JCBI
Journal of Computing & Biomedical Informatics Volume 10 Issue 02
ISSN: 2710 - 1606 2026
Research Article
https://doi.org/10.56979/1002/2026/1315

Prophetic Authorial Style Modeling for Detecting Fabricated Hadiths Using AraBERT

Mohammed Shaaban[1*], Ayman Elshenawy[1&2], and Shehab Gamal el-Din[1]

[1]Systems and Computer Engineering Department, Faculty of Engineering, Al-Azhar University, Cairo, 11884, Egypt.
[2]Networks and Cyber Security Department, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, 19328, Jordan.
[*]Corresponding Author: Mohammed Shaaban. Email:eng.mashaaban@yahoo.com

Received: December 11, 2025 **Accepted:** February 25, 2026

Abstract: Identifying the authenticity of the Hadiths attributed to the Messenger of Allah (ﷺ) is a science, known as "The Science of Hadith Terminology" (ʿIlm Muṣṭalaḥ al-Ḥadīth), This science establishes the principles and rules used to evaluate Prophetic Hadiths in terms of authenticity and acceptability by examining both the chain of narration (Isnad) and the text of the Hadith (Matn), as well as the reliability and qualifications of narrators, This research presents a deep learning–based approach for detecting fabricated (Mawdu') Hadiths using Matn-only analysis without dependence on narrators' chains, Our methodology focuses exclusively on the Prophetic speech (Kalām al-Nabī) (ﷺ) within the text, We fine-tune a pre-trained Arabic language model (AraBERT) on a carefully curated and balanced dataset of authentic (Sahih) and fabricated (Mawḍūʿ) Hadith texts. Experimental results show that the proposed approach succeeds in detecting fabricated hadith with an accuracy of 87%, and ROC-AUC of 0.92. This work emphasizes that the model is intended as an auxiliary analytical aid, and not a replacement for traditional scholarly verification.

Keywords: Hadith Authentication; AraBERT; Transformer Models; Text Classification; NLP; Fabricated Hadith Detection

1. Introduction

Writing style detection (also known as Stylometry) is the evaluation of a piece of text to identify the unique writing style of the author [1]. This can be used to verify identity, detect AI generation, or determine if multiple documents were written by the same person. In this research, we wanted to apply a writing style detection approach to Prophetic Hadith (ﷺ), although the Prophetic Hadiths were originally oral traditions. In the field of Computer Science, any digitally stored linguistic data is formally categorized as 'Text'. Consequently, the computational process of analyzing a speaker's or author's unique linguistic characteristics is technically referred to as Writing Style Detection or Authorship Attribution, and the Prophetic Hadiths were documented textual forms that preserve consistent structural and stylistic patterns. Therefore, this study employs writing style detection techniques to analyze the recorded 'textual' representation of these oral sayings, identifying the underlying linguistic of the Prophet (ﷺ) through the Matn. The integration of Artificial Intelligence (AI) for writing style detection in Hadith studies represents a significant added value in "Digital Forensics" for Islamic heritage.

The purpose of this research is to apply AI technology in the field of Natural Language Processing (NLP), specifically focusing on the Arabic language and the Hadith (Prophetic Traditions). Authenticating the Hadith ascribed to Prophet Muhammad (ﷺ) is a historical problem that has been carefully addressed by Muslim scholars. Classical machine learning techniques have been employed for Hadith classification problems using manually designed linguistic features such as n-grams, term frequency-inverse document frequency (TF-IDF), and statistical features [2, 3]. These traditional methods are inefficient in identifying the semantic and contextual patterns that exist in Classical Arabic religious texts.

Deep learning techniques, specifically transformer-based language models, have achieved outstanding results in natural language understanding applications [4]. Furthermore, a lot of recent research has demonstrated the efficiency of the transformer architecture in language understanding and carrying out tasks based on that understanding. For example, Youness et al. proposed an AraT5-based transformer model for Arabic dialogue generation, demonstrating the capability of encoder–decoder transformers to produce coherent Arabic conversational responses [5]. In another work, a bidirectional attentional mechanism was introduced to enhance contextual understanding in Arabic chatbot systems [6]. Additionally, deep neural architectures have been employed for dialog generation in Arabic conversational agents to improve human–computer interaction in Arabic environments [7]. Furthermore, Elshenawy et al. presented a comprehensive overview of the evolution of deep learning models and their applications across various intelligent systems, highlighting the growing importance of transformer architectures in modern artificial intelligence research [8]. These developments further support the use of transformer-based models such as AraBERT for Arabic text analysis tasks.

AraBERT is a transformer-based pre-trained language model that utilizes contextual embeddings to represent the semantic, syntactic, and stylistic properties of entire sentences, rather than individual words [9, 10], and this is particularly advantageous in Hadith Matn analysis.

As a result, this study uses AraBERT to introduce a Matn-based approach for Hadith authenticity classification. The proposed approach is designed to automatically classify Hadiths as authentic (Sahih) or fabricated (Mawḍūʿ) without using Isnad data, solely based on the text content. This method can be used in addition to the traditional sciences of Hadiths [11, 12], and is a scalable computational method for analyzing large Hadith datasets. The experiments were carried out on a selected dataset of the Prophet's (ﷺ) written sayings. The model was trained on two balanced datasets: authentic Hadiths and fabricated Hadiths. To preserve the integrity of the outcome, class balancing was ensured, both by limiting the study to 'sayings' (textual content) and by having an equal number of instances for each class. The results showed a strong ability of the model to classify authentic and fabricated Hadiths, with an accuracy of 87%.

While this study is inspired by the concept of writing style detection and authorial style analysis, the experimental setup implemented in this work primarily constitutes a binary text classification task applied to curated Hadith Matn texts. Although stylistic characteristics may implicitly contribute to the representations learned by the transformer model, the current experimental design mainly demonstrates the capability of deep learning models to capture discriminative linguistic patterns within Hadith texts. Therefore, the results should be interpreted primarily as evidence of effective Matn-based classification rather than a direct and complete modeling of Prophetic authorial style.

The main contribution of this paper can be summarized as follows: (i) A fabricated Hadiths detection framework is proposed using only the textual content (Matn) without relying on the chain of narration (Isnad). (ii) A fine-tuned AraBERT model is applied to capture linguistic patterns that may reflect stylistic characteristics of Prophetic speech.(iii) A high-quality dataset containing only the Prophet's direct speech with extensive preprocessing, filtering, and duplicate removal is built. (iv) Demonstrates the effectiveness of the proposed approach and establishes a baseline for future research. (v) Practically implements a model-based tool for testing Hadith texts, supporting AI-assisted analysis in Islamic studies.

1.1. Related work

1.1.1. Writing Style Detection and Authorship Attribution: General Overview

The increase in research on writing style detection and authorship attribution has expanded into various domains. The following review highlights some of these studies and their related domains:

- In the cross-lingual style analysis domain, Škorić et al. [13] proposed parallel stylometric document embeddings using deep language models across seven European languages. A bi-directional LSTM model for Kannada language cross-domain writing style identification was proposed by Chandrika and Kallimani [14], reaching an accuracy of 77.8% and proving effective for low-resource languages. Saputra and Riccosan [15] applied IndoBERT within a multi-label, multi-class framework to jointly analyse stylistic patterns and demographic attributes such as author gender of Indonesian news articles. In Japanese literary studies, Kanda et al. [16] introduced an ensemble approach that combines BERT-based embeddings with traditional stylistic features, achieving strong performance in small-sample scenarios.

- In the Large Language Models domain and distinguishing human-authored texts from machine-generated content: Alghamdi et al. proposed the ABERT approach, which helps to discriminate between the human-authored and AI-generated text with a reduction of trainable parameters by about 67.7%[17]. Kaushik et al. proposed the embedding fusion approach, which uses the semantic representations of the masked language models and the encoder-decoder models to achieve high accuracy, i.e., above 96%, to discriminate between the human-authored and AI-generated text [18]. Silva et al. [19, 20] proposed the model based on the GAN-BERT and Forged-GAN-BERT approaches to detect the forged literary text generated by LLMs like ChatGPT.
- In the software forensics domain, Czibula et al. [21] proposed an ensemble of deep autoencoder models (AutoSoft), developed to detect software coding style.
- In cybersecurity contexts: Shin et al. [22] proposed an approach to link users across Dark Web and Surface Web platforms by combining BERTopic with writing style analysis methods.
- In short texts such as tweets, Oliva et al. [23] proposed LSTM-based models with max-pooling mechanisms to capture salient stylistic features in short texts.
- In Historical and religious texts studies: AlZahrani and Al-Yahya [24] achieved accuracies of up to 96% by fine-tuning Arabic pre-trained models such as ARBERT and AraELECTRA to analyse stylistic patterns in Islamic legal texts and Classical Arabic. Alqurashi et al. [25] applied transformer-based models using CAMeLBERT and topic modeling to perform authorship attribution on a large dataset of classical Arabic poetry, achieving F1 scores ranging from 0.97 to 1.0. In a similar context, Ancient Greek texts, Schmidt et al. [26] applied fine-tuned GreBERT in the Pseudo-Dionysian Ars Rhetorica to detect stylistic multi-authorship
- In writing style change detection: Writing Style Change Detection (SCD) focuses on identifying the positions where the writing style changes within a document, often indicating a transition between authors. Alsheddi and Menai proposed a Boundary-Focused Large Language Model Adaptation (BF-LLMA-SCD) method that fine-tunes decoder-based large language models using QLoRA. Their approach achieved state-of-the-art performance on several PAN datasets and strong results on an Arabic SCD dataset, reaching F1 scores above 0.99 on easy instances [27]. In the same direction, Weerasinghe et al. investigated stylometric patterns in AI-generated texts by applying an authorship verification model trained on human-written data. Their study analysed texts generated by multiple large language models such as GPT-2, GPT-3, ChatGPT, and LLaMA, showing that each model exhibits distinctive stylistic characteristics that can be detected using stylometric analysis [28].

1.1.2. Writing Style Detection and Authorship Attribution in Hadith:

Writing Style Detection and Authorship Attribution is the process of analysing textual features to determine the relationship between an author and their written work. Therefore, applying authorship attribution to Hadith studies primarily focuses on the Matn (text), excluding Isnad (chain of narration) analysis. This section reviews prior work related to Hadith authenticity detection, including machine learning, deep learning, and Isnad-based approaches for comparison. Finally, highlighting how the proposed approach differs from existing studies.

Prior work in this area is divided into three sections as follows:

1. Traditional machine learning algorithms, such as Naïve Bayes, Support Vector Machines (SVM), and Decision Trees [2, 3, 29], these approaches were limited in capturing semantic relationships within Arabic texts.
2. Deep learning models, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks approaches applied to automatically learn textual representations from Hadith corpora [3, 29]. These models recognize sequential patterns as well as contextual dependencies.
3. Transformer-based models like AraBERT and multilingual BERT, which recognize the nuances of the Arabic language in semantics as well as syntax, for Hadith authenticity detection approaches [29,30]. Overall, the above studies are summarized in Table 1

Table 1. Summary of Prior Studies on Automated Hadith Authentication

Reference (Year)	Input Data Used	Model Type (Best Performing)	Classification Task (Classes)	Best Performance (Accuracy / F1)
Abdelaal & Youness (2019) [2]	Isnad-Only (Reporters' Reliability)	Naïve Bayes (NB)	Multi-class: Sahih, Hasan, Da'if, Mawḍūʿ	93.75% Accuracy (cross-validation)
Tarmom et al. (2022) [3]	Isnad-Only	CNN	Binary: Authentic vs. Non-authentic	93% Accuracy
Tarmom et al. (2022) [3]	Matn-Only	LSTM	Binary: Authentic vs. Non-authentic	85% Accuracy
Refaee (2022) [29]	Isnad + Matn (Full Hadith)	ARBERT (DL Model)	Binary: Sahih vs. Daif	91.56% Accuracy / 88.64% F1
Refaee (2022) [29]	Matn-Only	ARBERT (DL Model)	Binary: Sahih vs. Daif	75.30% Accuracy / 62.53% F1
Gaanoun & Alsuhaibani (2022) [30]	Matn-Only	CAMeLBERT-CA	Binary: Sahih vs. Mawḍūʿ (MH Detection)	92.47% F1
Alghamdi et al. (2025) [31]	Isnad + Matn (Full Hadith)	AraBERT (PLM)	Binary: Genuine vs. Fake	99.94% F1
Alghamdi et al. (2025) [31]	Matn-Only	CamelBERT (PLM)	Binary: Genuine vs. Fake	98.54% F1

In some previous studies, Hadith authenticity verification was limited by treating the entire Matn (text) as a single unit of analysis. This approach often included 'noise' within the dataset. Specifically, dialogues from companions or descriptive narrations of the Prophet's (ﷺ) actions and commands that do not form his actual direct speech. This research overcomes these limitations by enhancing the data granularity and isolating the Prophet's (ﷺ) direct speech from the noise in the context.

2. Materials and Methods

2.1. Model Architecture

The proposed system (Figure 1) utilizes a binary text classification task using a fine-tuned AraBERT (aubmindlab/bert-base-arabertv2) model. AraBERT is a language model pre-trained using a transformer architecture and a large corpus of Arabic data. It generates contextual word embeddings, meaning that each word in a sentence is represented differently based on its context.

The components of the proposed system are as follows:

- AraBERT encoder: This acts as the main component of the proposed system, where it transforms the input Hadith texts into word vectors containing semantic, syntactic, and stylistic features using a language model pre-trained on a large corpus of data. It then utilizes a classification head consisting of a fully connected layer followed by a softmax activation function, enabling it to differentiate between Sahih and Mawḍūʿ Hadiths.
- A classification head consisting of a fully connected layer.
- A SoftMax activation function
- The system outputs a probability distribution over two classes: Sahih/Mawḍūʿ Hadiths
- Fine-tuning of the pre-trained model using a balanced dataset to adapt it to the Hadith authenticity detection task.

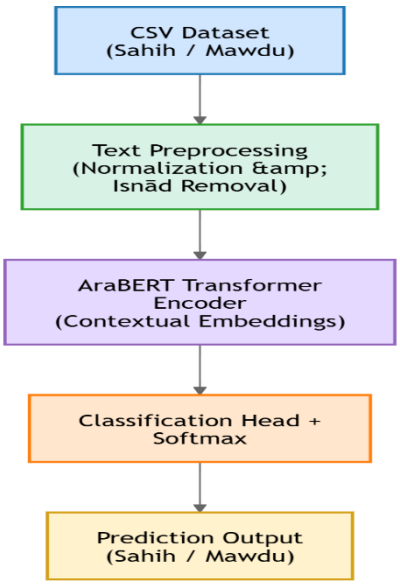


Figure 1. System architecture for hadith authenticity classification.

2.2. Training Procedure

The model was fine-tuned using the Hugging Face Transformers library configured with the following parameters (Table 2):

Table 2. Model configuration for the AraBERT-based approach

Parameter	Value
Validation Samples	389
Number of Epochs	10
Batch Size	8
Maximum Sequence Length	96 tokens
Tokenization	AraBERT Tokenizer
Optimizer	AdamW
Learning Rate	0.00002
Weight Decay	0.01
Loss Function	CrossEntropyLoss
Gradient Clipping	1.0
Warm-up Steps	0

2.3. Evaluation Strategy

To assess the efficiency of the proposed model for classifying Mawḍūʿ (fabricated) Hadiths, standard statistical measures were used. These measures are based on the following parameters obtained from the Confusion Matrix tool: True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), Accuracy, Precision, Recall, and F1 Score.

2.4. Dataset

2.4.1. Data Collection

Two primary datasets were used in this research:

1. **Authentic Hadiths (Sahih):** A collection of authentic Hadith texts extracted from Sahih al-Bukhari. The Sahih Hadith corpus employed in this work was sourced from the publicly available Sahih al-Bukhari dataset hosted on the Hugging Face platform.
2. **Fabricated Hadiths (Mawḍūʿ):** A dataset of fabricated Hadiths collected for the fabricated Hadith class, texts were sourced from the (MAHADDAT dataset provided by Gaanoun and Alsuhaibani, available via the MHDetection GitHub repository. To ensure high data accuracy, Matn manually selected (text of the Hadith that is the actual words spoken by Prophet Muhammad (ﷺ), visually inspecting and reading the content. A pre-processing applied to remove diacritics (Tashkeel) and punctuation. And developed an application (Figure 2) that facilitates manual selection of the Prophet's Hadiths, removing diacritics and punctuation to create a new dataset containing only the Matn. actual words spoken by Prophet Muhammad (ﷺ).

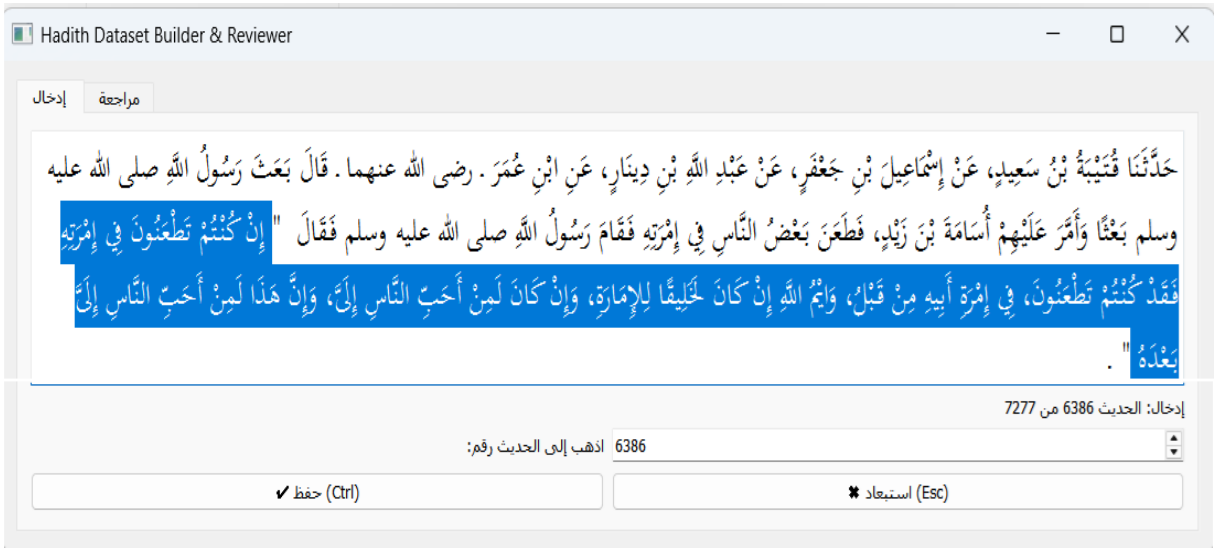


Figure 2. The used App for facilitating pre-processing and selection criteria.

2.4.2. Example for hadith before and after pre-processing

Hadith before pre-processing:

حَدَّثَنَا الْحُمَيْدِيُّ عَبْدُ اللَّهِ بْنُ الزُّبَيْرِ ، قَالَ : حَدَّثَنَا سُفْيَانُ ، قَالَ : حَدَّثَنَا يَحْيَى بْنُ سَعِيدٍ الْأَنْصَارِيُّ ، قَالَ : أَخْبَرَنِي مُحَمَّدُ بْنُ إِبْرَاهِيمَ التَّيْمِيُّ ، أَنَّهُ سَمِعَ عَلْقَمَةَ بْنَ وَقَّاصٍ اللَّيْثِيَّ ، يَقُولُ : سَمِعْتُ عُمَرَ بْنَ الْخَطَّابِ رَضِيَ اللَّهُ عَنْهُ عَلَى الْمِنْبَرِ، قَالَ : سَمِعْتُ رَسُولَ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ، يَقُولُ : " إِنَّمَا الْأَعْمَالُ بِالنِّيَّاتِ، وَإِنَّمَا لِكُلِّ امْرِئٍ مَا نَوَى، فَمَنْ كَانَتْ هِجْرَتُهُ إِلَى دُنْيَا يُصِيبُهَا أَوْ إِلَى امْرَأَةٍ يَنْكِحُهَا، فَهِجْرَتُهُ إِلَى مَا هَاجَرَ إِلَيْهِ

Hadith after pre-processing:

"إنما الأعمال بالنيات وإنما لكل امرئ ما نوى فمن كانت هجرته إلى دنيا يصيبها أو إلى امرأة ينكحها فهجرته إلى ما هاجر إليه"

2.4.3. Selection Criteria

- Excluded Content: Ignoring Hadiths that described the Prophet's actions or orders rather than his direct speech.
- Dialogue Handling: Some Hadiths were dialogues between the Prophet and others. Much of this was ignored.
- Conditions for Inclusion: Including only parts of a dialogue if the Prophet's words formed a complete and useful sentence on their own.
- Fragmented Speech: If the Prophet's words, when isolated, did not convey a complete meaning, the Hadith was ignored. Table 3 illustrates the selection criteria by examples

2.4.4. Exact Duplicate Removal

Exact textual duplicates were identified and removed separately for each class to prevent overrepresentation of identical Matn texts. For the Sahih class, the initial dataset contained 1683 Hadiths. After removing exact duplicates, 1637 unique Hadiths remained. For the Mawḍūʿ class, the dataset size was reduced from 1637 to 1630 Hadiths.

2.4.5. Near Duplicate and Semantic Similarity Filtering

To mitigate data leakage caused by highly similar Hadith texts, a near-duplicate filtering process based on textual similarity was applied. Before similarity filtering, the dataset consisted of 1637 Sahih and 1630 Mawḍūʿ Hadiths. After removing near-duplicate samples, the dataset was reduced to 1544 unique Sahih Hadiths and 1498 unique Mawḍūʿ Hadiths.

Table 3. Examples illustrate selection criteria

Matn Status	Decision Reason	Full Hadith Text (Arabic)
Rejected	**This Hadith describes an action and a state of the rather (ﷺ)Prophet than quoting his direct spoken words**	**حَدَّثَنَا مُوسَى بْنُ إِسْمَاعِيلَ، قَالَ حَدَّثَنَا أَبُو عَوَانَةَ، قَالَ حَدَّثَنَا مُوسَى بْنُ أَبِي عَائِشَةَ، قَالَ حَدَّثَنَا سَعِيدُ بْنُ جُبَيْرٍ، عَنِ ابْنِ عَبَّاسٍ، فِي قَوْلِهِ تَعَالَى {لاَ تُحَرِّكْ بِهِ لِسَانَكَ لِتَعْجَلَ بِهِ} قَالَ كَانَ رَسُولُ اللَّهِ صلى الله عليه وسلم يُعَالِجُ مِنَ التَّنْزِيلِ شِدَّةً، وَكَانَ مِمَّا يُحَرِّكُ شَفَتَيْهِ ـ فَقَالَ ابْنُ عَبَّاسٍ فَأَنَا أُحَرِّكُهُمَا لَكُمْ كَمَا كَانَ رَسُولُ اللَّهِ صلى الله عليه وسلم يُحَرِّكُهُمَا. وَقَالَ سَعِيدٌ أَنَا أُحَرِّكُهُمَا كَمَا رَأَيْتُ ابْنَ عَبَّاسٍ يُحَرِّكُهُمَا. فَحَرَّكَ شَفَتَيْهِ ـ فَأَنْزَلَ اللَّهُ تَعَالَى {لاَ تُحَرِّكْ بِهِ لِسَانَكَ لِتَعْجَلَ بِهِ * إِنَّ عَلَيْنَا جَمْعَهُ وَقُرْآنَهُ} قَالَ جَمْعُهُ لَهُ فِي صَدْرِكَ، وَتَقْرَأَهُ {فَإِذَا قَرَأْنَاهُ فَاتَّبِعْ قُرْآنَهُ} قَالَ فَاسْتَمِعْ لَهُ وَأَنْصِتْ {ثُمَّ إِنَّ عَلَيْنَا بَيَانَهُ} ثُمَّ إِنَّ عَلَيْنَا أَنْ تَقْرَأَهُ. فَكَانَ رَسُولُ اللَّهِ صلى الله عليه وسلم بَعْدَ ذَلِكَ إِذَا أَتَاهُ جِبْرِيلُ اسْتَمَعَ، فَإِذَا انْطَلَقَ جِبْرِيلُ قَرَأَهُ النَّبِيُّ صلى الله عليه وسلم كَمَا قَرَأَهُ.**

Rejected	**This Hadith describes the Prophet's (ﷺ) noble character and actions; it does not contain a direct quote spoken by him.**	**حَدَّثَنَا عَبْدَانُ، قَالَ أَخْبَرَنَا عَبْدُ اللَّهِ، قَالَ أَخْبَرَنَا يُونُسُ، عَنِ الزُّهْرِيِّ، ح وَحَدَّثَنَا بِشْرُ بْنُ مُحَمَّدٍ، قَالَ أَخْبَرَنَا عَبْدُ اللَّهِ، قَالَ أَخْبَرَنَا يُونُسُ، وَمَعْمَرٌ، عَنِ الزُّهْرِيِّ، نَحوَهُ قَالَ أَخْبَرَنِي عُبَيْدُ اللَّهِ بْنُ عَبْدِ اللَّهِ، عَنِ ابْنِ عَبَّاسٍ، قَالَ كَانَ رَسُولُ اللَّهِ صلى الله عليه وسلم أَجْوَدَ النَّاسِ، وَكَانَ أَجْوَدُ مَا يَكُونُ فِي رَمَضَانَ حِينَ يَلْقَاهُ جِبْرِيلُ، وَكَانَ يَلْقَاهُ فِي كُلِّ لَيْلَةٍ مِنْ رَمَضَانَ فَيُدَارِسُهُ الْقُرْآنَ، فَلَرَسُولُ اللَّهِ صلى الله عليه وسلم أَجْوَدُ بِالْخَيْرِ مِنَ الرِّيحِ الْمُرْسَلَةِ.**
Accepted	**This text constitutes the direct words spoken by the Prophet (ﷺ), which is the specific Matn required for the dataset.**	**حَدَّثَنَا عُبَيْدُ اللَّهِ بْنُ مُوسَى، قَالَ أَخْبَرَنَا حَنْظَلَةُ بْنُ أَبِي سُفْيَانَ، عَنْ عِكْرِمَةَ بْنِ خَالِدٍ، عَنِ ابْنِ عُمَرَ ـ رضى الله عنهما ـ قَالَ قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم "بُنِيَ الإِسْلاَمُ عَلَى خَمْسٍ شَهَادَةِ أَنْ لاَ إِلَهَ إِلاَّ اللَّهُ وَأَنَّ مُحَمَّدًا رَسُولُ اللَّهِ، وَإِقَامِ الصَّلاَةِ، وَإِيتَاءِ الزَّكَاةِ، وَالْحَجِّ، وَصَوْمِ رَمَضَانَ."**
Rejected	**The text is a dialogue from which a single, independent, and useful sentence of the Prophet's speech cannot be derived.**	**حَدَّثَنَا أَحْمَدُ بْنُ يُونُسَ، وَمُوسَى بْنُ إِسْمَاعِيلَ، قَالاَ حَدَّثَنَا إِبْرَاهِيمُ بْنُ سَعْدٍ، قَالَ حَدَّثَنَا ابْنُ شِهَابٍ، عَنْ سَعِيدِ بْنِ الْمُسَيَّبِ، عَنْ أَبِي هُرَيْرَةَ، أَنَّ رَسُولَ اللَّهِ صلى الله عليه وسلم سُئِلَ أَىُّ الْعَمَلِ أَفْضَلُ فَقَالَ "إِيمَانٌ بِاللَّهِ وَرَسُولِهِ". قِيلَ ثُمَّ مَاذَا قَالَ "الْجِهَادُ فِي سَبِيلِ اللَّهِ". قِيلَ ثُمَّ مَاذَا قَالَ "حَجٌّ مَبْرُورٌ."**
Accepted	**Although it is a dialogue, a complete and useful sentence of the Prophet's direct speech was successfully extracted.**	**حَدَّثَنَا أبو بكر بن أبي شيبة... عَنْ عَائِشَةَ قَالَتْ: جَاءَتْنِي امْرَأَةٌ، وَمعهَا ابْنَتَانِ لَهَا، فَسَأَلَتْنِي فَلَمْ تَجِدْ عِنْدِي شَيْئًا غَيْرَ تَمْرَةٍ وَاحِدَةٍ... فَدَخَلَ عَلَيَّ النبيُّ صَلَّى اللَّهُ عليه وَسَلَّمَ فَحَدَّثْتُهُ حَدِيثَهَا، فَقَالَ النبيُّ صَلَّى اللَّهُ عليه وَسَلَّمَ: "مَنِ ابْتُلِيَ مِنَ البَنَاتِ بشيءٍ، فأحْسَنَ إلَيْهِنَّ كُنَّ له سِتْرًا مِنَ النَّارِ."**

2.4.6. Dataset Balancing

To prevent class imbalance and biased predictions, the dataset was balanced before training. Before balancing, the dataset contained 1544 Sahih and 1498 Mawḍūʿ Hadiths. Random down-sampling was applied to the majority class (Sahih), removing 46 samples, resulting in a perfectly balanced dataset of 1498 Hadiths per class. And no samples were removed from the Mawḍūʿ class during this stage.

2.4.7. Final Dataset Statistics

After pre-processing and balancing, the final dataset contained 2,996 Hadith texts, evenly distributed Equally across both classes (Sahih and Mawḍūʿ), An additional external test set of 400 Hadith texts (200 Sahih, 200 Mawḍūʿ) was kept fully unseen during training and validation to evaluate the generalization capability of the model, the dataset used for model training consisted of a total of 2,596 Hadith Matn texts.

The dataset was divided into training and validation subsets using an approximate 85/15 split, as follows:

- **Training Set:** 2,207 samples Used for fine-tuning the AraBERT model parameters.
- **Validation Set:** 389 samples used for monitoring model performance and hyperparameter stability.

Figure 3 shows a flowchart for the proposed Hadith dataset processing

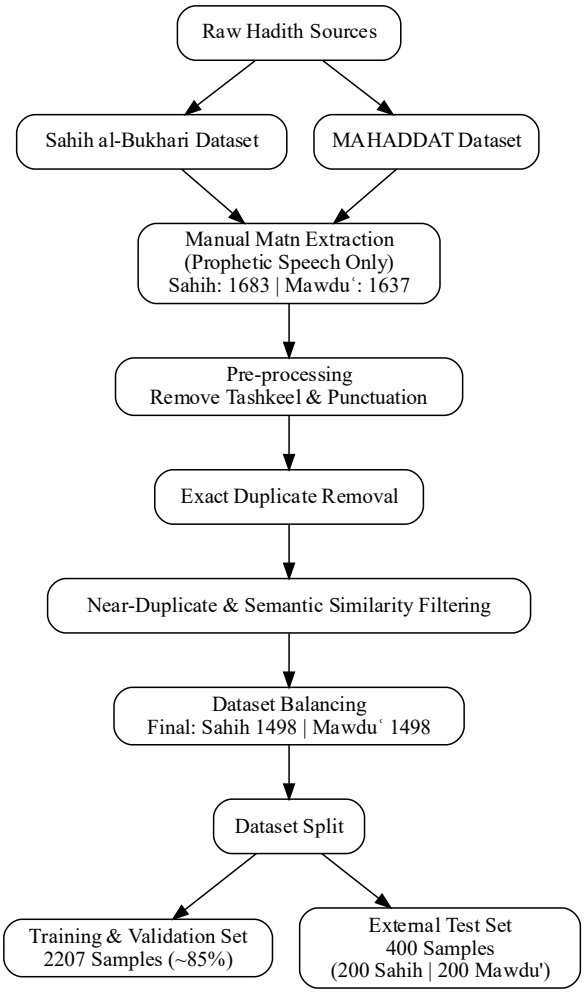


Figure 3. Data processing flowchart for the proposed Hadith dataset.

3. Experimental Results

3.1. Training Behavior

The training procedure shows stable and efficient convergence, as can be seen from the training loss curve (Table 4). There is a sudden drop in the loss value in the initial few epochs, which shows efficient learning of the discriminative textual features. After Epoch 4, the loss value drops slowly, which shows that the model has learned most of the linguistic patterns.

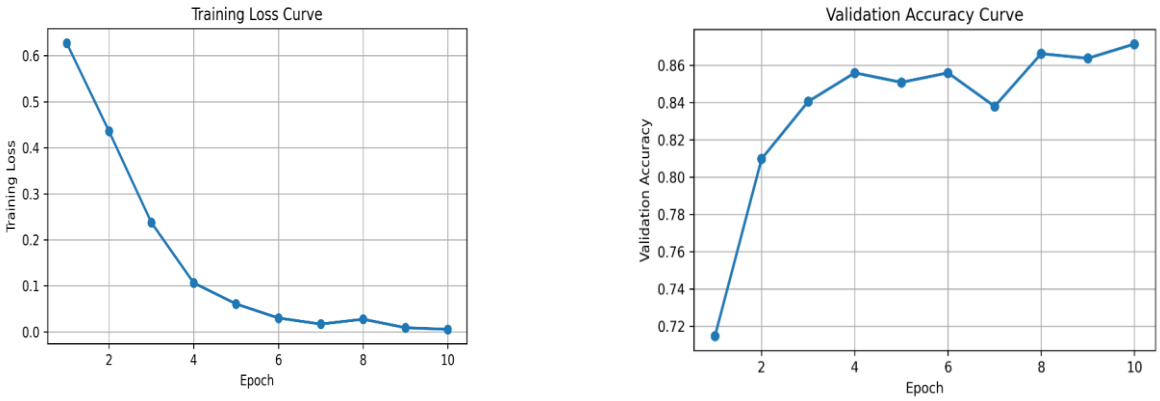
Validation accuracy follows a similar trend, increasing steadily during early epochs and stabilizing in later stages (from 71.47% in Epoch 1 to 87.15% in Epoch 10), while ROC-AUC values remain above 0.90 from Epoch 2 onward. This behavior indicates successful fine-tuning of the pre-trained AraBERT model without training instability. From Figure 4-a, it is observed that the training loss reduces sharply in the initial Epochs, implying successful learning of discriminative features, followed by a gradual stabilization, indicating successful convergence.

Table 4. Epoch-wise Performance Metrics and Confusion Matrix Results on the Validation Set

Epoch	Loss	Accuracy	ROC-AUC	Precision	Recall	F1-Score	TN	FP	FN	TP
1	0.6277	71.47	0.8240	0.656	**0.940**	0.773	89	99	12	189
2	0.4360	80.98	0.9067	0.772	0.896	0.829	135	53	21	180
3	0.2376	84.06	0.9110	0.872	0.811	0.840	164	24	38	163
4	0.1064	85.60	0.9191	0.876	0.841	0.858	164	24	32	169
5	0.0607	85.09	0.9262	0.839	0.881	0.859	154	34	24	177
6	0.0301	85.60	0.9243	0.847	0.881	0.864	156	32	24	177
7	0.0169	83.80	**0.9267**	0.787	**0.940**	0.857	137	51	12	189

8	0.0275	86.63	0.9230	**0.878**	0.861	0.869	164	24	28	173
9	0.0090	86.38	0.9246	0.866	0.871	0.868	161	27	26	175
10	0.0054	**87.15**	0.9260	0.861	0.895	**0.878**	159	29	21	180

Validation accuracy increases rapidly during the early epoch, then stabilizes in the later stages, as depicted in Figure 4-b. This indicates that the model has attained effective generalization without overfitting. The highest F1-score, 0.878, is attained at Epoch 10, which signifies the best compromise between precision and recall. This phenomenon reveals the inherent trade-off between maximizing the sensitivity of the detector and minimizing false positives, as might be anticipated in the linguistically ambiguous task of Hadith Matn analysis.



a. Training loss curve b. Validation accuracy curve

Figure 4. Training loss and validation accuracy curves

3.2. Confusion Matrix Analysis:

The confusion matrix (Figure 5) shows the classification model's performance in distinguishing Sahih from Mawḍūʿ Hadiths. In this analysis, the Mawḍūʿ class is considered the positive class. At the end of the 10th epoch, the model has the following measures:

- Precision (0.861): Measures the accuracy of the model in identifying Mawḍūʿ Hadiths among all the positive predictions made by the classifier.
- Recall (0.896): Measures the effectiveness of the model in identifying the majority of Mawḍūʿ Hadiths.
- F1-Score (0.878): Measures the balance of the model, which is high, indicating a stable model.
- True Positives (Mawḍūʿ): The model successfully classified 180 Mawḍūʿ Hadiths.
- True Negatives (Sahih): The model correctly identified 159 Sahih Hadiths.
- Misclassifications: There were only 29 false positives and 21 false negatives.

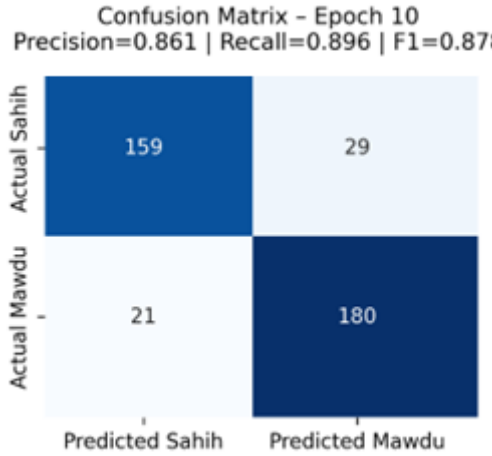


Figure 5. Confusion matrix evaluated on the validation set (Epoch 10).

3.3. Model Evaluation by External Test Set

An external test set was isolated from training and validation to ensure unbiased assessment of model generalization on previously unseen data. The evaluation results are shown in Table 5. The external test set was used only once after training was completed to provide an unbiased evaluation of the final model (Epoch 10).

Table 5. Confusion Matrix Results on the External Test Set.

accuracy	Precision	recall	f1_score	roc_auc	TN	FP	FN	TP
0.8525	0.8615	0.84	0.8506	0.9381	173	27	32	168

The external test results show performance values that are closely aligned with those observed on the validation set. While validation accuracy ranged approximately between 84% and 87% with ROC-AUC values around 0.92–0.93, the external test set achieved comparable accuracy values , the model achieved a validation accuracy of approximately 87.1% with a ROC-AUC of about 0.93, while the external test evaluation yielded an accuracy of 85.25% and a ROC-AUC of 0.938. The small and expected reduction in accuracy. Similarly, precision, recall, and F1-score on the external data follow patterns consistent with validation results, indicating stable classification behavior and limited performance degradation on unseen data. This close match between validation and external evaluation metrics shows that the model is performing well in terms of generalization and is not dependent on any specific features of the dataset.

3.4. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve was obtained after testing the trained model on the external test set using the checkpoint at Epoch 10, as depicted in Figure 6. The ROC curve has an AUC of approximately 0.94, which is a clear indication of excellent discrimination ability for genuine (Sahih) and forged (Mawḍūʿ) Hadith texts.

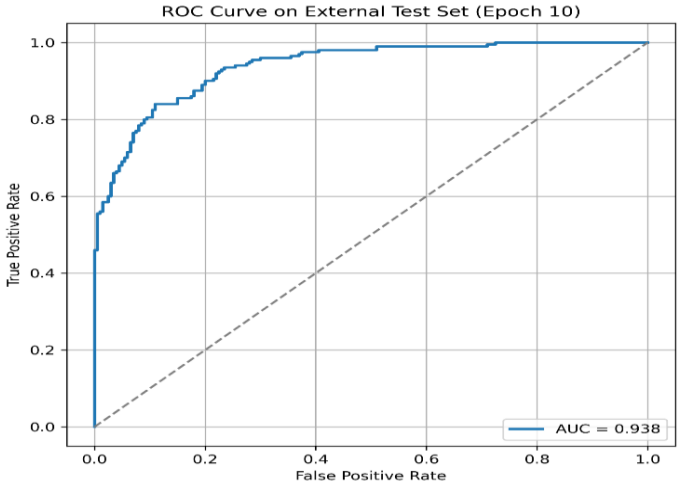


Figure 6. Receiver Operating Characteristic (ROC) curve of the model

This shows that the curve has a high true positive rate with a relatively low false positive rate, thereby confirming that the model is capable of identifying the fabricated Hadiths while being robust to false positives. The model's performance on the ROC curve further verifies that the model has good generalization capabilities.

4. Discussion

The experimental results show that the proposed approach using fine-tuning the AraBERT model is capturing Hadith authenticity using only the Matn text, with an accuracy reached to 87%. The training process showed a stable learning behavior where the loss function decreased rapidly during the initial stages and then gradually plateaued. This is an indication that the model has learned without instability during training. From the evaluation results, the model has achieved balanced performance across various evaluation metrics, with an F1-score has been achieved at 0.878. This is a good balance between precision and recall. This is important in Hadith authenticity detection tasks because both types of errors are critical in different ways. For instance, high recall means that most authentic Hadiths are correctly classified, whereas high precision means that the number of misclassifications is relatively small.

Table 6. Comparison between this experiment and previous work experiments.

Reference (Year)	Transformer name	Dataset Size	Sahih	Mawḍūʿ	Accuracy	F1-Score
Refaee (2022) [29]	ARBERT [29]	3944	3000	944	75.3	62.53

Gaanoun & Alsuhaibani (2022) [30]	AraBERTv2 [30]	26561	24109	2452	98.55	90.99
Gaanoun & Alsuhaibani (2022) [30]	ArabicBERT [30]	26561	24109	2452	98.67	92.01
Gaanoun & Alsuhaibani (2022) [30]	ARBERT [30]	26561	24109	2452	98.43	90.99
Gaanoun & Alsuhaibani (2022) [30]	CAMeLBERT MSA [30]	26561	24109	2452	98.1	88.74
Gaanoun & Alsuhaibani (2022) [30]	CAMeLBERT CA [30]	26561	24109	2452	98.68	92.47
Gaanoun & Alsuhaibani (2022) [30]	mBERT [30]	26561	24109	2452	97.1	80.62
This study	**AraBERT [this paper]**	**2996**	**1498**	**1498**	**87.15%**	**87.8**

Analysis of the confusion matrix is also important in understanding how well the model classifies Hadith texts. From the results, it is evident that the model has been able to correctly identify a large number of Sahih and Mawḍūʿ Hadiths. This shows that the transformer architecture can learn subtle stylistic features that may be used to distinguish between authentic and fabricated Hadiths. Although the study is motivated by the concept of authorial style modeling, the present work should be viewed primarily as a Matn-based classification framework. Future work may investigate more explicit stylistic modeling techniques, such as stylometric feature analysis, authorship verification setups, or explainable AI methods to better interpret stylistic signals learned by the model.

Table 6 compares the performance of the proposed approach with several previous transformer-based studies for Hadith authenticity detection using Matn text. The results show that earlier works, particularly the study by Gaanoun & Alsuhaibani (2022), achieved higher accuracy values (around 97–98%) and F1-scores above 0.90, . However, direct comparison between these results and the present study should be interpreted cautiously, as the datasets differ in size and preprocessing procedures. In the present study, a carefully curated dataset was constructed, focusing exclusively on the Prophet's direct speech within the Matn and applying strict filtering procedures to remove narrative descriptions and dialogue fragments. While the resulting dataset is smaller, it provides a controlled experimental setting designed to reduce noise and data leakage. In contrast, Refaee (2022) used a dataset of 3,944 samples, smaller than the dataset used by Gaanoun & Alsuhaibani (2022) but larger than the dataset used in this study, achieving 75.3% accuracy and an F1-score of 0.625, which is lower than the results obtained in this study. Consequently, the reported accuracy of 87.15% and F1-score of 0.878 should be interpreted within the context of this curated dataset. Rather than aiming to directly outperform previous studies, the present work primarily demonstrates that stylistic modeling of Prophetic direct speech using AraBERT can provide a reliable auxiliary signal for identifying potentially fabricated Hadith texts. Future work may enable more direct comparisons by evaluating multiple models on standardized datasets.

Figure 7 illustrates the implementation of the proposed AraBERT-based Hadith authenticity detection system through a prototype application interface. The application allows users to input the full Hadith text, after which the system performs a preprocessing step to extract and display the Matn (the Prophet's direct speech) after cleaning and filtering. The interface then enables the user to specify or review the actual label of the Hadith and run the classification process. Once the analysis is executed, the system displays the model's prediction, indicating whether the Hadith is Ṣaḥīḥ (authentic) or Mawḍūʿ (fabricated), along with a confidence score reflecting the model's certainty. The interface also shows whether the prediction matches the true label, providing a simple mechanism for evaluation and validation. This application demonstrates the practical deployment of the trained AraBERT model, enabling interactive testing of

Hadith texts and illustrating how the proposed approach can support AI-assisted analysis in Hadith studies.

Figure 7. The deployed model-based application.

5. Conclusions

This study proposes an automated approach for classifying the authenticity of Hadith Matn using a fine-tuned AraBERT model. A comprehensive preprocessing stage was performed to ensure high data quality before training. The experimental results demonstrate that the proposed method performs effectively and robustly across multiple evaluation metrics. During training, the validation accuracy steadily improved, reaching 87%, while the ROC-AUC exceeded 0.92, indicating strong class separability. Evaluation on an independent test dataset showed only a slight and acceptable decrease in performance, confirming the model's generalization capability. Furthermore, analysis of the confusion matrix, precision, recall, and F1-score indicates that the model achieves balanced performance across both classes, with slightly higher recall for the Mawḍūʿ class. This suggests that the model is capable of identifying many fabricated narrations within the evaluated dataset. However, this observation should be interpreted cautiously, as the experiment was conducted on a curated and balanced dataset, and further class-wise

evaluation on larger and more diverse corpora would be required to confirm consistent effectiveness in detecting fabricated Hadiths. Future work could enhance performance by leveraging more advanced Arabic language models such as Jais, expanding the dataset, and applying the approach to additional Hadith corpora to improve generalization. Another promising direction is the identification of Hadith fabricators (Wadda'ūn) through their distinctive writing styles. Overall, this research demonstrates the potential of transformer-based Arabic language models for automatic Hadith Matn authenticity classification and establishes a solid foundation for further interdisciplinary research at the intersection of artificial intelligence and Islamic studies.

References

1. Koppel, M.; Schler, J.; Argamon, S. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.* 2009, *60*, 9–26.
2. Abdelaal, H.M.; Youness, H.A. Hadith classification using machine learning techniques according to its reliability. Rom. J. Inf. Sci. Technol. 2019, 22, 259–271.
3. Gharaibeh, H., Al Mamlook, R.E., Samara, G. et al. Arabic sentiment analysis of Monkeypox using deep neural network and optimized hyperparameters of machine learning algorithms. Soc. Netw. Anal. Min. 14, 30 (2024). https://doi.org/10.1007/s13278-023-01188-4
4. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2020; pp. 38–45.
5. Youness, F.; Elshenawy, A.; Madkour, M.A. Arabic dialogue generation using AraT5 transformer. Int. J. Inf. Technol. 2025. https://doi.org/10.1007/s41870-025-02407-1.
6. Youness, F.; Elshenawy, A.; Madkour, M.A. Bidirectional attentional mechanism for Arabic chatbot. Int. J. Inf. Technol. 2024, 16, 3109–3120. https://doi.org/10.1007/s41870-024-01777-2.
7. Youness, F.; Madkour, M.A.; Elshenawy, A. Dialog generation for Arabic chatbot. Int. J. Inf. Technol. 2024, 16, 881–890. https://doi.org/10.1007/s41870-023-01519-w.
8. Elshenawy, A.; Mohammed, A.; Hamouda, S. The evolution of deep learning: Models, applications, and future directions. In Proceedings of the International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC); Cairo, Egypt, 2025; pp. 1–8. https://doi.org/10.1109/MIUCC66482.2025.11196834.
9. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT); 2020; pp. 9–15.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT; 2019; pp. 4171–4186.
11. Ibn Ḥajar al-Asqalānī. Nukhbat al-Fikar (Exquisite Thoughts on the Science of Hadith); Turath Publishing: London, UK, 2011.
12. Ibn al-Ṣalāḥ al-Shahrazūrī. An Introduction to the Science of the Hadith; Garnet Publishing: Reading, UK, 2006.
13. Škorić, M.; Bašić, B.; Škorić, L.; et al. Parallel stylometric document embeddings with deep learning-based language models in literary authorship attribution. Mathematics 2022, 10, 838.
14. Chandrika, C.P.; Kallimani, J.S. Authorship attribution on Kannada text using bi-directional LSTM technique. Int. J. Adv. Comput. Sci. Appl. 2022, 13.
15. Saputra, K.E.; Riccosan. Indonesian news article authorship attribution multilabel multiclass classification using IndoBERT. IAES Int. J. Artif. Intell. 2024, 13, 4688–4694.
16. Kanda, T.; Jin, M.; Zaitsu, W. Integrated ensemble of BERT- and feature-based models for authorship attribution in Japanese literary works. Front. Artif. Intell. 2025, 8, 1624900. https://doi.org/10.3389/frai.2025.1624900.
17. Alghamdi, J.; Lin, Y.; Luo, S. ABERT: Adapting BERT model for efficient detection of human and AI-generated fake news. Int. J. Inf. Manag. Data Insights 2025, 5, 100353. https://doi.org/10.1016/j.jjimei.2025.100353.
18. Kaushik, A.R.; Rufus, R.P.S.; Ratha, N. Enhancing authorship attribution through embedding fusion: A novel approach with masked and encoder–decoder language models. In Proceedings of the International Conference on Pattern Recognition (ICPR); Springer: Cham, Switzerland, 2025.
19. Silva, K.; et al. Authorship attribution of late 19th century novels using GAN-BERT. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL); 2023; pp. 310–320.
20. Silva, K.; et al. Forged-GAN-BERT: Authorship attribution for LLM-generated forged novels. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics; 2024; pp. 325–337.
21. Czibula, G.; Lupea, M.; Briciu, A. Enhancing the performance of software authorship attribution using an ensemble of deep autoencoders. Mathematics 2022, 10, 2572.
22. Shin, G.Y.; et al. Identifying similar users between dark web and surface web using BERTopic and authorship attribution. Electronics 2025, 14, 148.
23. Oliva, C.; et al. Improving LSTMs' under-performance in authorship attribution for short texts. In Proceedings of the ACM Conference; 2018.
24. AlZahrani, F.M.; Al-Yahya, M. A transformer-based approach to authorship attribution in classical Arabic texts. Appl. Sci. 2023, 13, 7255.

25. Alqurashi, L.; Watson, J.; Blakesley, J.; Sharoff, S. BERT-based classical Arabic poetry authorship attribution. In Proceedings of the 31st International Conference on Computational Linguistics (COLING); 2025; pp. 6105–6119.
26. Schmidt, G.; Vybornaya, V.; Yamshchikov, I.P. Fine-tuning pre-trained language models for authorship attribution of the Pseudo-Dionysian Ars Rhetorica. In Proceedings of the Computational Humanities Research Conference (CHR); 2024.
27. Alsheddi, A.S.; Menai, M.E.B. Boundary-focused large language model adaptation for style change detection in multi-authored text. Appl. Sci. 2026, 16.
28. Weerasinghe, J.; Seepersaud, O.; Smothers, G.; Jose, J.; Greenstadt, R. Be sure to use the same writing style: Applying authorship verification on large-language-model-generated texts. Appl. Sci. 2025, 15.
29. Refaee, E.A. Detecting Hadith authenticity using a deep-learning approach. Sci. J. King Faisal Univ. Basic Appl. Sci. 2022, 23, 80–84.
30. Gaanoun, K.; Alsuhaibani, M. Fabricated Hadith detection: A novel matn-based approach with transformer language models. IEEE Access 2022. https://doi.org/10.1109/ACCESS.2022.3217457.
31. Alghamdi, J.; Albukhari, A.; Al-Dala'in, T. Pre-trained models against traditional machine learning for detecting fake Hadith. Electronics 2025, 14, 3484.
32. Shaaban, M. Writing Style Detection Approach for Fabricated Hadith Using AraBERT. GitHub repository. Available online: https://github.com/engmohammadahmad (accessed on 10 March 2026).