-score as a reference metric to compare the different models since it is a robust metric concerning unbalanced datasets and offers a summary of the precision and recall. Nevertheless, because this study is more concerned with detecting Mawdu Hadith than Sahih Hadith, the metric adopted is the $F1_{MH}$ specific to Mawdu Hadith. In parallel, we provide the F1 to summarize the results of the F1 measure for the two classes of Hadiths.

AUC is based on the Receiver Operating Characteristics (ROC), a probability curve where True Positive Rates (TPR) are plotted against False Positive Rates (FPR) for different thresholds. The AUC is then the area under this ROC curve, and summarizes the curve with a range of threshold values as a single score. AUC is an other used metric for unbalanced data, that said, this metric remains controversial in the literature [69], [70], we therefore use it as complementary information to the F1-score.

B. CONFIGURATION

1) ML MODELS

We conducted experiments with four different scenarios depending on the settings of the TF-IDF and SVD. We list these four scenarios below:

- TF-IDF 1 : TF-IDF maximum features=5000, SVD number of components =15
- TF-IDF 2 : TF-IDF maximum features=8000, SVD number of components =20
- LogTF-IDF 1 : LogTF-IDF maximum features=5000, SVD number of components =15
- LogTF-IDF 2 : LogTF-IDF maximum features=8000, SVD number of components =20

The rest of the parameters are kept at their default value defined by the Sklearn library and we consider unigrams, bigrams and trigrams by setting the ngram_range parameter to the value (1,3). In addition, each of these four scenarios was evaluated using light-stemming [71] to assess the impact of this method. The whole parameterizations details of the ML models is presented in the supplementary file attached with the paper.

We split the dataset into a training set (Train), reserved for the training process, a development set (Dev) to select and validate the best system, and a test set (Test) to evaluate the adopted system and produce final results. We chose an 80/10/10 partition scenario for this split, while maintaining the distribution of the two labels. We openly share[9] these partitions for future researchers to ensure comparability of our results.

2) BERT MODELS

Due to the fact that the selected pre-trained models aren't originally intended to classify Hadiths, they have to be fine-tuned to fit our problem. To do so, we add a classification layer on top of the model architecture combined with a softmax function and feed it with the [CLS] token embedding of the model.

We use the Trainer class from the HuggingFace Transformers 4.5.1 library to train and evaluate the models. As the

[9] https://github.com/kamelgaanoun/mhdetection/tree/main/corpus

TABLE 4. Classification results on Dev set using ML models and second Log TF-IDF/SVD configuration.

Model	Acc.	$F1_{MH}$	AUC	F1	Prec.	Rec.
RF	92.96	**51.93**	69.72	**74.06**	82.20	69.72
LR	90.78	1.61	50.37	48.38	70.42	50.37
NB	77.48	37.71	**75.86**	61.98	61.01	**75.86**
SVM	91.23	10.04	52.63	52.71	**92.04**	52.63
GB	91.38	39.90	64.27	67.63	74.59	64.27
XGBoost	92.17	39.88	63.42	67.85	80.71	63.42
LGBM	**93.07**	50.54	68.50	73.41	84.02	68.50

problem is a single label classification one, the used loss function is cross-entropy.

Except for the batch size, all the models share the same hyperparameters. Due to GPU memory limitations, the batch size is set to 64 for base models and 16 for large models. As for the remaining hyperparameters, we set the learning rate to $2e^{-5}$ and the maximum input length to 128 tokens. NVIDIA Tesla P100-PCIE-16GB and NVIDIA Tesla T4 GPUs were used in these experiments.

In the following subsection, we present the best results obtained on the Dev set, as well as Test set results with the retained configuration. Results for the other configurations were also obtained and are presented in the supplementary file attached with this manuscript. The results of ML models are shown first, followed by the results of BERT models.

C. RESULTS

1) ML MODELS RESULTS

Table 4 presents results for second LogTF-IDF/SVD without stemming configuration, which produces the best scores on the Dev set. We evaluate this setting on the Test set, and illustrate the results in Table 5.

2) BERT MODELS RESULTS

We evaluate all BERT models on the Dev set with 1 and 3 epochs combined with the application of light-stemming. In this section, we present results obtained with the best configuration both on Dev and Test sets. The remaining results are available in the supplementary file.

Table 6 present results on the original Dev set (without stemming) after fine-tuning Models for 3 epochs. As the configuration with no stemming and 3 epochs is providing best results, we apply this setting on the Test set, and present final results in Table 7.

While this work focuses on training and evaluating on unbalanced datasets, we present three other scenarios based on training and evaluation dataset type in Table 8. The balanced datasets are obtained by downsampling the Sahih Hadiths to equal the number of Mawdu Hadiths. We make these datasets available[10] to researchers for future comparison.

TABLE 5. Classification results on non Stemed Test set using ML models and second LogTF-IDF/SVD configuration.

Model	Acc.	$F1_{MH}$	AUC	F1	Prec.	Rec.
RF	**92.89**	**49.60**	68.21	**72.89**	82.76	68.21
LR	90.74	2.38	50.53	48.76	66.86	50.53
NB	78.32	37.93	**75.41**	62.40	61.14	**75.41**
SVM	91.31	10.81	52.86	53.12	**95.63**	52.86
GB	91.31	39.37	64.04	67.34	74.20	64.04
XGBoost	92.10	37.87	62.46	66.82	80.88	62.46
LGBM	92.55	47.34	67.29	71.67	80.88	67.29

TABLE 6. Classification results on Dev set without stemming using BERT models and 3 epochs.

Model	Acc.	$F1_{MH}$	AUC	F1	Prec.	Rec.
AraBERTv2_base	97.89	88.19	92.24	93.51	94.89	92.24
ArabicBERT_base	98.23	90.11	93.34	94.57	95.88	93.34
ARBERT	98.49	91.45	93.49	95.31	97.34	93.49
QARIB	98.23	89.94	92.61	94.48	96.58	92.61
CAMeLBERT_MSA	98.12	89.63	93.65	94.30	94.97	93.65
CAMeLBERT_CA	**98.98**	**94.27**	**95.22**	**96.85**	**98.64**	**95.22**
mBERT_base	97.25	83.81	88.22	91.16	94.73	88.22
XLM_RoBERTa_base	96.72	80.45	86.10	89.33	93.41	86.10
AraBERTv2_large	97.74	86.90	90.32	92.83	95.77	90.32
ArabicBERT_large	98.61	92.01	93.37	95.62	98.20	93.37

TABLE 7. Classification results on Test set with BERT models.

Model	Acc.	$F1_{MH}$	AUC	F1	Prec.	Rec.
AraBERTv2_base	97.40	85.53	91.05	92.05	93.12	91.05
ArabicBERT_base	97.59	86.55	91.52	92.62	93.78	91.52
ARBERT	98.38	90.99	93.97	95.05	96.19	93.97
QARIB	98.01	88.79	92.48	93.85	95.33	92.48
CAMeLBERT_MSA	98.01	89.21	**94.13**	94.05	93.97	**94.13**
CAMeLBERT_CA	**98.68**	**92.47**	93.77	**95.88**	**98.25**	93.77
mBERT_base	96.73	80.62	86.46	89.42	93.06	86.46
XLM_RoBERTa_base	96.80	80.81	86.14	89.53	93.86	86.14
AraBERTv2_large	97.82	87.45	91.84	93.13	95.66	90.91
ArabicBERT_large	98.31	90.49	91.84	94.78	96.29	93.38

TABLE 8. CAMeLBERT_CA results on Balanced and Unbalanced Test sets.

Train type	Test type	Acc.	$F1_{MH}$	AUC	F1	Prec.	Rec.
Balanced	Balanced	95.31	95.14	95.31	95.30	95.53	95.31
Balanced	Unbalanced	97.21	85.88	94.80	92.17	89.90	94.80
Unbalanced	Balanced	94.49	94.17	94.49	94.47	95.04	94.49

D. QUALITATIVE ANALYSIS

The confusion matrix on Fig. 5 shows that only 5 SHs were predicted as Mawdu, and 30 Hadiths were falsely predicted as Sahih to make a total of 35 wrong predictions out of all 2,657 Test Hadiths.

To deepen the evaluation of our MH detection system, we created a new dataset of simulated MHs. The texts of those simulated MHs come from Muslim scholars religious opinions (Fatwa),[11] which are texts dealing with religious topics, containing semantics very similar to Hadiths, and quoting prophetic personalities. Considering that these three

[10] https://github.com/kamelgaanoun/mhdetection/tree/main/corpus

[11] We scrapped 45 Fatwas from https://binbaz.org.sa

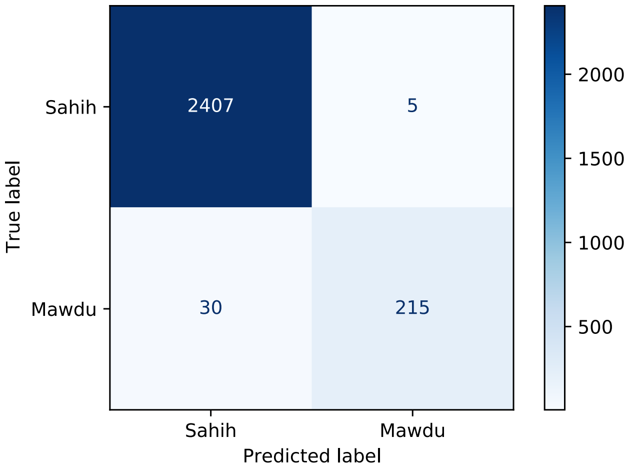


FIGURE 5. Confusion matrix for Test set results.

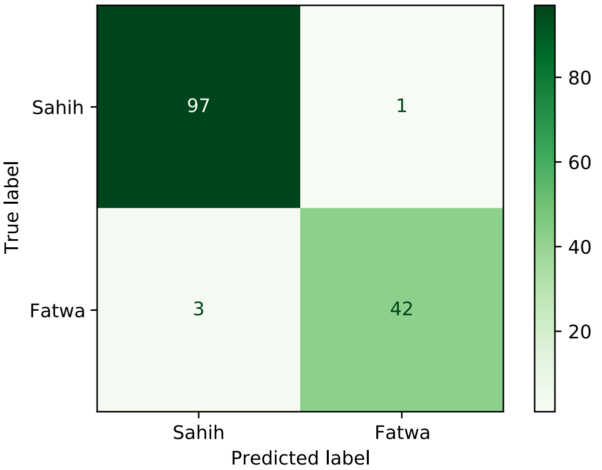


FIGURE 6. Confusion matrix for simulated MHs (Fatwa).

elements are omnipresent in the MHs, the integration of these texts may be seen as a simulation of MHs. Here, we create a dataset containing those Fatwas along with a set of SHs, and evaluate our system against this dataset to see if it can differentiate the simulated MHs from Hadith.

As for the sources, we used the 380 LK corpus Hadiths that do not have authenticity degrees (see subsection III-B). To obtain the degrees of authenticity, we follow the same method as in subsection III-B2. We retain 98 Hadiths that were identified as Sahih. As such, we are guaranteed that these Hadiths have not been seen by the model during training.

The model, as shown in Fig. 6, classifies almost all the examples correctly, except for one SH predicted as Mawdu and three Fatwas predicted as SHs. Based on this evaluation, $F1_{Fatwa}$ is 95.45% and Accuracy is 97.20%, which are very close to the scores obtained in the previous evaluations. This clearly further shows the ability of the model to detect MHs.

VI. DISCUSSION

BERT models considerably outperform ML models regardless of any configuration settings. In fact, the best BERT models, CAMeLBERT_CA, outperforms the best ML model, RF, by more than 42 percentage points in terms of $F1_{MH}$ score. The same is true for the lowest BERT score achieved by multilingual BERT (80.62%), which is 31 points higher than the ML highest score. Furthermore, the BERT models outperform the best ML model by an average of approximately 38 points. This finding justifies the proposal in this paper and is consistent with the literature conclusions about the superiority of BERT models in text classification [72], [73], [74].

Regarding ML models, the second LogTF-IDF/SVD configuration without light stemming produce the best results on the Dev set, with an $F1_{MH}$ score of 51.93% for the RF model and an average of 33% for all models. RF proves to be the best model for this problem, scoring highest in 5 of the 8 setups studied. LGBM comes second with an average of 47.99% against 48.56% for RF on all experimented configurations. The evaluation of the best configuration on the Test set gives the advantage to RF with an $F1_{MH}$ of 49.60%, outperforming LGBM, which achieves a score of 47.34%.

For the Dev set, BERT models achieve their highest scores with three epochs and without light stemming, with CAMeLBERT_CA attaining an $F1_{MH}$ of 94.27%, barely better than ArabicBERT_large, which achieves a score of 92.01%. When applying this configuration to the Test set, CAMeLBERT_CA gets an $F1_{MH}$ score of 92.47%, followed by ArabicBERT_large and ARBERT, with 92.01% and 90.99% $F1_{MH}$ scores, respectively.

Whether it is for the Dev set or the Test set, CAMeLBERT_CA achieves important results, and outperforms models trained on larger datasets. For example, CAMeLBERT_CA obtains better results than models trained on datasets 18 times larger such as CAMeLBERT_MSA and 16 times larger like ArabicBERT, and also AraBERT and ARBERT trained on 13 and 10 times larger datasets, respectively. This performance is attributable to the model's focus on the task at hand, as it is trained on classical Arabic texts unlike the other models. The other models were trained on MSA text, whereas Hadiths are mainly written in classical Arabic. Using a smaller model specialized on the treated text is more relevant than larger models trained on less specialized texts. Moreover, CAMeLBERT_CA also outperforms models with more complex architecture, like AraBERTv2_large and ArabicBERT_large based on the Large architecture BERT version, whereas CAMeLBERT is based on BERT base model.

Even though the dataset is unbalanced, the method achieves very high scores, including for the F1 metric, which is sensitive to this type of dataset. This result is important since the used dataset mimics the actual preponderance of MHs. By taking a balanced training and unbalanced evaluation datasets as a baseline, our results exceed this score by over 6 percentage points. The other two scenarios with a balanced evaluation dataset are considered as maximum comparison scores to encourage future research.

Furthermore, these results address a very present need in this field, which is either the absence of works addressing unbalanced datasets, or the use of metrics not appropriate for unbalanced datasets, like accuracy metric.

VII. CONCLUSION AND FUTURE DIRECTIONS

Hadiths are the second source of Islamic law after the Quran. However, some Hadiths may be fabricated and mislead the faithful. These Hadiths are called Mawdu Hadiths (MHs). In this work, we have developed a system for detecting fabricated Hadiths. For this purpose, we have created and released the first dataset specific to MHs, called MAHADDAT along with releasing a new enhanced version of an existed dataset (NAH), called NAH Plus. The work presented in this paper also study and understand the central text and content of Hadith, Matn, rather than solely focusing on the Sanad. Despite being trained in much smaller dataset as compared to other Arabic BERT models, our best system is based on CAMeLBERT_CA, a BERT-based model specializing in the classical Arabic variant. The proposed model achieves state-of-the-art results with an $F1_{MH}$ score of 92.47%. Moreover, a thorough comparison study in Hadith authentication between numerous classical ML algorithms and all available Arabic TLMs was also performed. Such comparison reveal that all Arabic TLMs are superior to all classical ML models. Future studies could refine the automatic authentication of Hadiths by improving the method used in this paper. Fine-grained authentication would be possible by detecting both the degree and type of Hadith. The Daif Hadith, for example, can be classified into up to ten types. In addition, the findings about the superiority of a BERT model specializing in CA as demonstrated by CAMeLBERT_CA could be explored by developing more specialized models trained on Hadith corpora, in order to create more effective Hadith analysis systems.

ACKNOWLEDGMENT

The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

REFERENCES

[1] L. Bozarth and C. Budak, "Toward a better performance evaluation framework for fake news classification," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, 2020, pp. 60–71.

[2] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 448–456.

[3] F. Al-Hawari and H. Barham, "A machine learning based help desk system for IT service management," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 6, pp. 702–718, Jul. 2021.

[4] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *Proc. 1st Workshop Social Media Analytics (SOMA)*, 2010, pp. 80–88.

[5] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102121.

[6] N. Boudad, R. Faizi, R. O. H. Thami, and R. Chiheb, "Sentiment analysis in Arabic: A review of the literature," *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 2479–2490, 2017.

[7] H. M. Abdelaal and H. A. Youness, "Hadith classification using machine learning techniques according to its reliability," *Rom. J. Inf. Sci. Technol.*, vol. 22, nos. 3–4, pp. 259–271, 2019.

[8] M. N. Al-Kabi, H. A. Wahsheh, and I. M. Alsmadi, "A topical classification of Hadith Arabic text," in *Proc. IMAN*, 2014, pp. 1–8.

[9] M. M. A. Najeeb, "A novel Hadith processing approach based on genetic algorithms," *IEEE Access*, vol. 8, pp. 20233–20244, 2020.

[10] M. A. Saloot, N. Idris, R. Mahmud, S. Jaafar, D. Thorleuchter, and A. Gani, "Hadith data mining and classification: A comparative analysis," *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 113–128, Jun. 2016.

[11] F. Binbeshr, A. Kamsin, and M. Mohammed, "A systematic review on Hadith authentication and classification methods," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 2, pp. 1–17, Apr. 2021.

[12] A. H. Usman and R. Wazir, "The fabricated Hadith: Islamic ethics and guidelines of Hadith dispersion in social media," *Turkish Online J. Design Art Commun.*, vol. 8, pp. 804–808, Sep. 2018.

[13] N. K. Ibrahim, S. Samsuri, M. Sadry Abu Seman, A. E. B. Ali, and M. Kartiwi, "Frameworks for a computational isnad authentication and mechanism development," in *Proc. 6th Int. Conf. Inf. Commun. Technol. Muslim World (ICT4M)*, Nov. 2016, pp. 154–159.

[14] J. A. C Brown, *Hadith: Muhammad's Legacy Medieval Modern World*. New York, NY, USA: Simon and Schuster, 2017.

[15] S. Hakak, A. Kamsin, W. Z. Khan, A. Zakari, M. Imran, K. B. Ahmad, and G. A. Gilkar, "Digital Hadith authentication: Recent advances, open challenges, and future directions," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 6, Jun. 2022, Art. no. e3977.

[16] T. Tarmom, E. Atwell, and M. Alsalka, "Deep learning vs compression-based vs traditional machine learning classifiers to detect Hadith authenticity," in *Proc. Annu. Int. Conf. Inf. Manage. Big Data*. Cham, Switzerland: Springer, 2022, pp. 206–222.

[17] I. Najiyah, S. Susanti, D. Riana, and M. Wahyudi, "Hadith degree classification for Shahih Hadith identification web based," in *Proc. 5th Int. Conf. Cyber IT Service Manage. (CITSM)*, Aug. 2017, pp. 1–6.

[18] H. M. Abdelaal, B. R. Elemary, and H. A. Youness, "Classification of Hadith according to its content based on supervised learning algorithms," *IEEE Access*, vol. 7, pp. 152379–152387, 2019.

[19] M. Ghanem, A. Mouloudi, and M. Mourchid, "Classification of Hadiths using LVQ based on VSM considering words order," *Int. J. Comput. Appl.*, vol. 148, no. 4, pp. 25–28, Aug. 2016.

[20] A. Hassaine, Z. Safi, and A. Jaoua, "Authenticity detection as a binary text categorization problem: Application to Hadith authentication," in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2016, pp. 1–7.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[22] T. B. Brown, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[23] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–18.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[25] J. Á. González, L.-F. Hurtado, and F. Pla, "Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter," *Inf. Process. Manage.*, vol. 57, no. 4, Jul. 2020, Art. no. 102262.

[26] D. Meškelė and F. Frasincar, "ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102211.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[29] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*.

[30] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media," in *Proc. 14th Workshop Semantic Eval.*, Dec. 2020, pp. 2054–2059.

[31] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training BERT on Arabic tweets: Practical considerations," 2021, *arXiv:2102.10684*.

[32] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," 2021, *arXiv:2103.06678*.

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[34] K. M. A. Jbara, A. T. Sleit, and B. H. Hammo, *Knowledge Discovery in Al-Hadith Using Text Classification Algorithm*. Amman, Jordan: Univ. Jordan, 2009.

[35] M. Alkhatib, "Classification of Al-Hadith Al-Shareef using data mining algorithm," in *Proc. Eur., Medit. Middle Eastern Conf. Inf. Syst., (EMCIS)*, Abu Dhabi, UAE, 2010, pp. 1–23.

[36] N. M. Al-Kabi, A. H. Wahsheh, M. I. Alsmadi, and A. M. A. Al-Akhras, "Extended topical classification of Hadith Arabic text," *Int. J. Islamic Appl. Comput. Sci. Technol.*, vol. 3, no. 3, pp. 13–23, Sep. 2015.

[37] M. F. Afianto and S. Al-Faraby, "Text categorization on Hadith Sahih Al-Bukhari using random forest," *J. Phys., Conf. Ser.*, vol. 971, Mar. 2018, Art. no. 012037.

[38] M. Y. A. Bakar and S. A. Faraby, "Multi-label topic classification of Hadith of Bukhari (Indonesian language Translation)using information gain and backpropagation neural network," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 344–350.

[39] N. A. P. Rostam and N. H. A. H. Malim, "Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 6, pp. 658–667, Jul. 2021.

[40] Gugun Mediamer, Adiwijaya, and Said Al Faraby, "Development of rule-based feature extraction in multi-label text classification," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 4, pp. 1460–1465, 2019.

[41] M. Ghazizadeh, M. H. Zahedi, M. Kahani, and B. M. Bidgoli, "Fuzzy expert system in determining Hadith validity," in *Advances in Computer and Information Sciences and Engineering*, 2008, pp. 354–359.

[42] K. A. Aldhlan, A. M. Zeki, A. M. Zeki, and H. A. Alreshidi, "Novel mechanism to improve Hadith classifier performance," in *Proc. Int. Conf. Adv. Comput. Sci. Appl. Technol. (ACSAT)*, Nov. 2012, pp. 512–517.

[43] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Know. Inf. Sys.*, vol. 14, pp. 1–37, Dec. 2008.

[44] M. Q. Shatnawi, Q. Q. Abuein, and O. Darwish, "Verification Hadith correctness in Islamic web pages using information retrieval techniques," in *Proc. Int. Conf. Inf. Commun. Syst.*, 2011, pp. 164–167.

[45] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[46] A. M. Azmi and A. M. AlOfaidly, "A novel method to automatically pass hukm on Hadith," in *Proc. 5th Int. Conf. Arabic Lang. Process. (CITALA)*, 2014, pp. 118–124.

[47] S. S. Balgasem and L. Q. Zakaria, "A hybrid method of rule-based approach and statistical measures for recognizing narrators name in Hadith," in *Proc. 6th Int. Conf. Electr. Eng. Informat. (ICEEI)*, Nov. 2017, pp. 1–5.

[48] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[49] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, vol. 3, no. 22, 2001, pp. 41–46.

[50] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[51] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Oct. 2014.

[52] E. T. Luthfi, Z. I. M. Yusoh, and B. M. Aboobaider, "BERT based named entity recognition for automated Hadith narrator identification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, pp. 1–9, 2022.

[53] T. Tarmom, E. Atwell, and M. Alsalka, "Non-authentic Hadith corpus: Design and methodology," in *Proc. IMAN*. Leeds, U.K., 2019, pp. 13–19.

[54] S. Altammami, E. Atwell, and A. Alsalka, "The Arabic-english parallel corpus of authentic Hadith," *Int. J. Islamic Appl. Comput. Sci. Technol.*, vol. 8, no. 2, 2020, pp. 1–10.

[55] A. Mohamed and M. A. Jamaoui. *Hadith-Data-Sets*. Accessed: Jun. 26, 2022. [Online]. Available: https://github.com/abdelrahmaan/Hadith-Data-Sets

[56] K. Grzegorczyk, "Vector representations of text data in deep learning," 2019, *arXiv:1901.01695*.

[57] A. I. Kadhim, Y.-N. Cheah, I. A. Hieder, and R. A. Ali, "Improving TF-IDF with singular value decomposition (SVD) for feature extraction on Twitter," in *Proc. 3rd Int. Eng. Conf. Develop. Civil Comput. Eng. Appl.*, 2017, pp. 1–9.

[58] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Rev.*, vol. 35, no. 4, pp. 551–566, 1993.

[59] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[60] J. S. Cramer, "The origins of logistic regression," Tinbergen Inst. Discuss. Paper 02-119/4, 2002.

[61] E. S. Pearson, "Bayes' theorem, examined in the light of experimental sampling," *Biometrika*, vol. 17, nos. 3–4, pp. 388–442, Dec. 1925.

[62] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.

[63] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[64] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3149–3157.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st NIPS*, Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[66] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "CCNet: Extracting high quality monolingual datasets from web crawl data," 2019, *arXiv:1911.00359*.

[67] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2021.

[68] L. Nigst, M. Romanov, S. B. Savant, M. Seydi, and P. Verkinderen, "OpenITI: A machine-readable corpus of islamicate texts," Oct. 2020, doi: 10.5281/zenodo.4075046.

[69] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Global Ecol. Biogeogr.*, vol. 17, no. 2, pp. 145–151, Mar. 2008.

[70] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.

[71] M. A. H. Omer and S.-L. Ma, "Stemming algorithm to classify Arabic documents," *J. Commun. Comput.*, vol. 7, no. 9, 2010.

[72] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," 2020, *arXiv:2005.13012*.

[73] D. Patel, P. Raval, R. Parikh, and Y. Shastri, "Comparative study of machine learning models and BERT on SQuAD," 2020, *arXiv:2005.11313*.

[74] Q. G. To, K. G. To, V.-A.-N. Huynh, N. T. Q. Nguyen, D. T. N. Ngo, S. J. Alley, A. N. Q. Tran, A. N. P. Tran, N. T. T. Pham, T. X. Bui, and C. Vandelanotte, "Applying machine learning to identify anti-vaccination tweets during the COVID-19 pandemic," *Int. J. Environ. Res. Public Health*, vol. 18, no. 8, p. 4069, Apr. 2021.

KAMEL GAANOUN received the M.Sc. degree in data science and statistical modeling from Université Bretagne Sud, France. He is currently a Data Scientist and a member of the Association of Business Intelligence (AMID). His current research interests include artificial intelligence, Arabic natural language processing, transformers models, and data-centric approaches.

MOHAMMED ALSUHAIBANI received the M.Sc. degree in advance computer science and the Ph.D. degree in computer science from The University of Liverpool, U.K. He is currently an Assistant Professor and the Head of the Department of Computer Science, College of Computer, Qassim University, Saudi Arabia. His research interests include artificial intelligence, machine learning, computational linguistics, and natural language processing fields.

• • •