Journal of Applied Artificial Intelligence
https://jaai.sabapub.com
ISSN: 2709-5908

2026 Volume 7, Issue 1: 103–120

DOI: https://doi.org/10.48185/jaai.v7i1.1897

A Hybrid CNN-BiLSTM Framework with Attention-Based Explainability for Interpretable Fake News Detection

Mosimabale Agbabiaka[1, *], Emeka Ogbuju[2], Francisca Oladipo[3]

[1]Department of Computer Science, Faculty of Sciences, Federal University Lokoja, Lokoja, 263101, Nigeria.

[2]Department of Computer Science, School of Computing, Miva Open University, Abuja, Nigeria
[3]Department of Computer Science, School of Computing and Applied Sciences, Thomas Adewunmi University, Oko, Nigeria.

Received: 21.11.2025 • Accepted: 29.03.2026 • Published: 04.06.2026 • Final Version: 30.06.2026

Abstract: The rapid spread of fake news on social media has far-reaching implications, affecting various aspects of national life, including political stability, governance, economic systems, public health, and education. These platforms not only facilitate information exchange but also shape global public opinion. Effective fake news detection on social media is therefore essential to protect democratic processes, maintain public trust in institutions, prevent election manipulation, mitigate social polarization and violence, reduce the spread of health misinformation during crises (such as pandemics), curb financial fraud, and preserve the integrity of public discourse. Despite the ongoing efforts in combating fake news, most existing solutions remain limited in scope and interpretability. This study evaluates deep learning (DL) models, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and a hybrid DL model (CNN-BiLSTM with attention mechanism) to improve both robustness and prediction interpretability. The proposed model (CNN-BiLSTM+Attn) was evaluated on a Nigerian social media news dataset (FN_data, 126,974 records after deduplication and preprocessing) across five independent training runs with different random seeds (42 - 46), yielding a mean accuracy of 82.86% ± 0.22% and mean F1-score of 77.12% ± 0.52% as the best performing model, with statistically significant improvement over the other models ($p < 0.05$). The proposed model was further validated on a standard benchmark dataset (The ISOT fake news dataset), achieving mean accuracy of 99.33% ± 0.10% and mean F1-score of 99.39% ± 0.09% over ten independent runs (seeds 42 - 51). This improvement on ISOT confirms the robustness of the proposed model. Attention visualizations provide token-level explainability, highlighting the model's focus on deceptive cues. This work provides a transparent and interpretable framework for fake news detection in emerging digital contexts such as Nigeria.

Keywords: Fake News, Deep Learning, Hybrid Model, Attention Mechanisms, Social Media

* Corresponding Author: mosimabale.agbabiaka@fulokoja.edu.ng

1. Introduction

News has long been essential to human life, delivering information through media such as newspapers, television, and radio. However, the rise of the internet has shifted news dissemination toward online and social media platforms, which offer faster, cheaper, and more convenient access. These changes have also enabled the rapid spread of fake news. The internet and social media are often described as the lifeblood of fake news [1]. Fake news spreads more quickly and widely than true news, reportedly traveling six times faster [2]. It is created with different motives, such as political, economic, or defamation purposes, and poses threat to nations and individuals alike [3]. For instance, during the first three months of the coronavirus outbreak, over 800 deaths and approximately 5,800 hospitalizations were attributed to fake news [4]. In Nigeria, fake news has aggravated the herder-farmer crisis, resulting in loss of lives and properties [5].

This study proposes an explainable hybrid CNN-BiLSTM model enhanced with attention mechanisms. Attention mechanisms have been used in deep learning for various purposes, including machine translation [6], resource allocation, and text classification [7]. In this work, the attention layer assigns weights to input tokens and reveals their contribution to predictions, thereby improving interpretability. This approach is driven by the need to develop a fake news classification model that is both efficient and explainable.

2. Related Works

Fake news is not a new phenomenon; it has long been a part of human communication and a staple of media history, even before the advent of social media. "The Great Moon Hoax" of 1835, which falsely reported the discovery of life on the moon [8], the "Yellow Journalism" that led to the Spanish-American War in 1898 [9]. The advent of internet technology and social media in the late 20th century has had a multifaceted impact on the dissemination of fake news, thereby amplifying the attendant risks. Contemporary examples include computational propaganda, facilitated by technological advancements that enable the mimicry of authentic news websites, the manipulation of multimedia content, and the creation of counterfeit duplicates [8]. Fake news remains a global challenge, it constitutes more than one-third of trending events on Chinese microblogs, and over one million tweets were linked to fake news during the 2016 US Presidential election [2].

Understanding the landscape of fake news detection requires a comprehensive examination of existing literature and methodologies to grasp its complexities.

Research in fake news detection spans machine learning and deep learning paradigms. Early efforts, such as those in [10], compared both machine and deep learning approaches on three different datasets. Their results show that BiLSTM consistently outperformed traditional classifiers in accuracy.

Hybrid architecture has proven especially effective. In an experiment by [11] , a hybrid of CNN and BiLSTM achieved 88.78% accuracy, surpassing standalone CNN, LSTM, and BiLSTM models. similarly, [12] evaluated LSTM and CNN models on two fake news datasets from Kaggle. LSTM achieved 77.16% and 73.01% accuracy before optimization, and 89.96%–86.48% after optimization, while CNN achieved 69.38% and 93.85% accuracy before optimization, and 93.20%–95.16% after optimization. Also, [13] explored optimization techniques using FastText embedding, achieving 98% accuracy for their proposed model.

[14] investigated CNN, BiLSTM, and ResNet variants; an augmentation technique was used to address data imbalance. Different embedding techniques (GloVe, FastText, and Word2Vec) were compared, with BiLSTM achieving the best results.

Furthermore, [15] proposed a hybrid neural network model that combines CNN and RNN, leveraging CNN's extraction of local features and LSTM's sequential dependencies to enhance performance, achieving 60% and 99% accuracy on the FA-KES and ISOT datasets, respectively.

[16] compared a hybrid approach and a non-hybrid approach using RNN-LSTM and CNN architectures, with accuracies of 93.41% and 85.16% respectively on the BuzzFeed dataset. On a larger dataset (ISOT), RNN-LSTM and CNN achieved 99.90% and 98.02% accuracy respectively, highlighting the hybrid model's superior performance compared to non-hybrid models.

[17] presented a framework for fake news detection using the Long-Short Term Memory (LSTM) model. They employed datasets from Kaggle and Nigerian dailies. Synthetic Minority Oversampling Techniques (SMOTE) was used to balance the locally acquired dataset. The model achieved 92.86% accuracy on the balanced dataset. [18] proposed a Machine-Human (MH) system integrating linguistic and network analysis for news prediction, though constrained by the absence of standardized Nigerian datasets. [19] contributed a novel dataset from Twitter and Facebook (751,876 entries). They implemented CNN and RNN models on the datasets, achieving accuracies of 81.96% and 82.34% respectively.

[20] proposed an attention-based LSTM and BiLSTM models, which recorded 97.66% accuracy, highlighting the impact of integrating advanced deep learning approaches in fake news detection. (Hung Vo et al., 2025;Mahmud et al., 2024) integrated NLP and deep learning using different architectures, with CNN-BiLSTM achieving better results. [23] utilized ISOT and FakeNewsNet for implementing the LSTM model, which achieved 99.95% and 98.64% accuracy respectively. [24] implemented different deep learning architectures like LSTM, CNN, BiLSTM, ResNet, COBRA, and GRU for fake news detection using the ISOT dataset. LSTM, BiLSTM, GRU, and CORBRA achieved 97% accuracy, while CNN and ResNet achieved 96% and 76%, respectively.

Despite these advances, several limitations persist, including performance variability due to small or imbalanced training sets [17], [18], suboptimal model generalization, and a lack of interpretability in predictions. These limitations often result in overfitting and poor performance on unseen data.

Detecting fake news on Nigerian social media poses additional challenges due to the variety of linguistic, socio-political, and cultural contexts, which make feature extraction more difficult.

To address these issues, this research introduces a hybrid CNN-BiLSTM model enhanced with an attention mechanism. The preprocessing steps include removing domain-specific stop words and expanding contractions for both English and non-English words. GloVe embedding is used for semantic representation, and SMOTE is applied to handle class imbalance. The attention layer assigns weights to each token based on its importance, which is visualized in bar charts, facilitating both precise classification and interpretable insights into predictions.

3. Methodology

The development of a fake news detection model requires a comprehensive methodological framework that encompasses multiple stages and processes, including dataset acquisition, data preprocessing, feature extraction, component selection, model building, evaluation, and prediction (as illustrated in Figure 1). This section elaborates on the steps and approaches used to develop the proposed model.

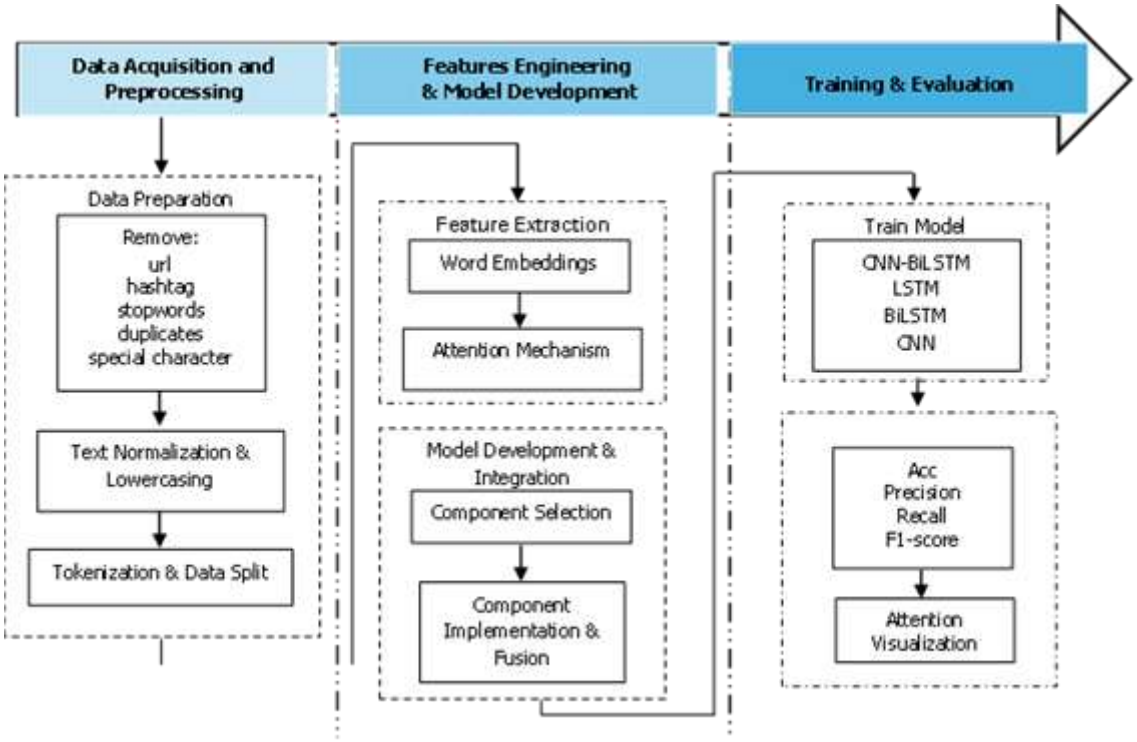

Figure 1. Architectural Design of the Hybrid Fake News Detection Model

3.1. Datasets

The primary dataset (FN_data) consists of 126,974 social media news records after deduplication and preprocessing, with a class distribution of 42,400 real news (labeled 1) and 84,574 fake news (labeled 0). It focuses on news about events like the EndSARS protests and the Fulani herdsmen attacks in Nigeria, extracted from Twitter and Facebook. The 'text' column encompasses news content, while the 'label' column indicates its authenticity.

The experiment was also validated on a standard and well distributed benchmark dataset (ISOT dataset), which contains news articles that focused on Government and Politics, made up of 23,8481 fake news denoted by 0 and 21,417 real news denoted by 1.

3.2. Data Preprocessing

Raw text was tokenized using Keras Tokenizer (num_words=5000), converted to sequences, and padded/truncated to a maximum length of 100 tokens. The cleaned dataset was split into training and validation sets using 80/20 stratified split with random_state=42 for both FN_data and ISOT fake news dataset to ensure reproducibility.

A series of preprocessing steps was applied to the dataset, including:

- Defining a custom list of stop words that do not add significant value to the meaning of the dataset. Examples are dey, na, wetin, dat, de, u, etc [25]. This enables the model to focus on the most essential features in the dataset.
- Expanding contractions in the text to reduce complexity in the dataset by installing and importing the contractions library. We also use custom expansion for non-English contraction.
- Removing stopwords, URLs, hashtags, mentions, and special characters.
- Eliminating duplicate records
- Converting all text to lowercase
- Dropping null values

- Tokenizing and applying lemmatization to reduce words to their base.

Duplicate values were eliminated to prevent biased models with poor generality.

Figures 2 and 3 show word clouds of the most frequent terms in fake and real news after preprocessing for FN_data and ISOT, respectively.

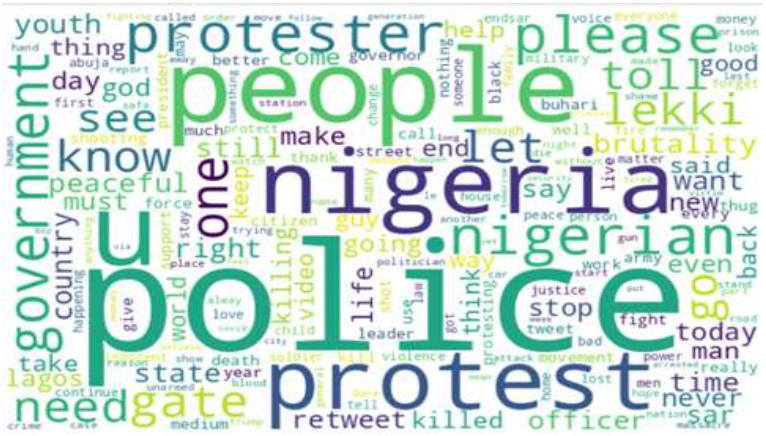

Figure 2. Word Cloud for Fake and Real News: FN_data dataset

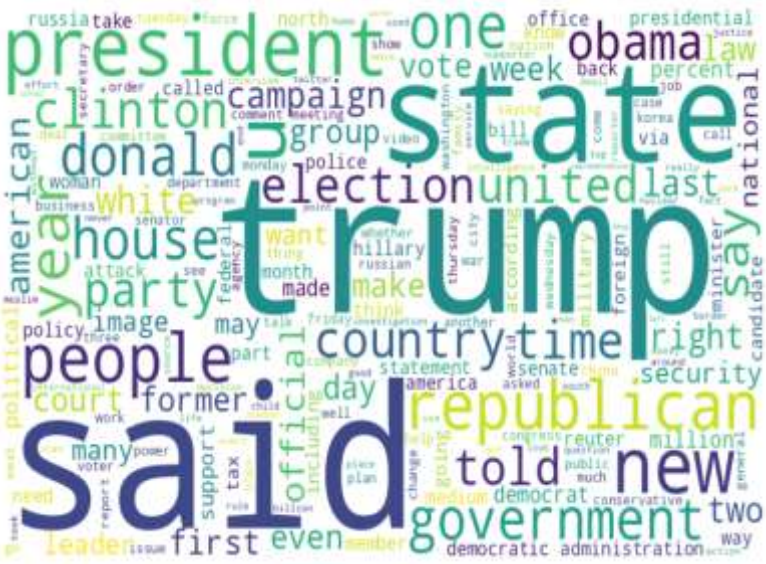

Figure 3. Word Cloud for Fake and Real News: ISOT dataset

3.3. Word Embedding

Word representations were generated using pre-trained GloVe (Global Vectors for Word Representation) embeddings [26]. The glove.6B.100d model, trained on a corpus of 6 billion tokens, was employed to map tokenized text into dense 100-dimensional vectors, capturing semantic relationships between words.

The embedding layer was configured with the following parameters:

- Input length: 100 (maximum sequence length, determined via padding/truncation of tokenized inputs)
- Embedding dimension: 100 (vector size per token)

The GloVe embeddings were loaded as a trainable layer, enabling fine-tuning of the pre-learned word representations to the domain-specific linguistic patterns of the FN_data dataset. This approach enhances semantic alignment with Nigerian social media discourse and misinformation cues.

3.4. Attention Mechanisms

To improve both prediction and model interpretability, an attention mechanism was integrated into the hybrid architecture [6]. This lightweight attention layer, known for its computational and memory-efficient [27], assigns weights to each token based on its contextual relevance, determining how much focus the model gives to each token.

The attention computation is as follows:

1. Compute attention score for each token using a tanh function, a learned weight matrix (W), and bias (b):

$$e = \tanh(x * W + b) \quad (1)$$

x represents the input sequence

2. Exclude padded tokens from influencing attention by applying a binary mask:

$$e = e + (mask - 1) \times 10^9 \quad (2)$$

3. Convert scores to attention weights using the softmax function:

$$a = \text{softmax}(e) \quad (3)$$

4. Computing the context vector.

$$\text{weighted output} = x * a \quad (4)$$

$$\text{context vector} = \text{sum}(\text{weighted output}) \quad (5)$$

The context vector is equivalent to:

$$\text{context vector} = \sum(\text{attention weight} * \text{input token}) \quad (6)$$

The context vector summarizes the input sequence and displays token importance using bar charts (Section 4.5).

3.5. Model Development

The proposed model exploits the complementary strength of each component. CNN extracts local linguistic patterns from the dataset, enabling the model to capture contextual information within the text. BiLSTM processes sequences in both forward and backward directions, allowing for a more comprehensive understanding of sequence relationships and dependencies. The attention layer

assigns weights to input tokens, enabling focused analysis of deceptive cues. Figures 3 provide a detailed illustration of the CNN-BiLSTM architecture.

3.6. Model Architecture and Hyperparameters

The model's components and parameters, as illustrated in Figure 4, were selected based on established practices in text classification for fake news detection. The configuration and computational parameters include:

- **Embedding Layer**: Pre-trained GloVe embeddings (100 dimensions) were chosen as the initialization for the embedding layer due to their proven effectiveness in capturing semantic and syntactic relationships in news-related text [26]. The embeddings were made trainable (fine-tuned) with a reduced learning rate to adapt to the stylistic patterns in Nigerian social media content while preserving general linguistic knowledge.
- **1D Convolutional Layer (64 filters, kernel size 3, ReLU activation)**: A single layer of 1D Conv with 64 filters and kernel size 3 was selected to extract local n-gram features. Kernel size 3 is a common choice in text CNNs, it captures meaningful multi-word patterns without excessive parameter growth. The 64 filters provide sufficient representational capacity while keeping the model lightweight. ReLU activation promotes sparsity and faster convergence compared to sigmoid or tanh.
- **Batch Normalization**: Stabilizes and accelerates training by normalizing activations from the Convolutional layer. Applied after convolution to normalize activations, reduce internal covariate shift, stabilize gradients, and accelerate training [28]. This is particularly beneficial in hybrid architectures combining convolutional and recurrent layers.
- **BiLSTM layers**: Comprises 34 units in each direction (forward and backward), processing the feature maps bidirectionally to model long-range dependencies and contextual relationships across the sequence. This enables the capture of subtle linguistic nuances critical for distinguishing authentic from fabricated content [21] and avoiding excessive parameters that could lead to overfitting on moderately sized datasets.
- **Mask Layer**: Excludes padded tokens from computations to prevent bias in attention weighting and gradient updates.
- **Attention Layer**: Computes level relevance scores (Equations 1–5, Section 3.5) and generates a context vector which aggregates weighted input representations, functioning as an interpretable "spotlight" on salient deceptive signals.
- **Dropout Layers**: Applied a 0.3 rate after the BiLSTM layer and before the output layer to regularize training, prevent overfitting, and improve generalization by randomly dropping units during training.
- **Output Layer**: A single dense layer (64 units) with sigmoid activation performs binary classification (real vs. fake).
- **Optimizer (Adam), Learning Rate (0.002), Batch Size (64)**: The Adam optimizer was selected for its adaptive learning rates and robustness to sparse gradients in deep networks [29]. A learning rate of 0.002 was chosen with a batch size of 64 to balance memory usage and gradient stability for generalization.

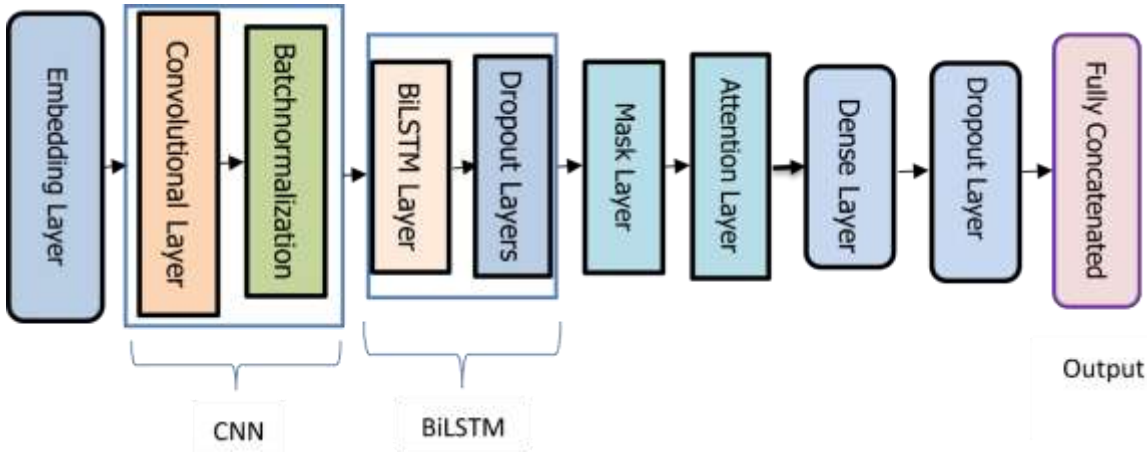

Figure 4. CNN-BiLSTM Component Architecture

4. Experimental Results and Discussion

All models were trained five times (FN_data) and ten times (ISOT) with different random seeds to evaluate stability of the models' performance. Results are reported as mean and standard deviation.

4.1. Training and Validation

Four deep learning architectures were trained using five (5) independent runs with different random seeds (42, 43, 44, 45, 46) to ensure statistical robustness and reproducibility. Adam optimizer, with a learning rate of 0.002 and a batch size of 64 to help in optimizing the training, learning, and validation to predict new data.

Table 1 presents the mean ± standard deviation across the five runs on the FN_data validation set.

Table 1. Performance Across Five Independent Runs (mean ± standard deviation)

MODELS	ACCURACY (%)	PRECISION (%)	RECALL (%)	F1 (%)
CNN	80.95 ± 0.18	67.50 ± 1.06	83.08 ± 3.72	74.42 ± 0.84
LSTM	82.29 ± 0.23	68.85 ± 0.88	85.88 ± 2.25	76.40 ± 0.45
BILSTM	82.23 ± 0.09	69.15 ± 0.81	84.55 ± 2.64	76.05 ± 0.60
CNN-BILSTM+ATTN	82.86 ± 0.22	69.60 ± 0.84	86.54 ± 2.31	77.12 ± 0.52

The proposed CNN-BiLSTM+Attn model achieved the highest average accuracy (82.86% ± 0.22%) and F1-score (77.12% ± 0.52%). The low standard deviations across all models (0.09–0.23% for accuracy, 0.45–0.84% for F1) indicate excellent stability and reproducibility, with BiLSTM showing the lowest variation in accuracy (±0.09%). The modest improvement of the hybrid model over BiLSTM and LSTM (≈0.6–0.9% in accuracy and F1) reflects realistic performance on short, noisy, abbreviation-rich Nigerian social media text.

The paired t-tests implemented for the five runs confirmed that the proposed hybrid model significantly outperformed all other models in accuracy and F1-score. For F1-score (Hybrid vs BiLSTM had $t = 3.74$, $p = 0.020$, Hybrid vs LSTM also had $t = 3.38$, $p = 0.028$, and Hybrid vs CNN with $t = 6.67$, $p = 0.003$) as seen in Table 2. For accuracy, (Hybrid vs BiLSTM has $t = 3.70$, $p = 0.003$, Hybrid vs LSTM recorded $t = 5.83$, $p = 0.004$, and Hybrid vs CNN $t = 13.89$, $p = 0.0002$) as presented in Table 3. For both f1-score and accuracy, $p < 0.05$ for all comparisons. These show statistically significant gains, validating the benefit of combining convolutional feature extraction, bidirectional sequential modeling, and attention weighting.

Table 2. Paired T-tests Across Five Runs for F1-score

MODELS	T-VALUE	P-VALUE
HYBRID VS CNN	6.67	0.003
HYBRID VS LSTM	3.38	0.028
HYBRID VS BILSTM	3.74	0.020

Table 3. Paired T-tests Across Five Runs for Accuracy

MODELS	T-VALUE	P-VALUE
HYBRID VS CNN	13.89	0.0002
HYBRID VS LSTM	5.83	0.004
HYBRID VS BILSTM	3.70	0.003

Figure 5 presents the performance metrics of the four models (CNN, LSTM, BiLSTM, and CNN-BiLSTM+Attn) across five (5) independent training runs on the FN_data dataset. The bar height represents the mean performance, and the error bars illustrate variability from run to run (standard deviation). Short error bars indicate low variability or high consistency, while long error bars indicate high variability or low consistency.

The CNN-BiLSTM+Attn model achieves the best result across all metrics. The models recorded lowest scores in Precision, indicating high rates of false positives, while Recall has the highest scores across all models, indicating the models are good at predicting positive class.

The models show low variability from run to run with standard deviation between 0.3 and 3.72%. Recall shows the largest variability across models, making it most sensitive to random seeds. Accuracy was the most stable metric across the models.

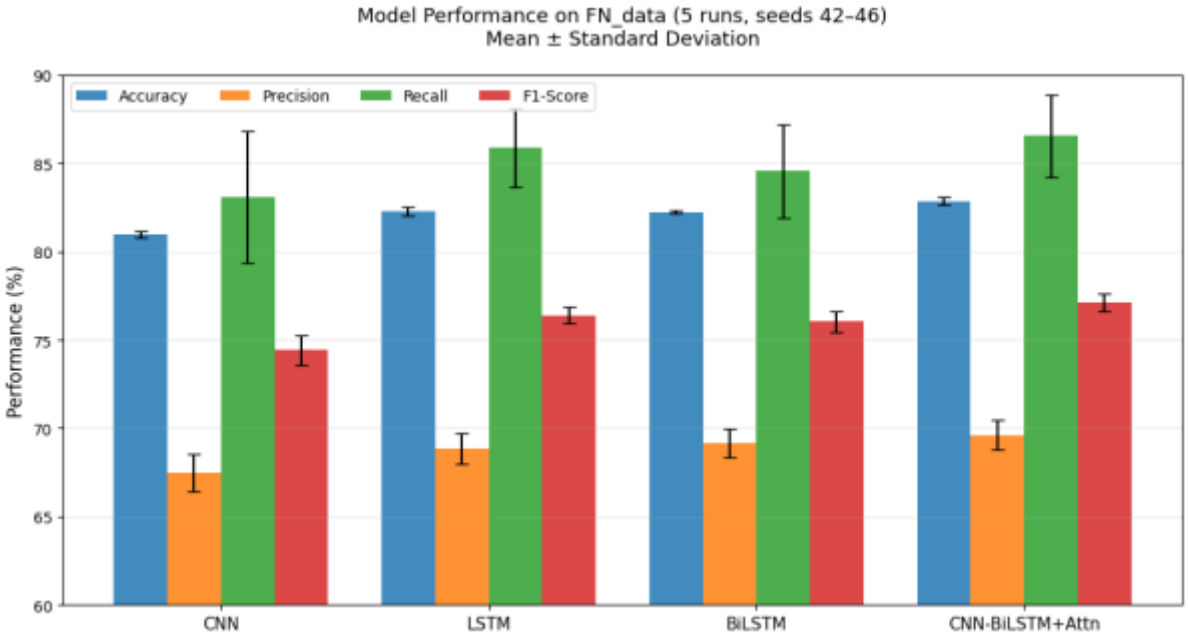

Figure 5. Models Performance Over (5) Training Run ()Mean Standard Deviation)

4.2. Performance Evaluation and Comparative Analysis

Performance of the models was assessed using standard classification metrics derived from the confusion matrix: accuracy, precision, recall, and F1-score, computed as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{TP+FN}, \text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$$

Where (TP), (TN), (FP), and (FN) denote true positives, true negatives, false positives, and false negatives, respectively. The TP and TN represent the correctly predicted classes while FP and FN represent the misclassified classes.

Figure 6-9 presents the confusion matrices for the best performing run of each model out of the five (5) independent runs (training), illustrating classification performance. Rows represent actual classes (fake or real), columns represent predicted classes.

Figure 6. Confusion Matrix for CNN-BiLSTM+Attn

TP = 13669, TN = 7451, FP = 1029, FN = 3246

- Accuracy = 0.8286 (82.86% correct predictions)
- Precision = 0.6960 (69.60% of predicted real news is accurate)
- Recall = 0.8654 (86.54% of actual real news are correctly identified)
- F1-score = 0.7712 (harmonic mean of precision and recall)

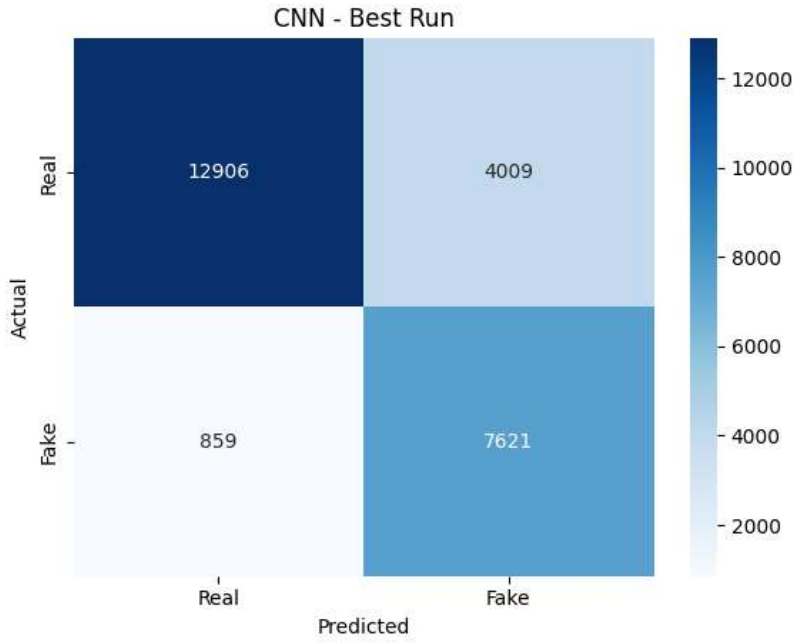

Figure 7. Confusion Matrix for CNN

TP = 12906, TN = 7621, FP = 859, FN = 4009

Figure 8. Confusion Matrix for BiLSTM

TP = 13395, TN = 7488, FP = 992, FN = 3520

Figure 9. Confusion Matrix for LSTM

TP = 13285, TN = 7582, FP = 898, FN = 3630

CNN-BiLSTM+Attn demonstrates superior diagonal performance, with 13,669 for real and 7,451 for fake, resulting in an accuracy of 83.17%. Which represents the overall performance of the model.

In F1-score, which is the harmonic mean of Precision and Recall, the proposed model had 77.12% score, making it the best performing model. The attention layer likely enhances the hybrid model by focusing on key tokens, leading to fewer misclassifications compared to BiLSTM and LSTM, which are closely matched with 82.23% and 82.29 accuracy % respectively. CNN had the least performance with 80.95% accuracy.

4.3. Performance of the Proposed Model on the ISOT Fake News Dataset.

To evaluate the generalization of the proposed model beyond the domain-specific Nigerian dataset, CNN-BiLSTM+Attn model was further evaluated on the standard ISOT fake news dataset across ten independent training runs with different random seeds (42-51), achieving a far better performance than that of FN-data dataset with mean accuracy of 99.33%, mean F1-score of 99.39%, precision 99.33%, and recall 99.46 as shown in Table 4.

Table 4. Proposed Model Performance on ISOT Dataset for 10 independent runs (mean ± std)

MODEL	ACCURACY (%)	PRECISION (%)	RECALL (%)	F1 (%)
CNN-BILSTM+ATTN	99.33 ± 0.10	99.33 ± 0.26	99.46 ± 0.18	99.39 ± 0.09

The high performance of the proposed model on the ISOT fake news dataset (99.33%) in contrast with its performance on FN_data (dataset from Nigerian social media), which is substantially lower (82.86% accuracy), highlights the impact of the quality of dataset on the performance of a model. Short sequences, abbreviations, and code-mixing, associated with social media text, reduce discriminative power compared to structured English news articles in ISOT [30], [31].

Figure 10 shows the confusion matrix for the best run (seed 5). The model correctly predicted 3,465 real classes (TP) and 4,208 false classes (TN), and misclassified 14 classes as FN and 26 classes as FP. This result validates the performance and robustness of the proposed model on fake news detection. The much lower standard deviation on the ISOT dataset across all metrics compared to FN-data dataset (± 0.10% vs ± 0.22) for Accuracy, (0.09 vs ± 0.52) for F1 score, (0.26 vs ± 0.84) for Precision, and (0.18 vs ± 2.31) for Recall, demonstrates the proposed model's training stability across different seeds.

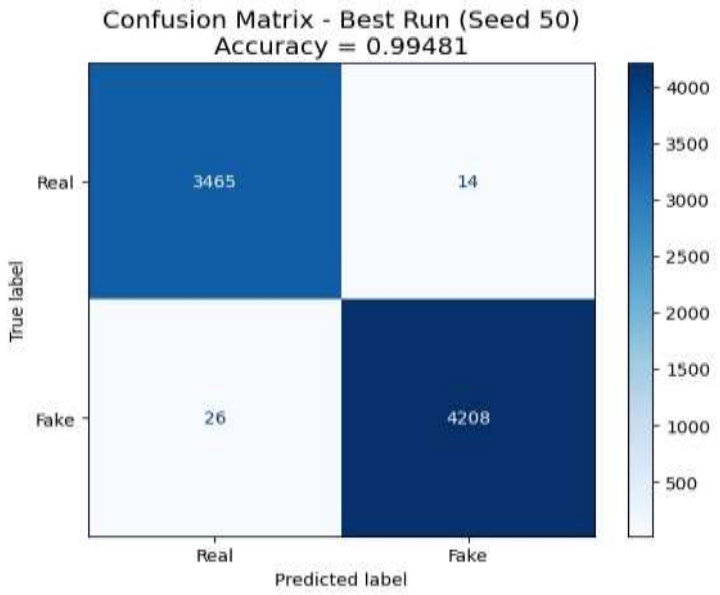

Figure 10. Confusion Matrix for CNN-BiLSTM+Attn on ISOT Dataset

4.4. Interpretable Prediction Analysis Using Attention Weights

We further plot the attention weights assigned to individual tokens to understand how the proposed model makes predictions. Attention weight distributions were extracted and visualized for sampled instances from the FN_data datasets. The aggregated attention scores were plotted in bar charts (Figures 11–16). Attention weights are normalized to [0,1] per instance and plotted in the order of appearance in the input sequence. We present both correctly classified and misclassified samples to highlight patterns and failure modes.

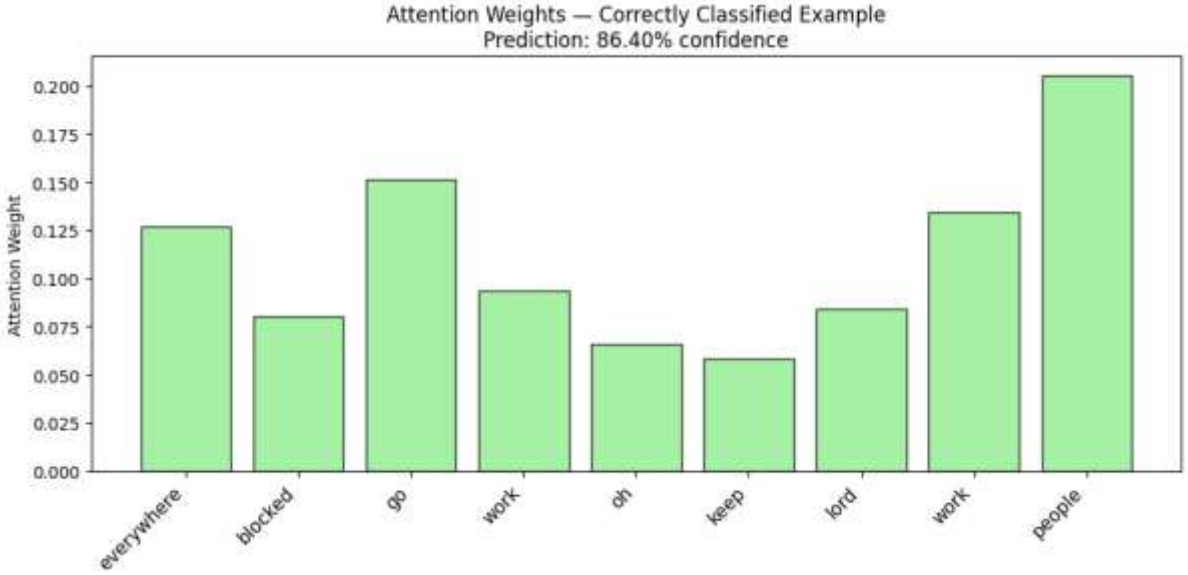

Figure 11. Correctly Classified Real News I

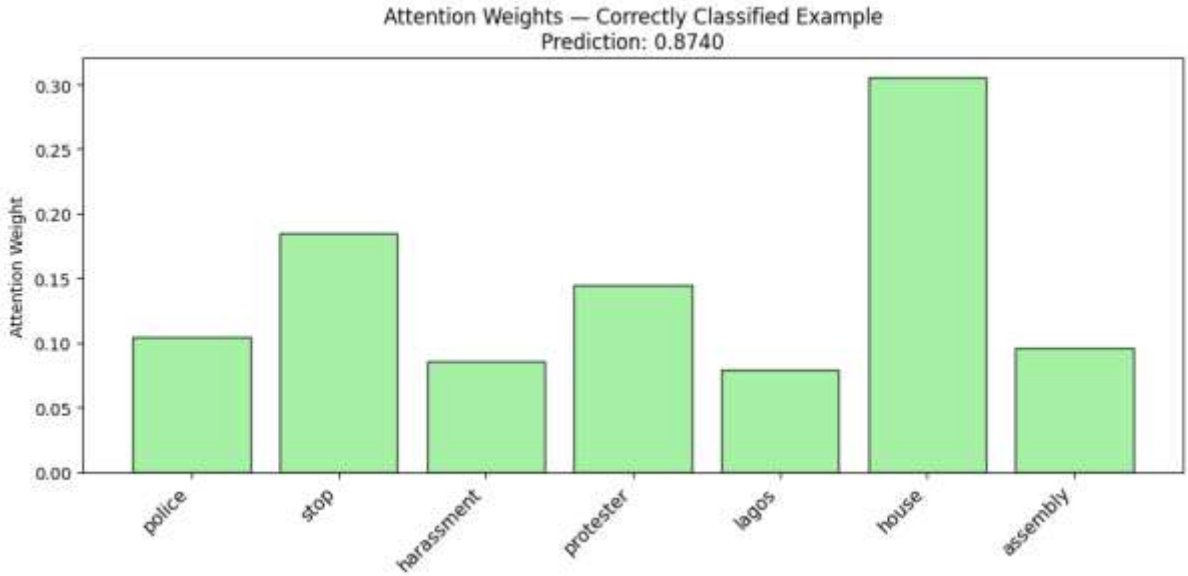

Figure 12. Correctly Classified Real News II

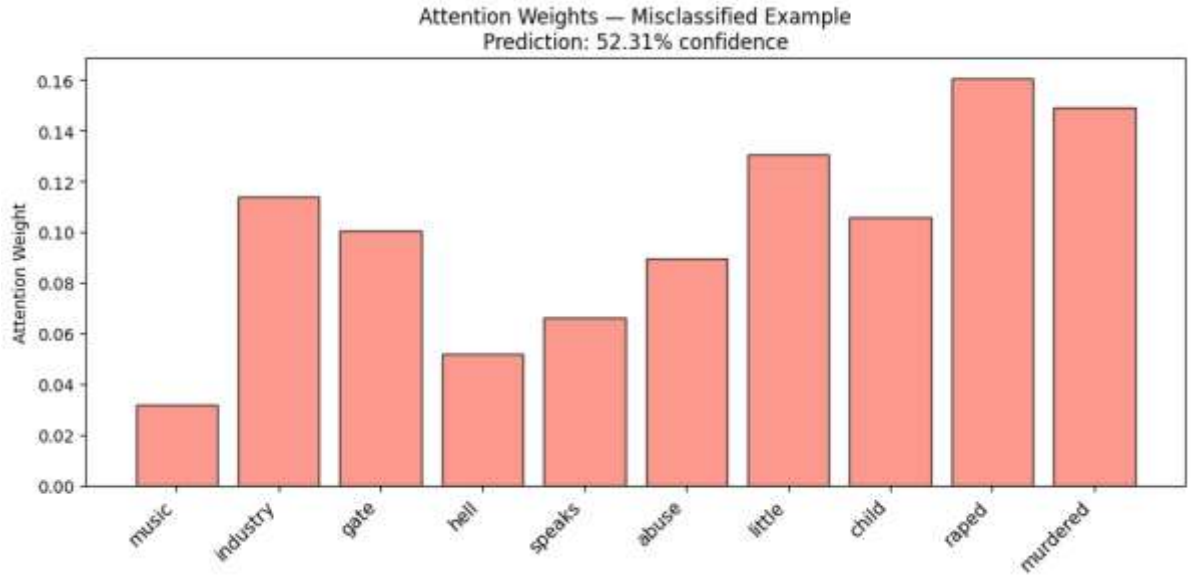

Figure 13. Misclassified Fake News I

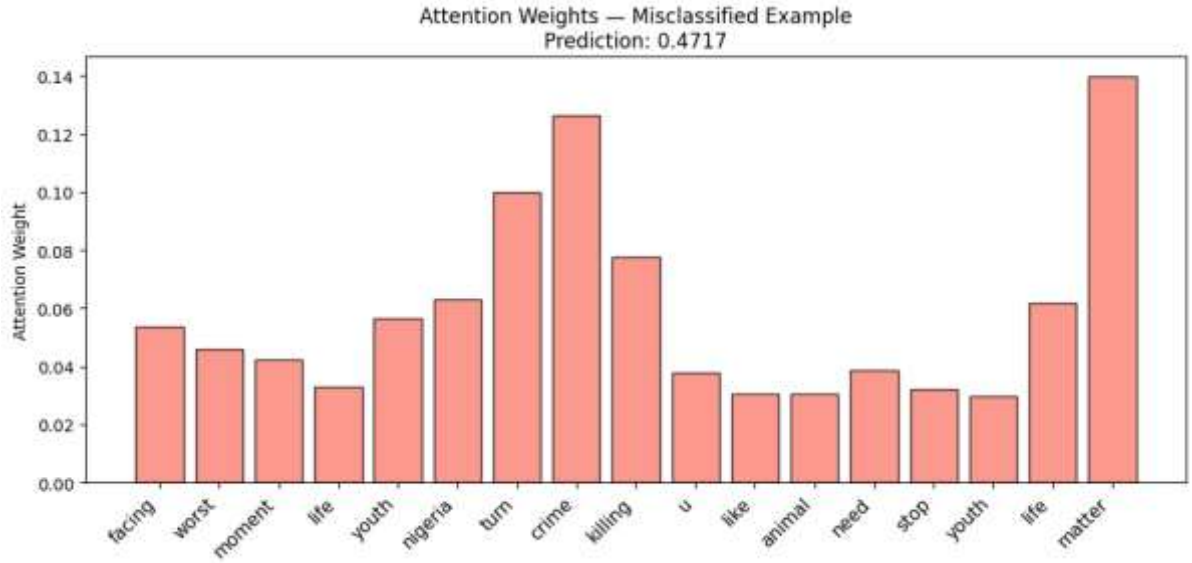

Figure 13. Misclassified Real News II

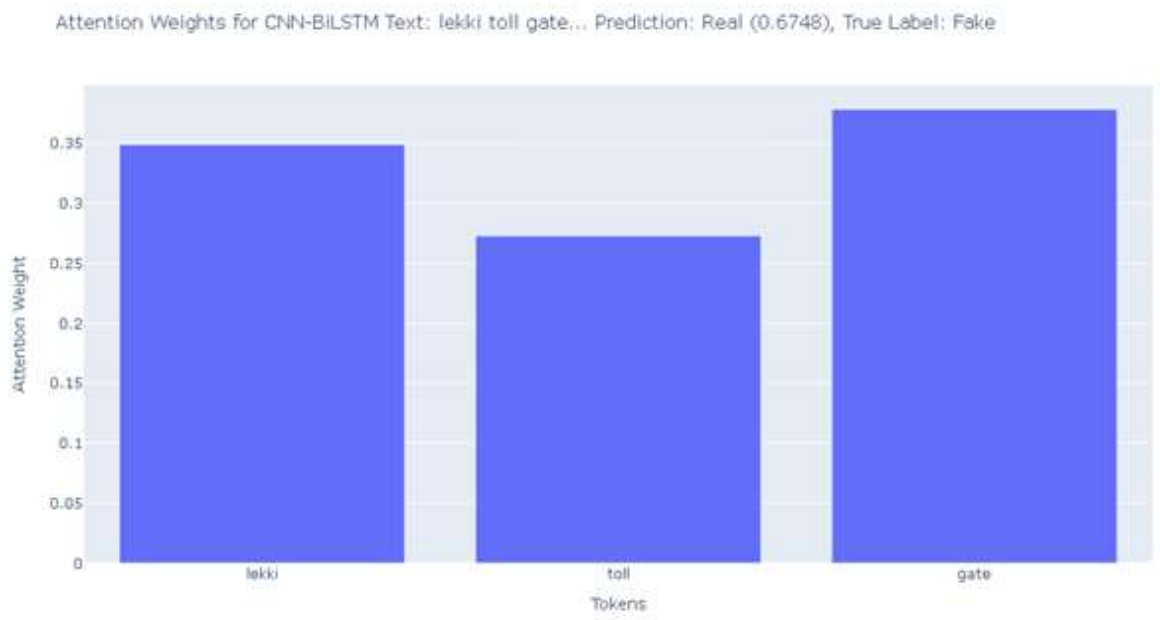

Figure 14. Misclassified Fake News

Figure 15. Misclassified Real News with abbreviations.

- Figure 11: Information predicted as real with confidence 86.40%.

Attention is high on tokens like “everywhere”, “go”, “people”, “lord”, “keep”, “work”)

Which are likely patterns that align with authentic information, an alert, or a call for an action. Location and action words (“everywhere”, “blocked”, “go”) received high weights, allowing the model to focus on substantive content for accurate classification as real.

- Figure 12: Protest-related, predicted Real, confidence 87.40%.

Attention concentrates on action-oriented and location-specific tokens (“house”, “assembly”, “protester”, “stop”, “police”). This focus on concrete actors and places that characterized the October 202 EndSARS protest in Lagos, Nigeria.

- Figure 13: Predicted Real, confidence 52.31%, actually Fake.

The model assigned high attention on verbs and nouns associated with violence (“raped”, “murdered”, “abuse”, “child”, “hell”). This over-emphasis on sensational terms, without sufficient context from surrounding tokens, likely contributing to the misclassification as real.

- Figure 14: Predicted Fake with confidence 47.17%, actually Real.

Misclassified a real news as fake, as the model assigned high attention to words like ”matter”, “crime,” “killing,” “worst,” “animal”, “life”, “youth”, and “nigeria”. The model emphasized on tokens carrying strong negative sentiment and tokens related to society to make decisions but failed to make the right prediction.

- Figure 15: Sentence misclassified as real for (“lekki”, “toll”, “gate”).

The model assigned higher weights to “lekki” and "gate," indicating that it emphasizes locations, leading it to incorrectly predict the news as real. The short sentence limits the ability to detect subtle terms that a longer sentence might reveal through sequential patterns.

- Figure 16: Misclassified real news (“r”,”figp”,”gp”).

The tokens in the sentence are likely abbreviations, which are treated as out-of-vocabulary (OOV) words with limited context, leading to misclassification.

These visualizations demonstrate the attention mechanism's effectiveness in classification task, by focusing on contextual and substantive tokens, for correct prediction. But the results also revealed vulnerabilities in misclassified cases, where over-emphasized weights on emotional or ambiguous terms led to misclassification.

These classification errors highlight broader challenges in DL-based fake news prediction. Its sensitivity to input length means that very short sentences like informal social media content often cause underfitting, as models like BiLSTM rely on long sequences for optimal performance. Abbreviations like “figp” or gp in Figure 16 introduced noise, leading the attention mechanism to focus on irrelevant tokens.

To mitigate these misclassifications, future models should incorporate advanced handling for short texts and noisy data, such as length-adaptive attention. Additionally, techniques such as character-level embeddings or sub-word tokenization (e.g., WordPiece or BPE) should be investigated to mitigate issues with abbreviations and Out-Of-Vocabulary (OOV) terms, which are common in social media text.

5. Conclusion

This study presents deep learning models (LSTM, BiLSTM, CNN, and CNN-BiLSTM + Attn on Nigerian social media content (FN_data). After five (5) training runs, the hybrid model achieves 82.86% ± 0.22% accuracy and 77.12% ± 0.52% F1-score, with statistically significant accuracy and F1 improvement over LSTM, BiLSTM and CNN. Validation on ISOT dataset (ten runs, seeds 42 - 51) yields an outstanding performance with 99.33% ± 0.01% accuracy and 99.39% ± 0.09% F1-score. The result demonstrates strong generalization to a more standard dataset and a consistent performance over the ten runs. Attention weights added robustness to the model by offering interpretable insights into deceptive cues. Limitations remain with short and abbreviated text, suggesting future directions in subwords tokenization and character-level embeddings.

Acknowledgment

I thank Prof. Francisca and Dr. Emeka for their guidance and commitment, and the Department of Computer Science, Federal University, Lokoja, for their invaluable support.

E.O. provided the idea and supervised the work; F.O. supervised and contributed valuable ideas and materials; M.A. conducted the experiment, analyzed results and wrote the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] O. D. Apuke and B. Omar, "Fake News Proliferation in Nigeria: Consequences, Motivations and Prevention Through Awareness Strategies," *Humanities & Social Sciences Reviews*, vol. 8, no. 2, pp. 318–327, Mar. 2020, doi: 10.18510/hssr.2020.8236.

[2] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," *Multimed. Tools Appl.*, 2023, doi: 10.1007/s11042-023-17470-8.

[3] E. Ogbuju, T. Abiodun, and F. Oladipo, "Text Analytics Solutions for the Control of Fake News: Materials and Methods," *International Journal of Open Information Technologies*, vol. 11, no. 3, pp. 69–74, 2023, Accessed: Apr. 29, 2024. [Online]. Available: https://cyberleninka.ru/article/n/text-analytics-solutions-for-the-control-of-fake-news-materials-and-methods

[4] A. Coleman, "'Hundreds dead' because of Covid-19 misinformation," BBC Monitoring. Accessed: Apr. 29, 2024. [Online]. Available: https://www.bbc.com/news/world-53755067

[5] A. S. Ogbette, M. O. Idam, A. O. Kareem, and D. N. Ogbette, "Fake News in Nigeria: Causes, Effects and Management," *Information and Knowledge Management*, Feb. 2019, doi: 10.7176/IKM/9-2-10.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE," 2015.

[7] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021, doi: 10.1016/j.neucom.2021.03.091.

[8] J. Posetti and A. Matthews, "A Short Guide to the History of 'Fake News' and Disinformation," *International Center for Journalists (ICFJ)*, 2018, doi: 10.1207/S15327728JMME1502_3.

[9] S. Barbara, "A Brief History of Fake News," Center for Information Technology and Society. Accessed: Jun. 04, 2024. [Online]. Available: https://cits.ucsb.edu/fake-news/brief-history

[10] K. M. Fouad, S. F. Sabbeh, and W. Medhat, "Arabic Fake News Detection Using Deep Learning," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3647–3665, 2022, doi: 10.32604/cmc.2022.021449.

[11] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, "Fake news detection using deep learning models: A novel approach," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 2, Feb. 2020, doi: 10.1002/ett.3767.

[12] E. D. Ajik, G. N. Obunadike, and F. O. Echobu, "Fake News Detection Using Optimized CNN and LSTM Techniques," *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 1044–1057, Aug. 2023, doi: 10.51519/journalisi.v5i3.548.

[13] Y. Taher, A. Moussaoui, and F. Moussaoui, "Automatic Fake News Detection based on Deep Learning, FastText and News Title," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130118.

[14] I. K. Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, "Detection of fake news using deep learning CNN–RNN based methods," *ICT Express*, vol. 8, no. 3, pp. 396–408, Sep. 2022, doi: 10.1016/j.icte.2021.10.003.

[15] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100007, Apr. 2021, doi: 10.1016/j.jjimei.2020.100007.

[16] G. GüLer and S. Gündüz, "Deep Learning Based Fake News Detection on Social Media," *International Journal of Information Security Science*, vol. 12, no. 2, pp. 1–21, Jun. 2023, doi: 10.55859/ijiss.1231423.

[17] A. Esan *et al.*, "Long-Short-Term Memory Model for Fake News Detection in Nigeria," 2023.

[18] E. M. Okoro, B. A. Abara, A. O. Umagba, A. A. Ajonye, and Z. S. Isa, "A hybrid approach to fake news detection on social media," *Nigerian Journal of Technology*, vol. 37, no. 2, p. 454, Jul. 2018, doi: 10.4314/njt.v37i2.22.

[19] S. Otor, B. Akumba, and J. Idikwu, "A Novel Fake-News Dataset and Detection System to Mitigate Cyber War with Emphasis on Nigerian News Events," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 24–32, Jun. 2023, doi: 10.32628/CSEIT23903146.

[20] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Front. Big Data*, vol. 6, Jan. 2024, doi: 10.3389/fdata.2023.1320800.

[21] T. Hung Vo, I. Felde, and K. C. Ninh, "Fake News Detection System, based on CBOW and BERT," *Acta Polytechnica Hungarica*, vol. 22, no. 1, pp. 27–41, 2025, doi: 10.12700/APH.22.1.2025.1.2.

[22] T. Mahmud, T. Akter, M. T. Aziz, M. Kamal Uddin, M. S. Hossain, and K. Andersson, "Integration of NLP and Deep Learning for Automated Fake News Detection," in *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, Jun. 2024, pp. 398–404. doi: 10.1109/ICICI62254.2024.00072.

[23] S. A. Al-obaidi and T. Çağlıkantar, "Automated Fake News Detection System," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 4, Nov. 2024, doi: 10.52866/2788-7421.1200.

[24] M. Neelamegan, S. Archanaa, N. V. Shree, and H. J. S. Sree, "Fake News Detection Using Deep Learning," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5089165.

[25] S. H. Muhammad *et al.*, "NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis," in *2022 Language Resources and Evaluation Conference, LREC 2022*, 2022.

[26] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[27] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," *Neural Comput. Appl.*, vol. 34, no. 16, pp. 13371–13385, Aug. 2022, doi: 10.1007/s00521-022-07366-3.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015.

[29] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[30] A. Altamimi, "Novel approach for predicting fake news stance detection using large word embedding blending and customized CNN model," *PLoS One*, vol. 19, no. 12, 2024, doi: 10.1371/journal.pone.0314174.

[31] M. Agbabiaka, E. Ogbuju, and F. Oladipo, "A Systematic Review of Deep Learning Approaches for Fake News Detection," *African Journal of Advances in Science and Technology Research*, vol. 21, no. 1, pp. 01–17, Nov. 2025, doi: 10.62154/ajastr.2025.021.01011.