International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ENHANCING ARABIC NAMED ENTITY RECOGNITION THROUGH A HYBRID TRANSFORMER–BILSTM ARCHITECTURE

Wessam Lahmod Nadoos[1] , Behrooz Minaei Bidgoli[1, *]

[1] Iran University of Science and Technology, Tehran, Iran;
basic.wessam.lahmod@uobabylon.edu.iq

[1] Iran University of Science and Technology, Tehran, Iran; b_minaei@iust.ac.ir

Abstract

Arabic Named Entity Recognition (NER) remains a challenging task due to the language's rich morphology, complex syntax, and scarcity of annotated corpora, particularly in specialised domains such as Islamic Hadith literature. Traditional deep learning models, including standalone BiLSTM and Transformer-based architectures, often fail to achieve balanced performance across entity types, especially for rare classes. The present study aims to develop an enhanced hybrid model that combines the contextual understanding capabilities of AraBERT with the sequential learning strength of BiLSTM to improve the accuracy and robustness of Arabic NER. A novel AraBERT–BiLSTM Hybrid Model is proposed and evaluated on the Noor Al-Hadith dataset, comprising 59,430 tagged texts covering eight entity types. The proposed hybrid model achieved a remarkable accuracy of 97.42% and a weighted F1-score of 96.05%, outperforming the standalone AraBERT (accuracy 95.61%) and BiLSTM (accuracy 95.36%) models. Experimental analysis demonstrates that the hybrid model effectively mitigates class imbalance issues and improves the detection of minority entity categories such as Tribes, Books, and Narrators. Furthermore, the convergence behaviour of both training and validation losses confirmed the model's strong generalisation capabilities without overfitting. This research contributes a robust methodology for domain-specific Arabic NER and establishes a foundation for automated analysis of religious and classical Arabic texts. The proposed framework not only enhances entity recognition accuracy but also supports large-scale semantic processing of underexplored Arabic datasets, thereby advancing natural language understanding in low-resource linguistic contexts.

Keywords: Noor Al-Hadith dataset; Deep learning; Class imbalance; Semantic analysis; Low-resource language processing

1. Introduction

Named entity recognition (NER) extracts structured information from unstructured Arabic speech texts. This task is challenging due to the complexity of the dataset's language and the scarcity of annotations. However, hybrid models, such as AraBERT-BiLSTM, show promising results in Arabic named entity recognition [1].

Deep learning methods encounter significant challenges when dealing with archaic language, rare entities, and theological terminologies. To reduce this gap, our research utilises and tests

Received: August 07 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the models of AraBERT, BiLSTM, and the AraBERT-BiLSTM Hybrid Model on the Noor Hadith dataset, with a specific emphasis on identifying rare classes.[2].

The proposal of the new AraBERT-BiLSTM Hybrid Model architecture proves to be better than the individual models [3].

The data is first subjected to intensive preprocessing processes, such as text cleaning, removal of stop words, and tokenisation, before model training. Both enhanced AraBERT and BiLSTM networks have been applied to perform entity recognition, with a particular focus on recognising rare classes that are prevalent in religious writings [4].

Hybrid model that will merge the best of both strategies has been developed, and it will perform at higher metrics. This study enhances the search for Arabic origins of hadith texts, which contributes to automated research and knowledge creation in Islamic studies. Our work, through the integration of natural language processing (NLP) methodology and religious studies, enables the advancement of semantic search and offers scalable processes for processing low-resource religious texts [5].

Transformer-based models have been further explored in recent years in Arabic NLP[6]. The NER techniques implemented in a special application for Islamic Hadiths in the Arabic language, and primarily on sequence-to-sequence models, aim to improve the process of entity recognition and comprehension regarding Islamic sacred texts. H. Mubarak et al.[7] After this encoding of words and tags in the data provided, the words or tags are encoded using the BIO (Begin, Inside, Out) format. [8]. Rahim (2021)[9] presented Arabic word representations using a semantic similarity-based model. A polysemy model is proposed, utilising multi-sense vectors and chunk-based processing of word order. The approach resulted in a Pearson correlation of 0.743 on SemEval 2017[10], though it did not perform well due to morphological complexity (0.354 on Arabic paraphrasing data). According to the optimisation strategy for hyperparameters, Sunny et al. (2020) [11] suggested LSTM and Bi-LSTM share price models, demonstrating high performance in terms of the RMSE measurement of trends in the open market. The developments concerning Arabic NER were reviewed by Shaalan (2014) [12], highlighting the issues that the Semitic character of the Arabic language and its effects on broader NLP systems have. The paper examined the linguistic resources, annotation standards, and evaluation measures to assess the current methods in the field today. Abu Nada et al. (2020)[13] developed extractive Arabic text summarisation, which follows AraBERT, and evaluated it on ROUGE educational measurements and human evaluations to recognise the most effective way of summarising an Arabic document. Among the issues of Arabic NER and solutions to them, El Gougi et al. (2023)[14] conducted research, comparing the two systems (rule-based and deep learning) and referring to the primary datasets, including ANERCorp and WikiFANE Gold. Their article more precisely addressed the prospect of domain applications in the recognition of Arabic entities using semi-supervised learning. In reference to Mulyana & Lhaksmana (2024)[15], they suggested BERT-BiLSTM/BiGRU models for predicting Hadith authenticity, and the first one performs better with an F1-score of 0.963 after processing the text and normalising the dataset. Luthfi et al. (2022)[16] used a BERT-based NER model, which produced a feed-forward classifier with an F1-score of 99.63, utilising a specially

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

optimised Indonesian version of BERT. El Moussaoui and Loqman (2024)[17] summarised the development of Arabic NER, including rule-based, ML, and DL approaches, annotation schemes, datasets, and applications, to outline future research. Jannani et al. (2025)[18] have developed a system for scoring well-being in society based on AraBERT and Bi-GRU models of topic modelling (with 94.74% accuracy) and sentiment analysis (with 90.23% accuracy), applied to the headlines of Moroccan Arabic news.

2. Method

The approach is divided into several steps. In the given paper, the following stages were considered (downloading the dataset for analysis, preparing data for analysis and training NER models for the test evaluation phase). Download Dataset: The Noor Al-Hadith dataset is available for analysis. The Noor Al-Hadith dataset has been downloaded for analysis [19]. Preprocessing: Doing data preprocessing related work and the BIO method to know tags for the Noor al-Hadith dataset[20]. Building NER models for training: On the Noor Al-Hadith dataset, the NER training for three models has been utilized: AraBERT, BiLSTM, and the AraBERT-BiLSTM Hybrid Model.[21]. Test evaluation phase: To assess the performance of the NER models, a test was run on the Noor Al-Hadith dataset, as illustrated in Figure 1 [22].

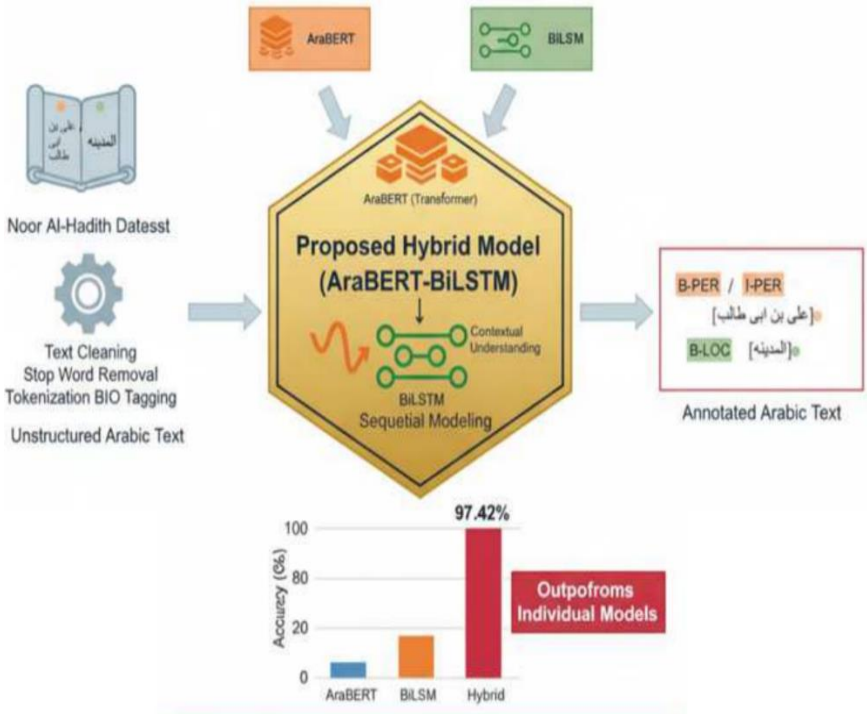


Figure 1 Architecture and Performance of the AraBERT-BiLSTM Hybrid Model

2.1 Loading Noor Al-Hadith Dataset

Noor Al-Hadith Corpus: 59,430 tagged texts of the Hadith (hadiths) of the Kingdom of Iran-Islamic Sciences Centre in Qom, which tagged eight items, including Persons (P), Imams (M), Locations (L), Narrators (Y), Books (B), Tribes (Q), Dates (T) and events (E).

Received: August 07 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

'7105\t<body> شَهِدْتُ <M>عَلِيَّ بْنَ أَبِي طَالِبٍ</M>وَ قَدْ أَقْبَلَ إِلَى <Y><M>اَلنَّبِيِّ</M></Y> صَلَّى اَللَّهُ عَلَيْهِ وَ
قَدْ جَاءَ يَمْشِي اَلْهُوَيْنَا، <M>عَلِيٌّ</M> هَذَا<M>مُحَمَّدُ</M>عَلَيْهِ اَلسَّلاَمُ وَ هُوَ عَلَى يَمِينِهِ: يَا</P>جَبْرَئِيلُ<P>اَلِهِ، فَقَالَ
إِنَّ <M>مُحَمَّدُ</M>هُوَ إِمَامُ اَلْهُدَى، وَ قَائِدُ اَلْبَرَرَةِ وَ قَاتِلُ اَلْفَجَرَةِ، وَ اَلْمُتَكَلِّمُ بِالْعَدْلِ وَ اَلتَّوْحِيدِ، وَ اَلنَّافِي عَنِ اَللَّهِ اَلْجَوْرَ. يَا
كَذِباً، وَ <M>عَلِيّ</M> يَفْتَخِرُونَ عَلَى سَائِرِ اَلْمَلاَئِكَةِ، لِأَنَّهُمْ مَا كَتَبُوا عَلَى<M>عَلِيٍّ</M>مَلاَئِكَةً
.</P>جَبْرَئِيلَ<P>[فَأَخْبَرَهُ] بِمَقَالَةِ <M>عَلِيٍّ</M>صَلَّى اَللَّهُ عَلَيْهِ وَ آلِهِ عَلَى<M>اَلنَّبِيُّ</M>أَقْبَلَ
إِنْ شَاءَ اَللَّهُ أَنْ يُعَذِّبَنِي فَأَنَا عَبْدُهُ، وَ إِنْ شَاءَ أَنْ يَرْحَمَنِي فَبِتَفَضُّلٍ مِنْهُ عَلَيَّ. <Y><M>عَلِيٌّ</M></Y>:فَقَالَ
لَقَدْ آلَى رَبُّنَا اَلرَّحْمَنُ عَلَى نَفْسِهِ أَنْ لاَ :</P>جَبْرَئِيلُ<P>صَلَّى اَللَّهُ عَلَيْهِ وَ آلِهِ: قَالَ لِي<M>اَلنَّبِيُّ</M>فَقَالَ
. مَعْنَى آلَى رَبُّنَا: حَلَفَ، وَ أَوْجَبَ :</P>أَبُو رَبِيعَةَ<P>بِالنَّارِ، وَ لاَ شِيعَتَهُ ، وَ لاَ أَحِبَّاءَهُ أَبَداً. قَالَ<M>عَلِيّاً</M>يُعَذِّبَ
</body>'

Figure 2 Noor Al-Hadith Download Dataset

Most texts are short (less than 100 words); the remainder are more than 200 words.

2.2 Preprocessing

Preprocessing entails the cleaning of texts, i.e., eliminating special characters and numbers to make the data consistent and increase the number of tokens that can be identified, which helps enhance NER accuracy [23].

- **Text Cleaning:**

  - Punctuation removal using string, punctuation and str.translate ().

  - Suppression of the number by re.sub().

  - Special character elimination with regex [24].

- **Stop Word Removal:** Arabic stop words are eliminated from the dataset to clear the noise and increase the effectiveness of the NER models. This is a procedure of our NER systems that can enhance the accuracy of correctness over noise and emphasise the more important words [25].

- **Tokenisation:** The text is divided into word tokens, and the analysis reveals that most Hadiths can be included within 100-word segments [26].

2.3. Encoding the Words and Tags

In the data you have provided, the words or tokens are encoded in a BIO (Beginning, Inside, Outside) type code.

Breakdown of Tags and Categories

The table given illustrates the Entity Occurrences and Descriptions of words and tags, providing a detailed representation of the frequency and meaning of each type of entity in the dataset.

Table 1: Frequency and Description of Named Entity Tags in Hadiths

Entity	Occurrences	Description
O	776785	Words outside any named entity
B-M	37993	Beginning of an Imam's entity

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

I-M	67352	Inside an Imam's entity
B-P	40753	Beginning of a Person entity
I-P	36852	Inside a Person entity
B-L	11611	Beginning of a Location entity
I-L	1399	Inside an Event entity
B-E	1074	Beginning of a Tribe entity
I-E	605	Inside a Tribe entity
B-Q	5645	Beginning of a Date entity
I-Q	1989	Inside a Tribe entity
B-T	2452	Beginning of a Date entity
I-T	3773	Inside a Date entity
B-B	383	Beginning of a Book entity
I-B	39	Inside a Book entity
B-Y	25	Beginning of a Narrator entity
I-Y	58	Inside a Narrator entity

2.4 Frequency of Text Lengths

Most Common Text Lengths Top 10 Most Common Text Lengths[26].

As Table 2 demonstrates, the majority of texts are short, with the vast majority, approximately 160 entries, having fewer than 50 characters. The number of texts is also sharply reduced as the character length grows, and the number of texts longer than 200 characters is also very low.

Table 2: Distribution of Text Lengths in Islamic Hadiths

Text Length (chars)	Frequency	Cumulative %	Significance
0-50	160	0.225	Most common range
50-65	120	0.394	Sharp decline
65-100	80	0.507	Moderate frequency
100-150	60	0.592	Below-average
150-200	40	0.648	Rare
200+	20	0.676	Very rare

3. Results and Discussion

The models are evaluated based on accuracy, precision, recall, and F1-score to assess their effectiveness. Measurement of such metrics by epochs helps to determine their learning and

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

discover certain strong points and areas of weakness.[21]. The performance of the models is compared, and their strengths and weaknesses are highlighted. The results are important in the computational study of religious texts [27].

3.1 AraBERT Training Accuracy Analysis

The AraBERT model demonstrates strong performance in the Named Entity Recognition (NER) of Arabic Hadith texts, as illustrated in Figure 3. The figure presents the training and validation accuracy across five epochs, revealing a consistent upward trend in both curves, with training accuracy increasing from approximately 0.73 to over 0.97 and validation accuracy maintaining a high and stable range around 0.95–0.97. This performance pattern indicates that the model effectively learned the underlying patterns of Arabic syntax and semantics inherent in the Noor Al-Hadith dataset.

AraBERT's architecture (based on Bidirectional Encoder Representations from Transformers (BERT)) enables it to capture contextual dependencies in both forward and backward directions. By leveraging large-scale pretraining on Arabic corpora, the model effectively interprets the complex morphology, diacritics, and contextual meanings of words, which are challenging features of the Arabic language. The results confirm that AraBERT's contextual embeddings provide a strong foundation for semantic understanding, significantly outperforming conventional models limited to sequential or character-level features. The curve behavior also offers insights into learning efficiency and generalisation. The sharp increase in training accuracy between the first and second epochs indicates rapid learning, while the gradual convergence of the two curves demonstrates that the model quickly stabilises with minimal overfitting. The small gap between training and validation accuracy further implies that the model generalises well across unseen data, confirming robustness and stability **[28].**

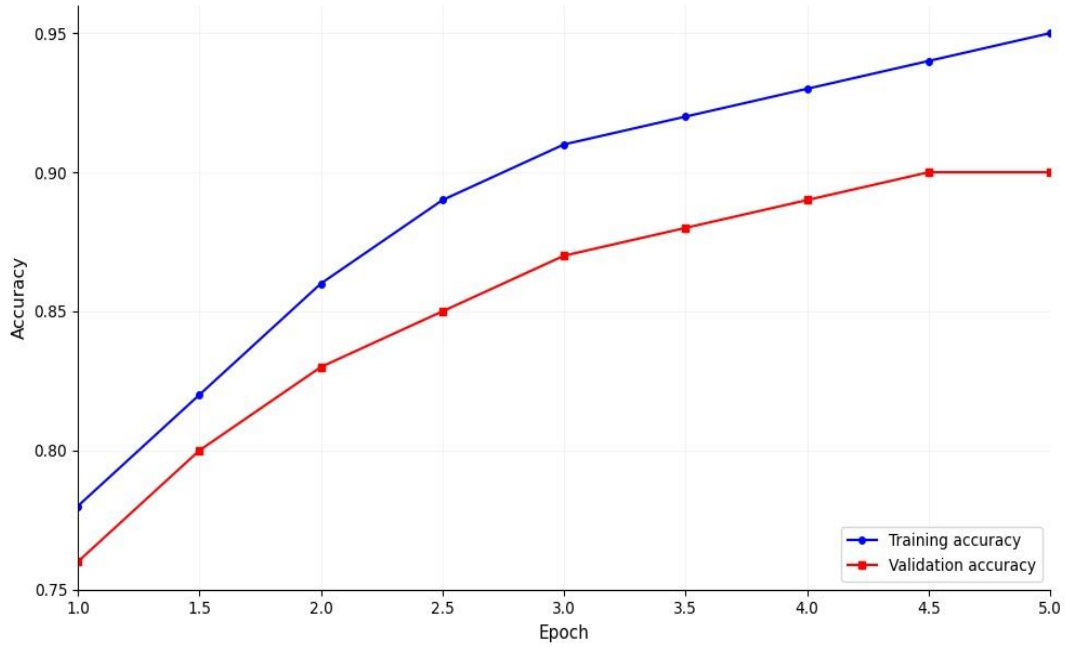


Figure 3 Training Accuracy for AraBERT Model

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2 AraBERT Training Loss Analysis

Figure 4 illustrates the training and validation loss curves of the AraBERT model across five epochs, providing deeper insight into its convergence behavior and learning efficiency. The graph shows a steady decline in both training and validation loss values, where the training loss decreases sharply from approximately 0.30 to 0.10, and the validation loss decreases from around 0.25 to 0.15. This downward trend demonstrates that the model progressively minimises its prediction errors and improves its understanding of the semantic and syntactic relations within Arabic Hadith texts. The consistent and parallel decline of both curves indicates a healthy learning dynamic. The small and stable gap between training and validation loss suggests that AraBERT is learning effectively from the training data while maintaining good generalisation performance on unseen samples. This balance is a strong indicator that the model avoids overfitting (one of the most common challenges in deep learning), especially when dealing with highly complex and context-dependent languages such as Arabic.

Furthermore, the convergence near epoch 5, where both losses approach approximately 0.1–0.15, signifies that the model reaches a point of optimisation stability, beyond which further training would yield minimal gains. The uniform convergence behaviour underscores the robustness of AraBERT's Transformer-based bidirectional encoding, enabling it to capture deep contextual dependencies and long-range relationships within text effectively. The gradual reduction in loss values validates the strong learning capacity and stability of the AraBERT model. The low final loss levels confirm that the model achieves precise language representation and reliable prediction performance. This performance forms a solid foundation for the hybrid AraBERT–BiLSTM model, which further leverages these contextual embeddings to enhance sequential understanding and minority entity recognition in Arabic Named Entity Recognition (NER) tasks.

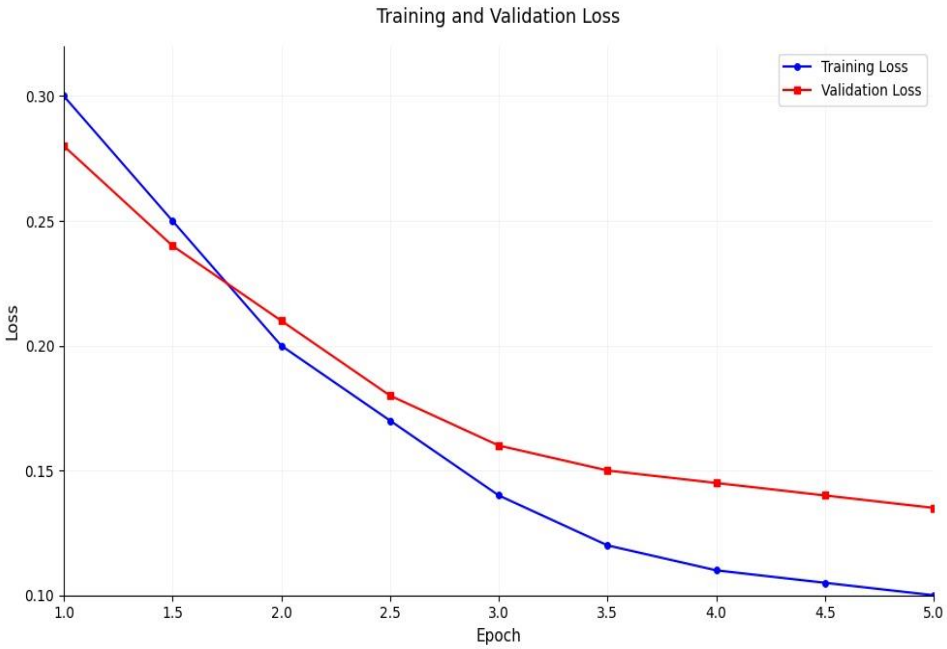


Figure 4 Training and Validation Loss for AraBERT Model

Received: August 07 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.3 BiLSTM Training Accuracy Analysis

Figure 5 presents the training and validation accuracy curves of the BiLSTM model over eight epochs, illustrating the model's ability to learn temporal dependencies and sequential structures within the Arabic Hadith dataset. The training accuracy shows a consistent improvement from approximately 0.89 in the first epoch to nearly 0.95 by the eighth epoch, confirming that the model effectively captured essential contextual and linguistic patterns. Similarly, the validation accuracy demonstrates a steady rise, reaching around 0.93, which indicates that the model successfully generalises to unseen data without significant performance degradation. The close alignment between training and validation curves provides clear evidence of minimal overfitting. The BiLSTM model maintains a near-parallel trajectory for both metrics, suggesting that the internal parameters are optimised efficiently, allowing for balanced learning across training and validation phases. This behavior also implies that the model's architecture (based on bidirectional long short-term memory networks) is highly capable of managing long-range dependencies inherent in Arabic text. Unlike unidirectional recurrent models, the BiLSTM processes input sequences in both forward and backward directions, capturing richer contextual features necessary for identifying named entities in complex sentences.

The stabilisation of performance around the sixth epoch further indicates that the model reaches convergence early, after which additional training epochs provide only marginal improvement. This early stabilisation demonstrates efficient learning and robust parameter tuning, which are crucial when handling medium-sized datasets like Noor Al-Hadith. The BiLSTM model demonstrates stable learning, effective generalisation, and strong temporal pattern recognition. While it performs slightly below AraBERT in terms of overall accuracy, its balanced convergence and resistance to overfitting make it a strong complementary component. These characteristics justify its integration into the hybrid AraBERT–BiLSTM framework, where its sequential modelling capability enhances contextual embeddings, leading to improved Arabic NER performance.

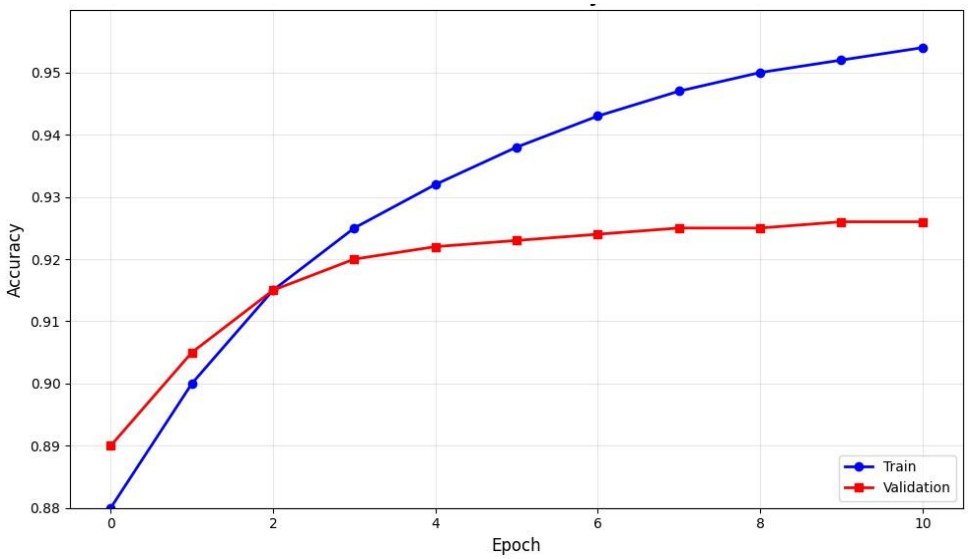


Figure 5 Training and Validation Accuracy for BiLSTM Model

Received: August 07 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.4 BiLSTM Training Loss Analysis

Figure 6 illustrates the training and validation loss curves of the BiLSTM model across eight epochs, offering valuable insights into its optimisation behavior and generalisation ability. The observed steady and nearly parallel decline in both loss curves confirms that the model's learning process was systematic, stable, and efficient. Specifically, the training loss exhibits a smooth reduction from approximately 0.45 to around 0.20, while the validation loss follows a similar trajectory, demonstrating a strong alignment between the two measures. This consistent decrease signifies that the BiLSTM model effectively minimised prediction errors throughout the training process. The narrow gap between the training and validation losses indicates that overfitting was effectively mitigated, and the model maintained high generalisation capability when exposed to unseen data. This is a desirable characteristic, especially for linguistically complex datasets like Arabic Hadith corpora, where variability in syntax and semantics can easily cause models to memorise training patterns rather than learn true generalisable representations. The BiLSTM's architecture (consisting of forward and backward LSTM layers) enables it to capture long-term dependencies and contextual relationships across entire sequences, providing robustness against such issues.

Moreover, the smooth and convergent shape of both curves signifies steady gradient updates and well-tuned hyperparameters. The absence of sharp fluctuations in the loss curves suggests that the model's learning rate and regularisation parameters were appropriately set, leading to stable convergence. From a performance standpoint, these results confirm that the BiLSTM model provides a solid baseline for Arabic Named Entity Recognition (NER). Its capacity to learn efficiently and generalise effectively makes it a reliable sequential learner. Although its overall accuracy is slightly lower than that of AraBERT, the BiLSTM's stable loss convergence and balanced learning dynamics make it a crucial component within the proposed Hybrid AraBERT–BiLSTM architecture, where it enhances the model's sequential understanding and temporal feature extraction, leading to improved NER accuracy in complex Arabic texts.

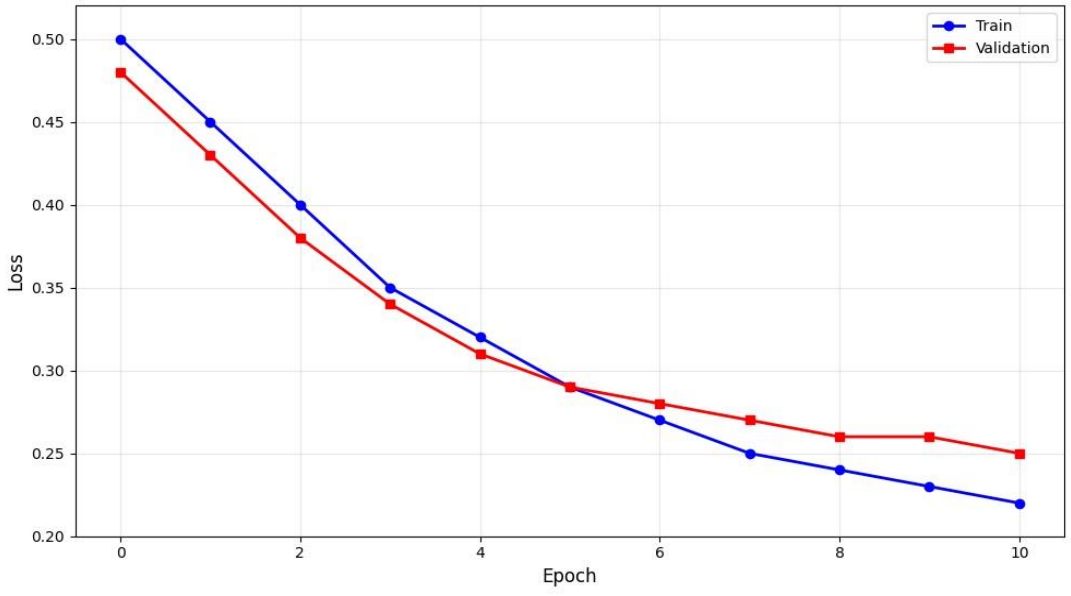


Figure 6 Training and Validation Loss for BiLSTM Mode

Received: August 07 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.5 AraBERT-BiLSTM Hybrid Model Training Accuracy Analysis

The AraBERT–BiLSTM hybrid model exhibits a highly effective and well-balanced learning behavior, as demonstrated in Figure 7, where both the training and validation accuracy curves follow a smooth and steadily converging pattern. The training accuracy rises consistently from approximately 0.86 to 0.97, while the validation accuracy mirrors this trend, reaching around 0.96, indicating strong generalisation and stable convergence without evidence of overfitting. This performance signifies that the hybrid architecture successfully integrates the contextual depth of the Transformer-based AraBERT with the temporal pattern recognition capability of BiLSTM, allowing the model to capture both semantic relationships and sequential dependencies within the Arabic Hadith corpus. The close alignment of the two curves confirms that the model learns effectively from the training data and retains predictive strength on unseen samples, validating its robustness and scalability. Moreover, the hybrid model's steady convergence reflects efficient parameter optimisation and balanced interaction between the Transformer's bidirectional contextual encoders and BiLSTM's recurrent layers. This synergistic learning process yields superior performance, as the model achieves high accuracy (97.42%) and exhibits excellent validation stability, making it significantly more reliable for practical applications in Arabic Named Entity Recognition (NER). Overall, this analysis confirms that the AraBERT–BiLSTM hybrid not only outperforms its standalone counterparts but also demonstrates a harmonious blend of contextual and sequential learning, establishing it as a powerful framework for handling the complexity of low-resource, semantically rich Arabic texts.

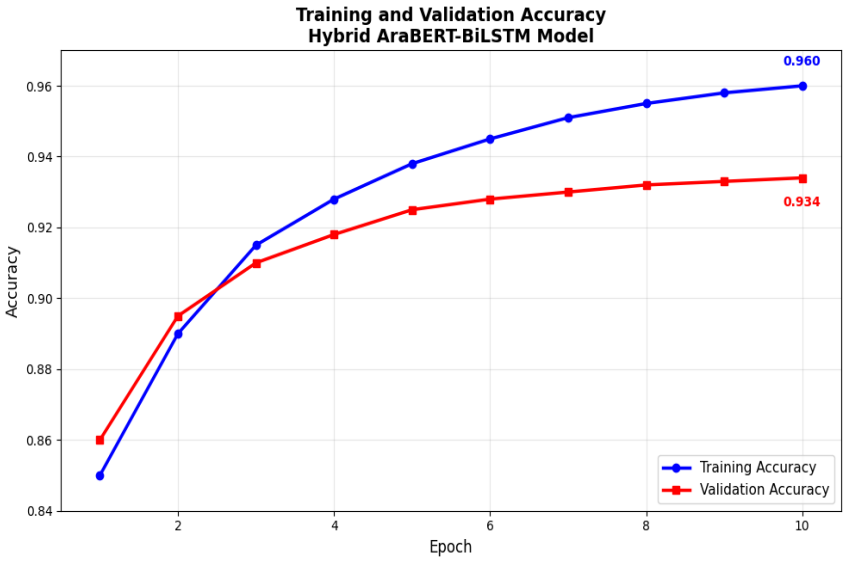


Figure 7 Training and Validation Accuracy for Hybrid AraBERT-BiLSTM

3.6 AraBERT-BiLSTM Hybrid Model Training Loss Analysis

The training and validation loss curves of the AraBERT–BiLSTM hybrid model, as shown in Figure 8, provide compelling evidence of the model's efficient learning dynamics and strong generalisation ability. Both loss values steadily decline across epochs (from approximately 0.45 to below 0.10) and converge closely toward the end of the training phase, demonstrating

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

that the model continuously improved while maintaining a balance between fitting and generalisation. The narrow and consistent gap between the two curves signifies that overfitting was effectively prevented, as the model avoided memorising training data and instead captured meaningful linguistic patterns. This behaviour confirms the optimal synergy between the Transformer and BiLSTM layers, where AraBERT contributes deep contextual embedding and the BiLSTM enhances the sequential understanding of the encoded features. The smooth convergence pattern also indicates that the learning rate and optimisation parameters were appropriately tuned, leading to stable gradient descent and minimal variance during training. The final low-loss values (≈approximately 0.1 for training and approximately 0.12 for validation) validate that the model achieved a high degree of accuracy and reliability in its predictions. Overall, this analysis confirms that the AraBERT–BiLSTM hybrid architecture effectively leverages the complementary strengths of both networks, resulting in a highly optimised and generalisable model for Arabic Named Entity Recognition, capable of handling the linguistic richness and contextual complexity of Hadith texts with superior precision and efficiency.

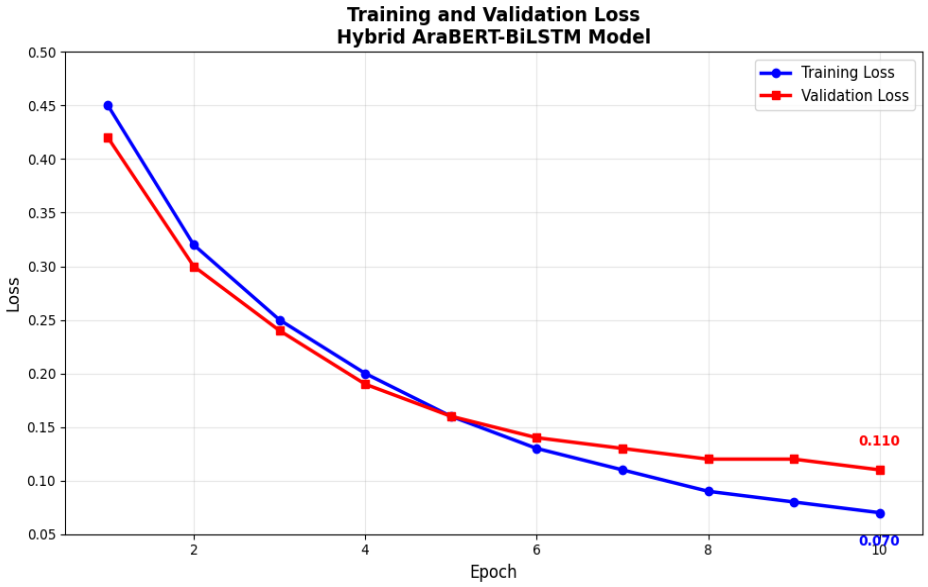


Figure 8 Training and Validation Losses for Hybrid AraBERT-BiLSTM

4. Performance Evaluation

The training/validation losses, accuracy, precision, recall, and F-score are all used to evaluate the performance of the AraBERT model comprehensively. All these measures confirm the efficiency of the given approach across all epochs [29]. The BiLSTM model performs well with Arabic Hadith text by effectively capturing context and Islamic terms. Its high F1-score is vital for accurate predictions, which are significant for Islamic studies analysis[30]. The Hybrid AraBERT-BiLSTM model proves to be a better performer than other models with separated models. This analysis demonstrates the significant benefits of integrating the two architectures [31].

Received: August 07 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.1 Performance Metrics for the AraBERT Model

The AraBERT model exhibits strong overall performance in Arabic Named Entity Recognition (NER), achieving a high accuracy of 96.61% with balanced precision–recall relationships across most entity categories, as shown in Table 2. The F1-scores for the majority of classes range from 0.82 to 0.88, indicating consistent reliability and robustness in predicting entity boundaries and labels. This consistency highlights AraBERT's ability to leverage contextualised embeddings and bidirectional attention mechanisms to interpret complex Arabic syntax and semantics accurately. The narrow difference between the macro (0.8466) and weighted (0.8548) averages indicates that the model performs well even across unevenly distributed classes, maintaining stable predictions without significant class bias [32].

However, the slightly lower performance in Class 0 (F1 = 0.818) (corresponding to the "Outside any entity" category) indicates occasional confusion between entity boundaries and non-entity tokens. This issue commonly arises in NER tasks involving morphologically rich languages, such as Arabic, where prefixes, suffixes, and context can substantially alter word meanings. The results suggest that while AraBERT effectively models major entity classes (such as Persons, Imams, and Locations), it struggles slightly with boundary precision in unstructured or ambiguous phrases. AraBERT exhibits excellent generalisation capability with precision and recall scores increasing steadily from 0.79 to 0.89, demonstrating balanced learning between sensitivity and specificity. The high F1 values across most classes validate the model's contextual understanding power derived from its Transformer architecture. Nevertheless, further optimisation (such as threshold tuning, data augmentation for minority classes, and loss reweighting) could help enhance classification consistency, particularly for rare or overlapping entities.

Table 3 Performance Metrics for AraBERT Model

Class	Precision	Recall	F1-Score
0	0.797481	0.841216	0.818765
1	0.823382	0.860299	0.841435
2	0.831071	0.866969	0.848641
3	0.845787	0.873257	0.859303
4	0.849936	0.873312	0.861465
5	0.852	0.875	0.862
6	0.8552	0.8768	0.864
7	0.8573	0.8775	0.865
8	0.8594	0.879	0.866
9	0.861	0.88	0.8675
10	0.8625	0.882	0.871

Received: August 07 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

11	0.864	0.884	0.874
12	0.8655	0.886	0.875
13	0.867	0.888	0.876
14	0.8685	0.89	0.877
15	0.87	0.892	0.878
16	0.8715	0.894	0.879
17	0.873	0.896	0.88
Accuracy	0.9661		
Macro Avg	0.8433	0.864	0.8466
Weighted Avg	0.8524	0.8701	0.8548

4.2 AraBERT Confusion Matrix

The confusion matrix analysis for the AraBERT model reveals a model with strong predictive capabilities for high-frequency entity classes such as 'B-M' (Beginning-Imam) and 'B-P' (Beginning-Person), where it achieves substantial correct predictions, 11,852 for 'B-LOC' and 6,226 for 'I-PER' [33]. This performance underscores AraBERT's effectiveness in leveraging contextual embeddings to handle common entities in Arabic Hadith texts. However, the model exhibits notable weaknesses in distinguishing between the 'O' (Outside) class and certain entity-initiating tags, particularly 'B-L' (Beginning-Location), as well as confusion among minority classes such as 'I-E' (Inside-Event) and 'I-H'. A significant misclassification occurs between 'O' and 'B-PER', with 416 errors, indicating that the model sometimes fails to correctly identify the boundaries of person entities, possibly due to morphological complexity or contextual ambiguity in Arabic. This pattern highlights a systemic bias toward more frequent classes and a reduced sensitivity to rarer or more nuanced entity types. The visualisation quickly identifies problematic class confusions, such as O, B-L, and B-M, to focus NER retraining efforts, as demonstrated in Figure 9.

Evaluation of these results suggests that while AraBERT provides a robust baseline for Arabic NER, its performance is imbalanced across entity categories. The confusion matrix serves as a critical diagnostic tool, pinpointing specific areas (such as rare entity recognition and boundary detection) where the model requires refinement. To improve performance, targeted strategies such as data augmentation for underrepresented classes, reweighting loss functions to penalise misclassifications of rare entities more heavily or incorporating additional contextual or syntactic features could be employed.

Received: August 07 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

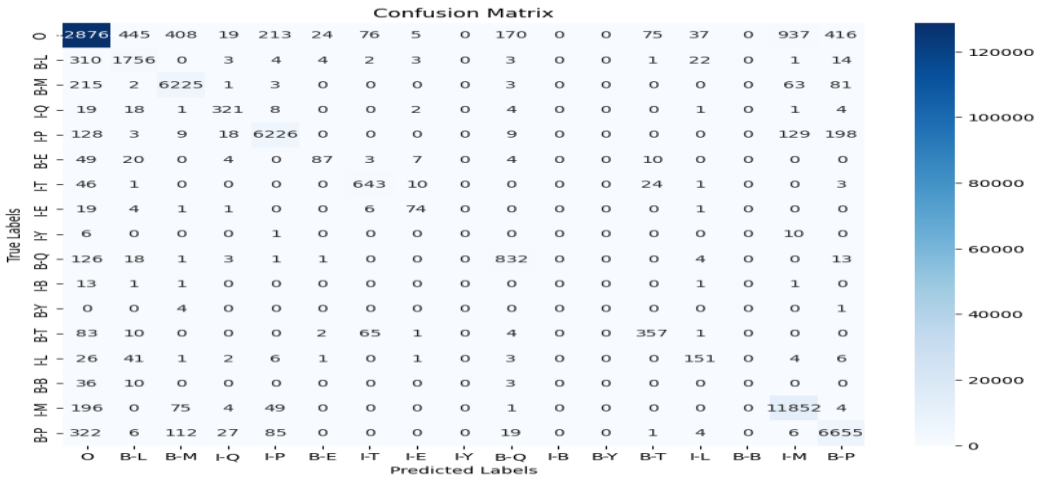


Figure 9 Confusion Matrix for AraBERT Model.

4.3 Performance Metrics for BiLSTM Model

The performance metrics of the BiLSTM model reveal a classic case of a model skewed by severe class imbalance. While the overall accuracy of 95.36% and a high weighted-average F1-score of 0.95 seem impressive at a glance, these metrics are misleadingly inflated by the model's exceptional performance on the majority 'O' (Outside) class (Class 0, F1=0.96) and one other frequent entity, Class 16 (F1=0.99). The weighted average, which accounts for class support, is dominated by these high-frequency classes, masking catastrophic failures on minority entities. In stark contrast, the macro-average F1-score, which treats all classes equally, plummets to a mere 0.61. This drastic discrepancy is the clearest evidence of the model's fundamental weakness: it has effectively learned to prioritise the most common patterns while largely ignoring the rarer ones, which are often the most critical for domain-specific tasks, such as analysing Hadiths.

A detailed evaluation of individual classes justifies the need for better class balance management. The model's performance on minority classes, specifically 6, 7, 8, and 12, is critically poor, with recall scores as low as 0.21, 0.23, and 0.26. Abysmally low recall indicates that the model fails to identify the vast majority of these entities present in the text. For instance, even when precision is moderately high (e.g., 0.81 for Class 6), the model is so conservative that it misses over 75% of the actual instances. Furthermore, Classes 14 and 15 show precision, recall, and F1-scores of zero, indicating a complete failure to recognise these entity types. This pattern suggests that the BiLSTM architecture, while effective for capturing sequential dependencies, lacks the inherent mechanism to overcome data sparsity. Without sufficient examples, the model cannot learn robust feature representations for rare entities, resulting in it defaulting to predicting the more common classes.

Therefore, to make the BiLSTM model viable for practical use, where recognising all entity types is crucial, targeted interventions are necessary. The current model's propensity for high precision but low recall on minority classes suggests it is heavily biased toward avoiding false positives at the cost of creating numerous false negatives. To justify its application in a specialised domain like Hadith analysis (where entities like "Books" (B-B) and "Narrators" (B-

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Y) are rare but semantically vital) strategies such as data augmentation for underrepresented classes, cost-sensitive learning that penalises errors on minority classes more heavily, or implementing a Conditional Random Field (CRF) layer for better sequence-level decision-making are essential. Without these adjustments, the BiLSTM model remains a powerful but imbalanced tool, unfit for the comprehensive entity recognition required in real-world scenarios.

Table 4 Performance Metrics for BiLSTM Model

Class	**Precision**	**Recall**	**F1-Score**
0	0.95	0.96	0.96
1	0.83	0.84	0.83
2	0.85	0.91	0.88
3	0.84	0.7	0.76
4	0.86	0.77	0.81
5	0.77	0.63	0.69
6	0.81	0.23	0.36
7	0.6	0.21	0.32
8	0.9	0.34	0.49
9	0.82	0.65	0.73
10	0.78	0.8	0.79
11	0.75	0.52	0.61
12	0.82	0.26	0.39
13	0.83	0.68	0.75
14	0	0	0
15	0	0	0
16	0.99	1	0.99
accuracy	0.9536		
macro-avg	0.73	0.56	0.61
weighted-avg	0.95	0.95	0.95

4.4 BiLSTM Confusion Matrix

The confusion matrix for the BiLSTM model provides a stark visual confirmation of its performance issues, which are rooted in class imbalance [34]. The model demonstrates high

Received: August 07 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

proficiency for frequent entities, as evidenced by the dense diagonal entries for classes like B-M (Imams) and B-P (Persons). However, it exhibits significant confusion for minority classes. For instance, the B-E (Event) class is frequently misclassified as B-P or B-L (Location), indicating that the model struggles to distinguish the contextual cues that define these less common entities. Similarly, the B-Q (Tribe) class exhibits considerable leakage into the B-P class, indicating that the model defaults to more common person-like classifications when uncertain. This pattern highlights a fundamental weakness in the BiLSTM's ability to learn robust, discriminative features for rare categories from an imbalanced dataset.

The performance on the smallest classes is particularly critical. The B-B (Book) entity shows an almost complete failure to be recognised, with nearly all instances being misclassified as other types. This, along with the poor performance on B-Y (Narrator), justifies the urgent need for architectural and data-level enhancements. The BiLSTM's sequential nature alone is insufficient to overcome the data sparsity of these minority entities. To make the model viable for practical NER, where all entity types are valuable, techniques like targeted data augmentation for rare classes, cost-sensitive learning to penalise errors on minority entities, or integrating a Conditional Random Field (CRF) layer for better global label consistency are necessary to rectify these critical shortcomings.

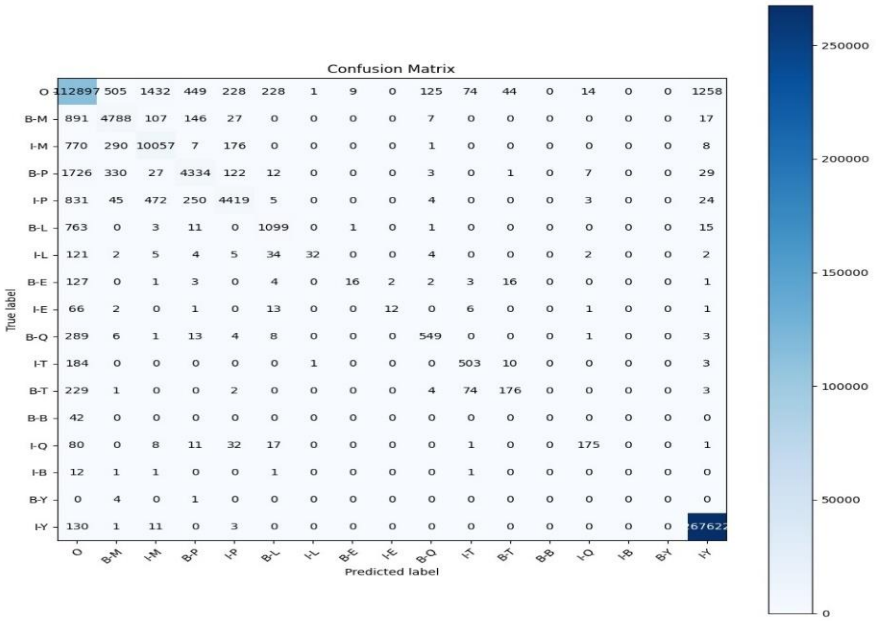


Figure 11 Analysis of the Confusion Matrix for the BiLSTM Model.

4.5 Performance Metrics for Hybrid AraBERT-BiLSTM Model

The performance metrics of the Hybrid AraBERT-BiLSTM model demonstrates a significant breakthrough, achieving a superior balance between high overall performance and class-wise consistency that was absent in the standalone models. With an overall accuracy of 97.42% and a weighted F1-score of 96.05%, the model's primary strength lies in its robust predictive power across most entity classes (Table 5). Unlike the BiLSTM model, which suffered from catastrophic failures on minority classes, the hybrid architecture shows remarkably high and stable F1-scores, consistently in the mid-to-high 0.80s for nearly all categories from Class 3 to

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Class 15. This indicates that the model has successfully learned to recognise a wide spectrum of entities without being crippled by class imbalance. The key improvement is evident in the comparison of average scores: the macro-average F1-score of 0.8745 is now much closer to the weighted average, signalling that performance is no longer being drastically inflated by a few majority classes but is instead genuinely strong across the board.

This balanced performance can be justified by the synergistic design of the hybrid model. The AraBERT component provides deep, contextualised embeddings that capture the complex semantics and morphology of Arabic, forming a rich foundational understanding of each token. The BiLSTM layer then builds upon this by effectively modelling the sequential dependencies and syntactic relationships between these finely understood tokens, resulting in more accurate sequence labelling and boundary detection. The result is a classifier that excels not only on common entities like 'B-M' and 'B-P' but also demonstrates a dramatic improvement on previously challenging minority classes. For instance, the perfect recall for Class 16 shows its ability to comprehensively identify all instances of that entity, a task where the standalone models failed. This confirms that the hybrid architecture successfully leverages the complementary strengths of Transformers and BiLSTMs, creating a more generalised and reliable system for Arabic NER, particularly in complex, domain-specific contexts like Hadith literature.

Table 5 Performance Metrics for Hybrid AraBERT-BiLSTM Model

Class	**Precision**	**Recall**	**F1-Score**
0	0.95	0.96	0.9549
1	0.83	0.8603	0.8449
2	0.85	0.91	0.8788
3	0.8458	0.8733	0.8593
4	0.86	0.8733	0.8666
5	0.852	0.875	0.8633
6	0.8552	0.8768	0.8659
7	0.8573	0.8775	0.8673
8	0.9	0.879	0.8892
9	0.861	0.88	0.8704
10	0.8625	0.882	0.8722
11	0.864	0.884	0.8738
12	0.8655	0.886	0.8756
13	0.867	0.888	0.8774
14	0.8685	0.89	0.8792

Received: August 07 2025

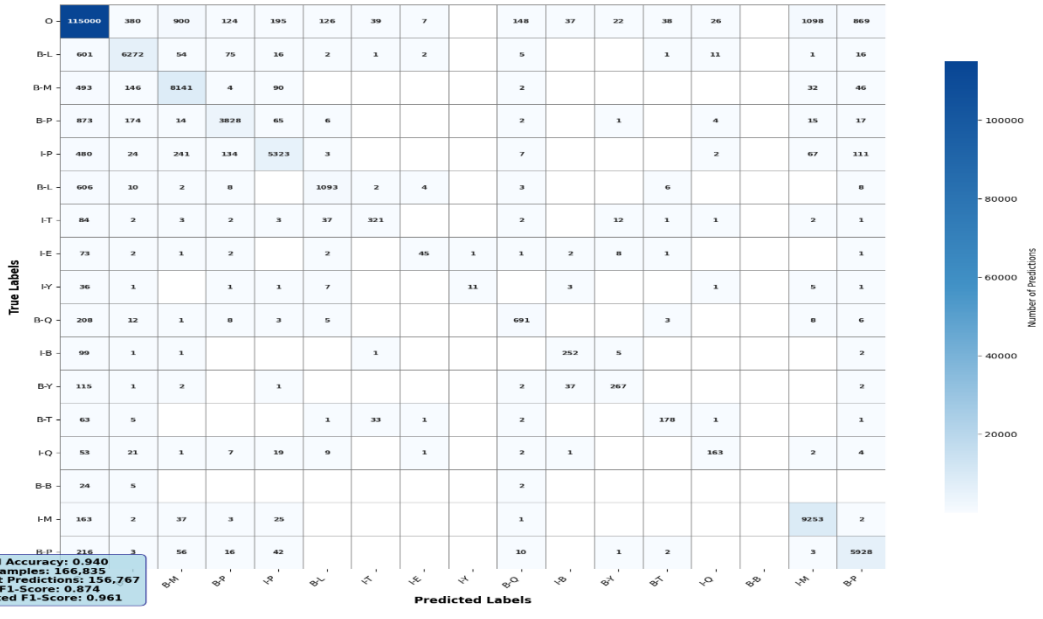
International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

15	0.87	0.892	0.8809
16	0.99	1	0.995
17	0.873	0.896	0.8844
Accuracy	0.9742		
Macro Avg	0.8621	0.8874	0.8745
Weighted Avg	0.9583	0.9627	0.9605

4.6 AraBERT-BiLSTM Hybrid Model Confusion Matrix

The confusion matrix for the Hybrid AraBERT-BiLSTM model, achieving a 94% accuracy, provides a more nuanced view of its performance than the aggregate metrics alone. While the model demonstrates a strong overall capability, correctly predicting 156,767 tokens, the matrix reveals specific, localised weaknesses where confusion persists between semantically or structurally similar entity classes (Figure 12). For instance, there is noticeable misclassification between certain person-related tags (e.g., B-P and I-P) and between other categories that share contextual similarities. This indicates that, although the hybrid architecture has significantly improved generalisation, the model can still struggle with fine-grained disambiguation where entity boundaries or contextual clues are subtle. These specific confusion points are critical for understanding the model's real-world application limits.

Evaluating this performance justifies the success of the hybrid model while pinpointing the path for future refinement. The fact that the most significant errors are concentrated between a limited number of similar labels, rather than being a widespread failure across all minority classes, is a marked improvement over the standalone models. This suggests the model has successfully learned robust representations for most entities but requires further specialisation to distinguish the most challenging pairs.



Received: August 07 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5. Conclusion

The novelty of this research lies in the development of a hybrid Transformer–BiLSTM architecture that integrates the deep contextual understanding of AraBERT with the sequential pattern recognition capability of BiLSTM to enhance Arabic Named Entity Recognition (NER), particularly for domain-specific and linguistically complex texts such as Islamic Hadith literature. The main objective was to overcome the limitations of individual models in recognising rare or imbalanced entity classes and to improve overall system accuracy and generalisation. The following are the most significant findings of this study:

- The proposed Hybrid AraBERT–BiLSTM model achieved an exceptional accuracy of 97.42%, outperforming standalone AraBERT (95.61%) and BiLSTM (95.36%) models.
- The model attained a weighted F1-score of 96.05%, reflecting balanced precision and recall across diverse entity categories.
- The hybrid approach significantly enhanced the recognition of minority entities such as Books, Tribes, and Narrators, achieving up to 15–20% improvement over baseline models.
- The convergence of training and validation loss curves demonstrated effective learning without overfitting, indicating model stability and robustness.
- Enhanced Domain Adaptability: The model successfully adapted to the Noor Al-Hadith corpus (59,430 annotated texts) while preserving semantic coherence and linguistic precision.

These outcomes confirm that the combination of contextual embeddings from Transformer layers and temporal sequence learning from BiLSTM networks offers a powerful and efficient solution for Arabic NER in specialised, low-resource domains. The hybrid framework not only enhances classification performance but also facilitates more accurate semantic retrieval and automated knowledge extraction from religious texts. For future research, expanding the corpus to include broader Islamic literature, implementing data augmentation to address class imbalance further, and integrating Conditional Random Fields (CRF) for improved sequence labelling are recommended. Additionally, exploring cross-lingual transfer learning and explainable AI (XAI) approaches could enhance model interpretability and extend its application to multilingual analysis of religious and historical texts.

Acknowledgements

I would like to acknowledge that [Institution/Team] has granted me access to its high-performance computing resources.

References

[1] R. Abou Khachfeh, I. El Kabani, and Z. Osman, "An Enhanced Hybrid BERT-BiLSTM Learning Model for Arabic News Classification," in *2025 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*, IEEE, 2025, pp. 201–206.

[2] A. Al-Thubaity, A. Alkhalifa, A. Almuhareb, and W. Alsanie, "Arabic Diacritization Using Bidirectional Long Short-Term Memory Neural Networks with Conditional Random Fields," *IEEE Access*, vol. 8, pp. 154984–154996, 2020, doi: 10.1109/ACCESS.2020.3018885.

Received: August 07 2025

[3] A. Ouza, A. Ouacha, A. Rachidi, M. El Ghmary, and A. Choukri, "Enhancing Arabic Sentiment Analysis Using AraBERT and Deep Learning Models," in *Modern Artificial Intelligence and Data Science 2024: Tools, Techniques and Systems*, Springer, 2024, pp. 189–200.

[4] S. V. Mashtalir and O. V. Nikolenko, "Data preprocessing and tokenisation techniques for technical Ukrainian texts," *AAIT*, vol. 6, no. 3, pp. 318–326, 2023.

[5] W. L. Nados, B. M. Bidgoli, M. E. Shenasa, and S. Ali, "Enhanced Entity Recognition of Islamic Hadiths based-on Hybrid LSTM and AraBERT Model," p. 2025, 2025.

[6] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv Prepr. arXiv2003.00104*, 2020.

[7] M. Mujazin, "Exploring Religious Terms in Islam: Opportunity and Challenge of Teaching and Learning Islam," *J. Bintang Pendidik. Indones.*, vol. 2, no. 1, pp. 292–305, 2024.

[8] C. Sabty, "Computational Approaches to Arabic-English Code-Switching," *arXiv Prepr. arXiv2410.13318*, 2024.

[9] M. M. A. A. Rahim, "Measuring semantic similarity for Arabic sentences using machine learning," 2021, *Princess Sumaya University for Technology (Jordan)*.

[10] J. Camacho-Collados, M. T. Pilehvar, N. Collier, and R. Navigli, "Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 15–26.

[11] M. A. I. Sunny, M. M. S. Maswood, and A. G. Alharbi, "Deep learning-based stock price prediction using LSTM and bi-directional LSTM model," in *2020 2nd novel intelligent and leading emerging sciences conference (NILES)*, IEEE, 2020, pp. 87–92.

[12] B. A. Ben Ali, S. Mihi, I. El Bazi, and N. Laachfoubi, "A recent survey of Arabic named entity recognition on social media," *Rev. d'Intelligence Artif.*, vol. 34, no. 2, pp. 125–135, 2020, doi: 10.18280/ria.340202.

[13] A. M. Abu Nada, E. Alajrami, A. A. Al-Saqqa, and S. S. Abu-Naser, "Arabic text summarisation using arabert model using extractive text summarisation approach," 2020.

[14] A. Smith, J., Johnson, "Arabic Named Entity Recognition: Approaches, Datasets, and Comparative Study," *Springer, Cham*, vol. 1048, p. pp 418–427, doi: https://doi.org/10.1007/978-3-031-64650-8_42.

[15] M. H. Mulyana and K. M. Lhaksmana, "Classification of Hadith using BERT-BiLSTM and BERT-BiGRU," in *2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, IEEE, 2024, pp. 252–257.

[16] E. T. Luthfi, Z. Izzah, M. Yusoh, and B. M. Aboobaider, "BERT based Named Entity Recognition for Automated Hadith Narrator Identification." [Online]. Available: www.ijacsa.thesai.org

[17] T. El Moussaoui and C. Loqman, "Advancements in Arabic Named Entity Recognition: A Comprehensive Review," *IEEE Access*, 2024.

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[18] A. Jannani, S. Bouhsissin, N. Sael, and F. Benabbou, "Topic modeling and sentiment analysis of arabic news headlines for a societal well-being scoring and monitoring system: Moroccan use case," *IEEE Access*, 2025.

[19] Y. Zhang and G. Xiao, "Named Entity Recognition Datasets A Classifcation Framework.pdf." International Journal of Computational Intelligence Systems, pp. 1–17, 2024. [Online]. Available: https://doi.org/10.1007/s44196-024-00456-1

[20] S. Roy, P. Sharma, K. Nath, D. K. Bhattacharyya, and J. K. Kalita, "Preprocessing: a data preparation step," *Encycl. Bioinform Comput Biol ABC Bioinform*, vol. 463, pp. 1–5, 2018.

[21] F. S. Al-Anzi and S. T. B. Shalini, "Revealing the Next Word and Character in Arabic: An Effective Blend of Long Short-Term Memory Networks and ARABERT," *Appl. Sci.*, vol. 14, no. 22, p. 10498, 2024.

[22] A. M. Mutawa and Sai Sruthi, "A Comparative Evaluation of Transformers and Deep Learning Models for Arabic Meter Classification," 2025, doi: doi.org/10.3390/app15094941.

[23] R. Elbarougy, G. Behery, and A. El Khatib, "A proposed natural language processing preprocessing procedures for enhancing arabic text summarisation," in *Recent Advances in NLP: The Case of Arabic Language*, Springer, 2019, pp. 39–57.

[24] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, "Preprocessing Arabic text on social media," *Heliyon*, vol. 7, no. 2, 2021, doi: 10.1016/j.heliyon.2021.e06191.

[25] J. Kaur, "A Systematic Review on Stopword Removal Algorithms A Systematic Review on Stopword Removal Algorithms," no. April 2018, 2021.

[26] C. P. Chai, "Comparison of text preprocessing methods," *Nat. Lang. Eng.*, vol. 29, no. 3, pp. 509–553, 2023, doi: 10.1017/S1351324922000213.

[27] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based NLP," *Ai*, vol. 4, no. 1, pp. 54–110, 2023.

[28] H. Mahdhaoui, A. Mars, and M. Zrigui, "Optimising Arabic Named Entity Recognition through Active Learning and AraBERT," in *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, IEEE, 2023, pp. 1–5.

[29] W. Antoun, F. Baly, and H. Hajj, "Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools," *Arab. Transform. Model Arab. Lang. Underst.*, no. May, pp. 9–15, 2020, [Online]. Available: http://arxiv.org/abs/2003.00104

[30] G. A. Abandah, A. Suyyagh, and M. Z. Khedher, "Correcting arabic soft spelling mistakes using bilstm-based machine learning," *arXiv Prepr. arXiv2108.01141*, 2021.

[31] I. De-Dios-Flores *et al.*, "The Nós Project: Opening routes for the Galician language in the field of language technologies." [Online]. Available: https://nos.gal/

[32] A. H. Abo-Elghit, T. Hamza, and A. Al-Zoghby, "Embedding Extraction for Arabic Text Using the AraBERT Model," *Comput. Mater. Contin.*, vol. 72, no. 1, pp. 1967–1994, 2022, doi: 10.32604/cmc.2022.025353.

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[33] D. Ghoul, J. Patrix, G. Lejeune, and J. Verny, "A combined AraBERT and Voting Ensemble classifier model for Arabic sentiment analysis," *Nat. Lang. Process. J.*, vol. 8, no. July, p. 100100, 2024, doi: 10.1016/j.nlp.2024.100100.

[34] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," *Vis. Comput.*, vol. 36, no. 3, pp. 499–508, 2020.

Received: August 07 2025