The Scientific Journal of King Faisal University

Basic and Applied Sciences

Detecting Hadith Authenticity Using a Deep-learning Approach

Eshrag Ali Refaee

Department of Information Technology and Security, College of Computer Sciences and Information Technology, Jazan University, Jazan, Saudi Arabia

LINK	RECEIVED	ACCEPTED	PUBLISHED ONLINE	ASSIGNED TO AN ISSUE
https://doi.org/10.37575/b/sci/210084	19/12/2021	14/04/2022	14/04/2022	01/06/2022
NO. OF WORDS	**NO. OF PAGES**	**YEAR**	**VOLUME**	**ISSUE**
4860	5	2022	23	1

ABSTRACT

Hadith is a collection of texts containing sayings of the prophet Muhammad, which, along with accounts of his daily practice, constitute the second major source of legislation for Muslims after the Holy Koran. The Hadith collection comprises thousands of text pieces transferred over the years by many narrators with varying degrees of credibility. Hadith scholars are faced with the challenge of assessing the degree of a specific Hadith's authenticity to classify the Hadith as Sahih (fully authentic and accepted) or Daif (rejected). Automatic Hadith classification has been addressed in the literature; however, the results vary and are not directly comparable, as no dataset has been made available for benchmarking. In addition, no previous work has utilised deep-learning (DL) approaches for Hadith classification. This work contributes by 1) collecting and publicly releasing a benchmark Hadith dataset of almost 4,000 Hadith texts to facilitate future research, 2) exploring DL model performance on binary Hadith classification tasks, and 3) benchmarking traditional machine learning against DL models. Our best results were recorded with an ARBERT DL model that provided an accuracy score of 91.56%.

KEYWORDS

Hadith classification; deep learning; Classical Arabic; machine learning; Hadith science; Hadith authenticity

CITATION

Refaee, E.A. (2022). Detecting Hadith authenticity using a deep-learning approach. *The Scientific Journal of King Faisal University: Basic and Applied Sciences*, **23**(1), 80–4. DOI: 10.37575/b/sci/210084

1. Introduction

According to the United Nations Educational, Scientific and Cultural Organisation UNESCO (2021), Arabic is the language of more than 400 million people around the world. Arabic can be divided into three major classes: Classical Arabic (CA), Modern Standard Arabic, and Dialectal Arabic (Habash, 2010). Hadith is one of the most well-known CA texts. The literal meaning of the word Hadith in Arabic is "anything spoken or told among people" (Al Ma'ni Online Dictionary, 2021). The Oxford Dictionary definition of Hadith is "a collection of traditions containing sayings of the prophet Muhammad which, with accounts of his daily practice, constitute the major source of guidance for Muslims apart from the Koran". Hadith science is a branch of the Islamic sciences concerned with studying the sayings and actions of the prophet Mohammad (Al Ma'ni Dictionary, 2021). It is crucial since Hadith is the second primary source of legislation, namely a constitution, after the Holy Koran for almost 1.6 billion Muslims worldwide (Desilver and Masci, 2017). The Hadith corpus is vast and has been recognised in the universal European Language Resources Association catalogue (European Language Resources Association, 2021). Each Hadith consists of two main parts: a Matn, which is the body text of the Hadith, and an Isnad, which is the chain of narrators who have transmitted the referenced Hadith from the days of the prophet until the day that the Hadith was documented in one of the significant Hadith books (Duderrija, 2021). The most well-known Hadith books are *Imam Al Bukhari*, *Imam Muslim* and *Ibn Majah*. During Hadith collection and before any Hadith is written in one of the crucial Hadith books, Hadith collectors perform an extensive verification process, checking the degree of Hadith authenticity to avoid documenting any Hadith that is not genuine. Due to the Hadith's vital role as a leading source of legislation, the Hadith collectors – *Imam Al Bukhari*, *Imam Muslim* and *Ibn Majah*, among others – take their responsibility for verifying Hadith authenticity seriously. For instance, they check whether all the chains of narrators of a particular Hadith are reliable, that is, they verify the absence of a reason to doubt their credibility. They also identify any gap between narrators, for example, if the lives of two consecutive narrators overlap and whether they have met in their lifetime (Azmi *et al.*, 2019). Based on its level of authenticity, the Hadith corpus can be classified into three major categories: Sahih, which is fully verified to be genuine; Hasan, which is highly likely to be genuine and usually accepted by most scholars but does not qualify to be Sahih because of a minor issue; and Daif, or weak Hadith. The latter category refers to Hadith that does not have the qualifications of the Sahih or Hasan Hadiths. Weak Hadith is not used as evidence since at least one of its narrators has committed a transgression or been accused of lying or another act that negatively affects the narrator's credibility.

In this work, the focus is on the automatic classification of Hadith based on its level of authenticity. Specifically, a deep-leaning (DL) approach is employed to perform automatic classification to determine whether a specific Hadith text is Sahih, namely genuine, or Daif. To do this, a publicly available dataset of the Hadith corpus is utilised. To the best of our knowledge, no previous work has focused on the use of DL models for automatic detection and classification of Hadith categories based on the Hadith's level of authenticity. This work also contributes to the literature by making the collected and pre-processed dataset publicly available for further research. To the best of our knowledge, there is no publicly available benchmark Hadith dataset that has been prepared for this task.

The remainder of this paper is structured as follows. A review of previous natural language processing (NLP) research that focuses on Hadith science is provided in section 2. The subsequent section outlines the dataset used in this work and the experimental setup. This is followed by the methodology section, which describes our methodology in detail, explains the experimental results, and discusses error analysis. The final section summarises the main findings and highlights potential directions for future research.

2. Related Work

Computational NLP research has approached Hadith science in several ways, including automatic Hadith topic classification, automatic Hadith question answering systems, and a graphing narration tree designed to ascertain how the Hadith text spread (Azmi

Corresponding Author: Eshrag Ali Refaee 00966502240692, erefaie@jazanu.edu.sa

et al., 2019). In this work, the investigation is focused on exploring the use of a machine learning-based approach that utilises DL models to classify a specific Hadith as Sahih or Daif automatically, a task typically completed manually by Hadith scholars. To perform such classifications, the Hadith scholars must study several circumstantial factors, such as narrators' biographies and their levels of reliability and credibility.

Ghazizadeh *et al.* (2008) report their work on a dataset of 999 Hadiths collected from three books: *Sahih Al Bukhari*, *Jamiu Al-Termithi*, and *Silsilat Al-Ahadith Al-Daeifah Wal Al-Mawdhuah*. The authors employed a fuzzy system that utilised a set of rules extracted from their dataset. The system was implemented using expert system software. Subsequently, the collected Hadith sample was inserted into the database to be assessed utilising documentary information. The best performance achieved by this system included an accuracy rate of 94%. Similar work was conducted by Najeeb (2021); however, the author did not provide the results.

Aldhaln *et al.* (2012) used the decision tree (DT) algorithm C4.5 to extract classification rules. This step was followed by applying Naïve Bayes (NB) with the goal of improving overall performance. The authors reported that the best accuracy rate achieved by their system was 97%, with a similar finding reported by Aldhaln *et al.* (2012b).

Fadele *et al.* (2021) proposed a taxonomy of four levels. Their approach included a document classification phase that used features extracted in the form of keywords to train the model. In addition, a keyword list was created at each stage of hyperactive rectangle tree formation. Thus, the keyword lists that correlated with the document categories were combined and fed into the classifier for document category prediction. Finally, a classification technique was applied to each document to determine the authenticity of a specific Hadith.

Table 1 summarises previous studies that investigated various methods of automatic Hadith authenticity detection. A comprehensive review of Hadith-related NLP tasks is presented in Azmi *et al.*'s (2019) paper.

Table 1: Summary of previous Hadith classification studies

Paper	Dataset	Performance	Approach
Ghazizadeh *et al.* (2008)	KAFI dataset	94% accuracy	Fuzzy system, extracted rules
Aldhaln *et al.* (2012a)	999 Hadiths	Recall: 97%, accuracy: 97%	DT C4.5 for rule extraction and NB
Aldhaln *et al.* (2012b)	999 Hadiths	Recall: 97%, accuracy: 97%	DT C4.5
Najeeb (2021)	Not reported	Not reported	Association rule mining
Najiyah *et al.* (2017)	Sample of 346 Hadiths	False negative = 0%, False positive = 0.097%	Expert system and DT algorithm
Abelaal *et al.* (2019a)	Hadith collection	93.69% accuracy	DT, linear support vector classifier (linear SVC), Stochastic Gradient Descent (SGD) classifier
Abelaal *et al.* (2019b)	Hadith collection	93.75% accuracy	NB
Fadele *et al.* (2021)	Not reported	Not reported	Taxonomy, keyword extraction, classifier

3. Experimental Framework

This section describes the experimental dataset, architecture and parameters used in this study and reviews the parameter tuning performed. Figure 1 shows our workflow pipeline and indicates the significant data pre-processing steps and the DL algorithms employed.

3.1. Data Collection:

This study utilised a publicly available Hadith dataset. It is worth mentioning that the datasets used in previous work were not available for result benchmarking. As such, random Hadith samples were collected from *Sahih Al Bukhari* and *Daif Ibn Majah* to represent our Sahih vs Daif binary classes.

3.1.1. Example of Hadith Sahih

> *Hadathana alhumaydiu eabd allah bn alzubayr qal hadathana sufyan qal hadathana yahyaa bn saeid al'ansariu qal 'akhbarani muhamad bn 'iibrahim altaymiu 'anah samie ealqamat bn waqaas allaythia yaqul samiet eumar bn alkhataab radaa allah eanh ealaa alminbar qal samiet rasul allah salaa allah ealayh wasalam yaqul " 'iinama al'aemal bialniyaat wa'iinama likuli amri ma nawaa faman kanat hijratuh 'iilaa dunya yusibuha 'aw 'iilaa amra'at yankihuha fahijratuh 'iilaa ma hajar 'iilayh".* (Source: Al Bukhari)

3.1.2. Example of Hadith Daif

> *Hadathana suid bin saeid hadathana aibn 'abi alrijal ean eabd alrahman bin eamrw al'awzaei ean eabdat bin 'abi libabat ean eabd allah bin eamriw bin aleas qal samieat rasul allah salaa allah ealayh wasalam yaquli: "lam yazal 'amr bni 'iisrayiyl muetadilaan hataa nasha fihim aalmualadun 'abna'an sabaya aal'umam faqaluu bialraay fadaluu wa'adalwa".* (Source: 4760, Daif ibn Majh)

The dataset was split as follows: 70% training, 15% validation, and 15% test. Statistics that relate to the dataset collected are provided in Table 2. As one of the contributions of this work, the dataset used in this set of experiments, with the same training-validation-test distribution, is being released to allow replication and benchmarking of our results. The dataset distribution shows that Sahih Hadith instances are the majority class and constitute almost 76% of the dataset. This is common, as data are imbalanced or skewed towards a majority class in many real-world applications (Avati *et al.*, 2018). The preliminary experiments found no significant impact of class imbalance on the overall performance. As such, no class balancing mechanisms were employed during the experiments described in this paper.

Table 2: Dataset description

Dataset	No. of instances	No. of Sahih instances	No. of Daif instances	Total
Training	2761	2100	661	5522
Validation	592	450	142	1184
Test	591	450	141	1182
Total	3944	3000	944	

3.2. Data Pre-processing:

The obtained data were pre-processed using computationally motivated steps designed to reduce the feature space:

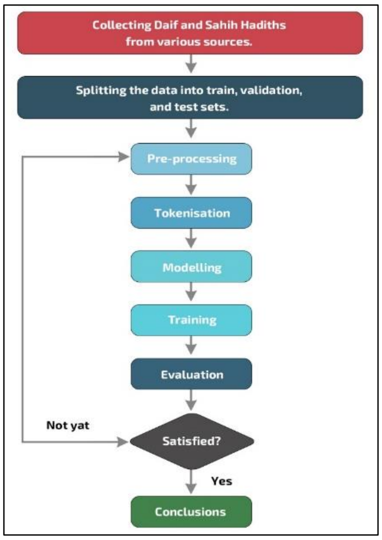
- Normalising exchangeable Arabic letters: Mapping letters with various forms – Alef, Hamza and Yaa – to their representative characters. The data was pre-processed using the AraBERT-base pre-processor (Antoun *et al.*, 2020).
- Text segmentation: This step was performed to separate tokens based on spaces and punctuation marks. The dedicated tokenisers were used for BERT-base, ARBERT and ARABER. For classical machine learning (ML) models, the tokeniser provided by the PyArabic [1] package was employed (Zerrouki, 2010).
- Removing special characters and punctuation that separate Matn from Isnad.
- Text stemming: This is a further text pre-processing step that aims to alleviate the high dimensionality of the text data by using reduced forms of words such as stems. Previous work has shown the importance of employing such a technique (Refaee, 2017). The problem of high dimensionality becomes very pronounced when dealing with a morphologically rich and highly derivative language like Arabic. Refaee (2017) highlights the significance of this text pre-processing step and argues that Arabic text classification tasks can be problematic when the compressed forms of words are not used, as trained models can be exposed to many previously unseen features (words) that might actually be present in training and testing but in different forms. This study employed the ARLSTem2 [2] Arabic light stemmer (Abainia and Rebbani, 2019).

The following step in our workflow pipeline involved extracting numerical features from tokens using the term frequency-inverse document frequency with n-grams from one to four (see Figure 1). Each instance is represented as a 50,000-dimensional vector.

[1] PyArabic is a publicly available Python library explicitly designed for the Arabic language. Available at: https://pypi.org/project/PyArabic/ and accessed on 20/10/2021

[2] The implementation of the ARLSTem2 Arabic stemmer is available at: https://www.nltk.org/_modules/nltk/stem/arlstem2.html

Figure 1: Workflow pipeline



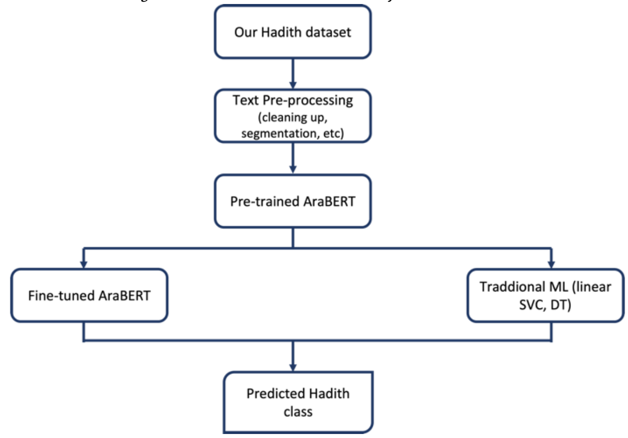
3.3. ML Models:

This section describes the ML models used in this work. To the best of our knowledge, no previous work has utilised DL models to perform automatic classification of Hadith texts based on their degree of authenticity, namely Sahih or Daif. As such, several state-of-the-art Arabic DL models were employed, specifically, the AraBERT model with a classification layer on top of it (Antoun *et al.*, 2020). In addition, and to have common ground with previous work that used traditional ML models (see Section 2), several traditional models were used. It is argued that having comparative results for traditional vs DL ML models on a benchmark dataset is important in exploring potential Hadith classification performance variation. The datasets used in previous studies were not available for the purpose of conducting comparative investigations; therefore, this paper presents a comprehensive set of experiments that cover traditional vs DL models and clarifies how different categories of ML models perform.

First, experiments were conducted with AraBERT, a common DL model (Antoun *et al.*, 2020). The model was fine-tuned using two stages. The first stage included freezing the AraBERT model and training only the classification layer using the following configuration settings: a maximum sequence length of 128, a batch size of 32, 10 epochs, a learning rate of 0.0001, and dropouts for the attention and feed forward layers of 0.3. In the second stage, the best model was selected – based on its F1 score on the validation set – from the previous stage, and AraBERT was unfreezed using a learning rate of 0.00001. Following El-Alami *et al.* (2021), AraBERT as a pre-trained model was used and fine-tuned for Arabic Hadith text classification. The text representation was computed using the contextual pre-trained AraBERT to introduce Arabic text into a fixed-length vector. This was accomplished by connecting the AraBERT output to an additional layer involving the Softmax classifier to predict the Hadith class. First, each text was tokenised to N tokens, and an input representation was generated for each token constructed by adding the vector embeddings corresponding to the token, the segment, and the token position. Then, the generated vectors were fed into the AraBERT model.

The second set of experiments were conducted using several traditional ML models, specifically, LinearSVC with a Radial Basis Function (RBF) kernel and a DT. This second set of experiments sought to compare the results from DL models – used for the first time in this work – with those from a few traditional ML models – used in previous work but without data provided to enable reproduction of results. Figure 2 shows the overall system architecture of our Hadith classification approach.

Figure 2: The Arabic Hadith classification system architecture



3.4. Evaluation Metrics:

The following evaluation measures were used to assess ML model performance:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (2)$$

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

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

True positive (TP) and true negative (TN) denote numbers of positive and negative examples that are classified correctly, while false negative (FN) and false positive (FP) represent the number of misclassified positive and negative examples, respectively.

In classification problems, the overall performance is measured by identifying the success rate, which is the proportion of correctly classified instances divided by the entire set of instances. The results are reported using two metrics: the macro-F-score and accuracy. The F-score is defined as the harmonic average of precision and recall; accuracy is one of the most widely reported metrics in the literature.

A simple problem formulation was used, treating Hadith classification as a single-level or flat binary classification. Previous studies wherein the major task was to determine whether a given Hadith is genuine (i.e. Sahih) treated the Hadith classification problem as a binary classification. Alternatively, the classification can be performed as a two-level, multi-way classification. In the latter method, the Hadith is first classified as genuine or not. Whenever a Hadith is classified as genuine, its text is then classified based on its degree of authenticity, such as Hasan. This work experimented with the first setting, namely using binary classification, as it seems that the major concern of Hadith scholars is to distinguish Sahih vs Daif. Since Hadith is a major source of legislation for millions of Muslims globally, the ability to distinguish Sahih Hadith vs Daif is a vital task for Hadith scholars (Azmi *et al.*, 2019). As such, this work focuses on the task of binary classification of Hadith into Sahih and Daif (Baru *et al.*, 2017).

4. Results

This section presents the results of the experimental investigations conducted using our Hadith dataset. This work studied the binary classification of Hadith text into Sahih vs Daif using DL models. Various experiments were performed to investigate several DL models. Table 3 presents the main findings of the first set of experiments. The ARBERT model achieves the best performance, with

91.56% accuracy and an F-score of 88.64%, using embeddings and two encoder layers. Figure 2 shows the performance variation among various DL models and different settings within the same model. To the best of our knowledge, no previous work has utilised DL models for Hadith classification; therefore, the current results are considered a benchmark for future work, especially since our dataset is being released as a part of this publication.

Most previous work has focused on using the entire Hadith, namely Matn + Isnad (see Introduction). Therefore, the impact of isolating Matn from Isnad on the overall performance of our top-performing model was studied to determine whether Isnad has the positive impact of being informative for the model.

Example of Hadith Sahih with Matn + Isnad

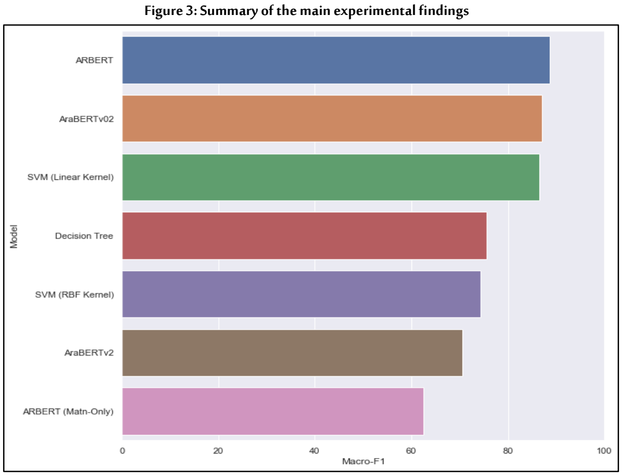
> *Hadathana alhumaydiu eabd allah bn alzubayr qal hadathana sufyan qal hadathana yahyaa bn saeid al'ansariu qal 'akhbarani muhamad bn 'iibrahim altaymiu 'anah samie ealqamat bn waqaas allaythia yaqul samiet eumar bn alkhataab radaa allah eanh ealaa alminbar qal samiet rasul allah salaa allah ealayh wasalam yaqul " 'iinama al'aemal bialniyaat wa'iinama likuli amri ma nawaa faman kanat hijratuh 'iilaa dunya yusibuha 'aw 'iilaa amra'at yankihuha fahijratuh 'iilaa ma hajar 'iilayh ".* (Source: Al Bukhari)

Example of Hadith Sahih with Matn only

> *Yaqul " 'iinama al'aemal bialniyaat wa'iinama likuli amri ma nawaa faman kanat hijratuh 'iilaa dunya yusibuha 'aw 'iilaa amra'at yankihuha fahijratuh 'iilaa ma hajar 'iilayh ".* (Source: Al Bukhari)

The last row of Table 3 indicates a significant drop in performance as a result of eliminating Isnad and experimenting with only Matn. This finding is in line with the theories formulated by Hadith scholars that highlight the vital role of Isnad in detecting Hadith authenticity (Azmi *et al.*, 2019). This suggests that better performance is expected if an additional set of features that show whether a narrator has been discredited by Hadith scholars can be linked and utilised.

Figure 3: Summary of the main experimental findings



In the last set of experiments, to place our findings in a meaningful context alongside previous works, this work experimented with a number of traditional ML algorithms, namely LinearSVC, a support vector machine (SVM) with an RBF kernel function, and a DT. The maximum depth for the DT was set to 20 to avoid overfitting. For the LinearSVC and the SVM with an RBF kernel, the random seed was 42. A simple majority baseline algorithm for majority class prediction was used. The algorithm achieved an accuracy score of 76.14% and an F-score of 43.23%. Table 4 shows that the best performance is achieved by the LinearSVC, which achieves an accuracy score of 90.69% and an F-score of 86.55%, as shown in Figures 2 and 3. It is interesting to note that previous work has also reported that the SVC is among the best-performing models on the same Hadith classification task (see Table 1). Overall, DL models achieve better performance than traditional ML models, as shown in Figure 3, which is probably due to the advances that DL models, such as ARBERT, achieve by learning from a huge number of word features. This is in line with recent findings that DL can notably outperform traditional ML models on many NLP tasks, such as sentiment analysis (Patwa *et al.*, 2020). Figure 4 shows the precision and recall scores attained by all the models employed and reflects a steady and gradual performance improvement achieved by the DL models, with a slightly varying performance achieved by traditional ML models, especially DT. Overall, DL models show a better tendency for further improvements , for example, by utilising larger datasets or more tuning of the models. Traditional ML models, on the other hand, seem to attain a flat performance, which has been reported in previous text classification tasks with the same traditional ML models (Patwa *et al.*, 2020).

Figure 4: Comparative summary of DL and traditional ML model results

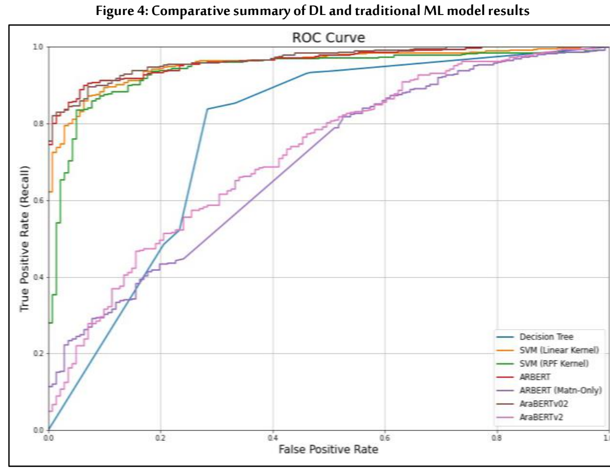


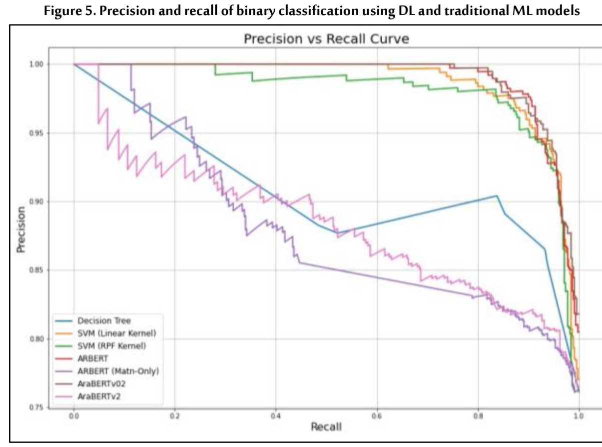
Table 3: Results of binary Hadith classification using DL models

Trial No.	Model	Frozen layers	Dropout	Performance
1 (Isnad + Matn)	AraBERTv2	None	0.3	79.5% Acc. 70.60% F1
2 (Isnad + Matn)	AraBERTv02	Embeddings + 2 encoder layers	0.1	88.32% Acc. 85.12% F1
3 (Isnad + Matn)	AraBERTv02	None	0.1	89.85% Acc. 87.00% F1
4 (Isnad + Matn)	ARBERT	None	0.1	91.03% Acc. 88.34% F1
5 (Isnad + Matn)	ARBERT	Embeddings + 2 encoder layers	0.1	**91.56% Acc.** **88.64% F1**
6 (Isnad + Matn)	ARBERT	Embeddings + 4 encoder layers	0.1	90.86% Acc. 88.00% F1
7 (Isnad + Matn)	ARBERT	Embeddings + 6 encoder layers	0.1	89.17% Acc. 85.70% F1
8 (Matn only)	ARBERT	Embeddings + 2 encoder layers	0.1	75.30% Acc. 62.53% F1

Table 4: Results of binary classification using traditional ML models

Data	Traditional ML Model	Performance
Isnad + Matn	DT	83.76% Acc. 75.51% F1
Isnad + Matn	LinearSVC	90.69% Acc. 86.55% F1
Isnad + Matn	SVC with RBF kernel	84.60% Acc. 74.32% F1

Figure 5. Precision and recall of binary classification using DL and traditional ML models



5. Conclusion and Future Work

This work explored the NLP task of classifying Hadith text as Sahih or Daif based on its authenticity. A dataset of almost 4,000 Hadith samples was collected and pre-processed. It was not possible to obtain datasets from previous work on the same task; therefore, by collecting Hadith samples, this work created a dataset, which is divided into training, validation and test data, that is available as a testbed to facilitate research. Two different sets of experiments were performed: DL models with various settings, and traditional ML models to determine whether to they could compete with the DL models. This work also experimented with two different dataset settings: Isnad + Matn and Isnad only. The best performance was achieved using a DL model (ARBERT) trained on the Isnad + Matn dataset, which achieved an accuracy score of 91.56%.

Future research directions may include experimenting on an English dataset of Hadith text. It would be interesting to determine whether model performance can be improved further by experimenting with English translations of Hadith text. Another direction may involve expanding more features regarding Isnad. Specifically, it is expected that better results can be achieved by using information – features – available from Hadith science regarding the status of the narrator. One of the reasons that Hadith can be treated as Daif is the presence of a narrator who has been chastened as having a weak memory or has been discredited. Such information can be utilised to provide informative features that may help to improve the model's ability to distinguish Sahih from Daif. Another future direction may involve expanding Hadith classification from binary to three-way classification by including the third category of Hadith Hasan. It is anticipated that the dataset collected and cleaned as a part of this work will be helpful for future research by providing a testbed for exploring further tools, methodologies, and approaches.

Biography

Eshrag Ali Refaee

Department of Information Technology and Security, College of Computer Sciences and Information Technology, Jazan University, Jazan, Saudi Arabia

Dr Refaee is a Saudi assistant professor and has a PhD in Computer Sciences (in computational linguistics) from Heriot-Watt University, Edinburgh, UK. She is the head of the Information System department and vice dean of the college. She has published 16 peer-reviewed papers and a book chapter titled Sematic Analysis in the book "Introduction to Computational Linguistics" published by King Abdullah International Centre for the Arabic Language. Her research interests include Arabic NLP, text mining, and ML. She has participated in several international competitions in NLP and achieved first place in SemEval-16, Arabic sentiment analysis track, San Diego, USA. ORCID: 0000-0003-2344-0693.

References

Abainia, K. and Rebbani, H. (2019). Comparing the effectiveness of the improved ARLstem algorithm with existing Arabic light stemmers. In: *Proceedings of the 2019 International Conference on Theoretical and Applicative Aspects of Computer Science* (ICTAACS), (pp. 1–8), Skikda, Algeria, 15–6/12/2019.

Abdelaal, H.M., Ahmed, A.M., Ghribi, W. and Alansary, H.A.Y. (2019a). Knowledge discovery in the Hadith according to the reliability and memory of the reporters using machine learning techniques. *IEEE Access*, **7**(n/a), 157741–55.

Abdelaal, H.M. and Youness, H.A. (2019b). Hadith classification using machine learning techniques according to its reliability. *Science and Technology*, **22**(3), 259–71.

Duderija, A. (2009). Evolution in the canonical Sunni Hadith body of literature and the concept of an authentic Hadith during the formative period of Islamic ought as based on recent Western scholarship. *Arab Law Quarterly*, **23**(4), 389–415.

Al Ma´ani Dictionary. (n/a). In *Al Ma´ani Dictionary*. Available at: https://www.almaany.com/en/ dict/ar-en/hadith/ (Accessed on 10/10/2021).

Aldhaln, K., Zeki, A., Zeki, A. and Alreshidi, H. (2012a). Improving knowledge extraction of Hadith classifier using decision tree algorithm. In: *Proceedings of the 2012 International Conference on Information Retrieval and Knowledge Management*. (pp. 148–152), Kuala Lumpur, Malaysia, 13–5/03/2012.

Aldhlan, K.A., Zeki, A.M., Zeki, A.M. and Alreshidi, H.A. (2012b). Novel mechanism to improve Hadith classifier performance. In: *Proceedings* of the *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*. (pp. 512–517), Kuala Lumpur, Malaysia, 13-15/03/2012.

Antoun, W., Baly, F. and Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In: *Proceedings* of *LREC 2020 Workshop Language Resources and Evaluation Conference*. (p. 9), Marseille, France, 13–5/05/2020.

Avati, A., Jung, K., Harman, S., Downing, L. and Ng, A. (2018). Improving palliative care with deep learning. *BMC Medical Informatics and Cecision Making*, **18**(4), 55–64.

Azmi, A.M., Al-Qabbany, A.O. and Hussain, A. (2019). Computational and natural language processing-based studies of Hadith literature: A survey. *Artificial Intelligence Review*, **52**(2), 1369–414.

Baru, R., Omar, S.H.S., Ag, I.M., Fuad, A.N. and Mohd, M.F. (2017). Consolidation of Ulum Al-hadith to the society. *International Journal of Academic Research in Business and Social Sciences*, **7**(10), 2222–6990.

Desilver, D. and Masci, D. (n/a). *World's Muslim Population*, *Pew Research Centre*. Available at: https: //www.pewresearch.org/fact-tank/2017/01/31/worlds-muslim-population-more-widespread-than-you-might-think/ (Accessed on 12/11/2021).

El-Alami, F.Z., El Alaoui, S.O. and Nahnahi, N.E., (2021). Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. *Journal of King Saud University-Computer and Information Sciences,* In press.

European Languages Resources Association (ELRA). (n.d.). *ELRA Universal Catalogue: ELRA–WC0134*. Available at: http://universal.elra. info/product_info.php?cPath=42_43 and products_id=353/ (Accessed on 03/11/2021).

Fadele, A.A., Kamsin, A., Ahmad, K. and Hamid, H. (2021). A novel classification to categorise original Hadith detection techniques. *International Journal of Information Technology*, **n/a**(n/a), 1–15. DOI: https://doi.org/10.1007/s41870-021-00649-3

Ghazizadeh, M., Zahedi, M.H., Kahani, M. and Bidgoli, B.M. (2008). Fuzzy expert system in determining Hadith validity. In: *Advances in Computer and Information Sciences and Engineering* (pp. 354–359). Bridgeport, CT, USA. 03–12/12/2007.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. United States: Synthesis Lectures on Human Language Technologies, Morgan and Claypool Publishers.

Johnson, J.M. and Khoshgoftaar, T.M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, **6(**1), 1–54.

Najeeb, M.M. (2014). Towards innovative system for Hadith isnad processing. *International Journal of Computer Trends and Technology*, **18**(6), 257–59.

Najiyah, I., Susanti, S., Riana, D. and Wahyudi, M. (2017). Hadith degree classification for Shahih Hadith identification web based. In: *Proceedings of the 2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1–6), Denpasar, Indonesia, 08–10/08/2017.

Patwa, P., Aguilar, G., Kar, S., Pandey, S. and PYKL, S. (2020). Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. (pp. 774–790). Barcelona, Spain, 12–3/12/2020.

Refaee, E. (2017). Sentiment analysis for micro-blogging platforms in Arabic. In: *Proceedings International Conference on Social Computing and Social Media* (pp. 275–294). Proceedings: Cham, Springer, Vancouver, Canada, 15–20/07/2017.

UNESCO. (n.d.). *UNESCO: World Arabic Language Day*. Available at: https://en.unesco.org/commemorations/worldarabiclanguageda y (Accessed on 02/10/2021).

Zerrouki, T. (2010). *PyArabic, an Arabic Language Library for Python*. Available at: https://pypi.org/project/PyArabic/ (Accessed on 25/09/2021).