



Fake news detection through multi-level contextual and sequential modeling

Pimpa Cheewaprakobkit* and Rindra Razafinjatovo

Faculty of Information Technology, Asia-Pacific International University, Saraburi 18180, THAILAND

*Corresponding author: pimpa@apiu.edu

ABSTRACT

The spread of false information on digital platforms destabilizes democratic societies and undermines public trust. Conventional machine-learning models often fail to capture the subtle contextual cues and long-range linguistic dependencies required for accurate verification. To address these limitations, we propose a hybrid deep-learning architecture that integrates three modules: A Contextual Encoder Module (CEM) using RoBERTa to generate relational embeddings for deep semantic extraction; a Bidirectional Sequence Learner (BSL) to model long-range temporal dependencies; and a Context Refinement Layer (CRL) that employs attention mechanisms to highlight the most salient deceptive markers. We evaluated the model on the WELFake dataset, which contains 72,134 news articles. Our proposed method achieved 99% accuracy, precision, recall, and F1-score, significantly outperforming state-of-the-art baselines, including BERT (95%) and BiLSTM (97%). The CRL's attention outputs provide transparency into the model's decision-making process, an essential feature for applications in misinformation detection and automated content moderation. Beyond its technical contributions, this research supports society by enabling early and reliable detection of false information, reducing harm from misinformation, and promoting a safer, more trustworthy digital environment.

Keywords: Fake news detection, WELFake dataset, Hybrid deep learning, Text classification, RoBERTa

INTRODUCTION

The proliferation of online communication platforms has reshaped, often quite rapidly, how people encounter, circulate, and interact with news. While digital media has broadened participation and lowered barriers to information, it has simultaneously created a permissive environment for the swift diffusion of fake news and fabricated narratives intended to mislead audiences and steer public opinion. At this point, the social and political fallout is, by most accounts, far-reaching: such content corrodes trust in journalistic institutions, exacerbates societal polarization, distorts electoral outcomes, and, in the longer run, imperils democratic resilience. Consequently, responding to this challenge entails more than media literacy initiatives and policy reform; it also, notably, calls for computational infrastructures capable of reliably discriminating authentic reporting from deceptive material with a degree of precision commensurate with the speed and scale of contemporary platforms. [1].

Early lines of work in fake news detection were anchored in conventional machine-learning classifiers, Support Vector Machines, Logistic Regression, and Decision Trees, rather than more expressive neural architectures. These systems drew on hand-engineered textual descriptors, most notably TF and IDF [2], along with n-gram statistics and a range of readability indices.

Although such feature-driven pipelines yielded moderate success on many benchmarks, their accuracy seldom transferred well across domains, which can be inferred to stem from a shallow grasp of context and a limited capacity to encode linguistic nuance [3]. Put differently, and as others have documented, models keyed to surface cues struggle with pragmatics and semantics, thereby undermining robustness under domain shift.

With the advent of deep learning, performance advanced markedly. To better capture sequential dependencies and local regularities, researchers introduced architectures based on Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) [4], and Convolutional Neural Networks (CNNs) [5]. Yet, even as these models delivered stronger benchmarks, they arguably remained limited in two respects: they struggled to encode rich, long-range semantics, and they offered insufficient interpretability for use in real-world applications.

Transformer-based architectures, most notably BERT and RoBERTa, have arguably reshaped natural language processing (NLP) by formalizing self-attention, thereby enabling the modeling of long-range dependencies and yielding more expressive semantic representations [6, 7]. In turn, and not least for fake news detection, these advances have produced markedly improved performance [8].

Nonetheless, the same transformers carry well-known liabilities: they cope poorly with extended input sequences and, by tending to distribute attention relatively evenly across tokens, can blur the faint deceptive cues on which detection often hinges [9].

To address these gaps, researchers have turned to hybrid deep learning designs that fuse heterogeneous architectures to capture both local cues and broader discourse-level structure. Notably, CNN-BiLSTM combinations underscore how convolutional pattern detectors can be profitably paired with sequence-aware learners [10, 11]. Yet a recurring shortfall remains: many such hybrids forgo fine-grained attention that would highlight the most contextually salient tokens and spans. At this point, it is worth noting that transformer-based hybrids, while generally more robust, are computationally intensive and, arguably, less transparent in their decision-making.

To address these limitations, we propose a hybrid framework comprising three mutually interdependent components:

- 1) Contextual Encoder Module (CEM), producing deep, semantically nuanced contextual embeddings by leveraging RoBERTa;
- 2) Bidirectional Sequence Learner (BSL), learning temporal dependencies in both forward and backward directions; and
- 3) Context Refinement Layer (CRL) employs fine-grained attention to foreground semantically informative content.

Together, these modules enhance contextual understanding, strengthen sequence modeling, and improve interpretability, thereby increasing accuracy and robustness in fake news detection.

Overall, this study introduces an integrated architecture that effectively unites contextual encoding, bidirectional learning, and attention-based refinement. An empirical evaluation on the WELFake dataset demonstrates that the proposed model achieves near-perfect accuracy, precision, recall, and F1 scores, outperforming state-of-the-art methods.

MATERIALS AND METHODS

This study introduces a hybrid deep learning framework for fake news detection, illustrated in Figure 1. The model integrates contextual encoding, sequential learning, and attention-based refinement through three key components: The Contextual Encoder Module (CEM) for deep semantic extraction, the Bidirectional Sequence Learner (BSL) for capturing long-range temporal dependencies in news narratives, and the Context Refinement Layer (CRL), which employs attention mechanisms to highlight the most salient deceptive markers. Together, these modules enable a comprehensive understanding of linguistic, semantic, and sequential patterns within textual data.

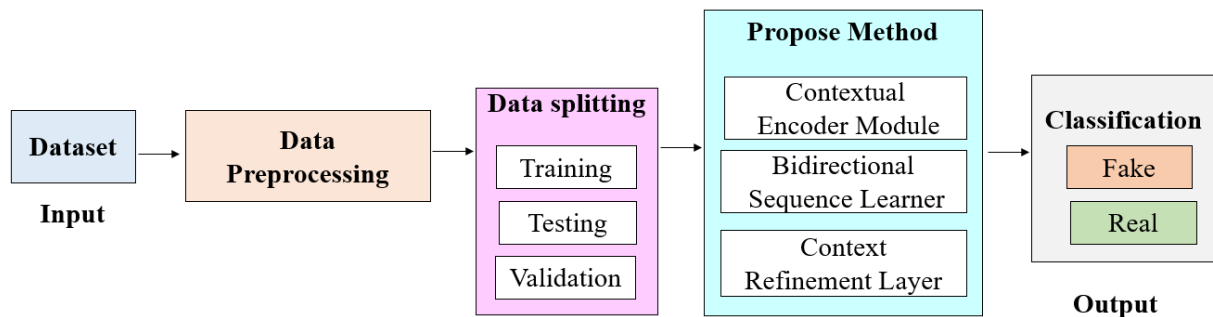


Figure 1 Our proposed architecture.

Data preprocessing

The WELFake dataset consists of raw textual content aggregated from diverse online sources, including news portals and social media platforms. Consequently, the raw data contains typical web-based artifacts and formatting irregularities that must be addressed prior to model training. To ensure high data quality while maintaining full compatibility with the RoBERTa encoder, we adopted a structured preprocessing strategy focused on noise reduction without sacrificing semantic integrity.

First, non-semantic artifacts introduced during web scraping, such as HTML tags, embedded hyperlinks, email addresses, and extraneous metadata, were removed, as they do not contribute meaningful linguistic information. Redundant whitespace was normalized

to ensure textual consistency. This preprocessing stage was explicitly designed to align with the RoBERTa tokenizer. Once cleaned, the text is tokenized and passed to the Contextual Encoder Module (CEM) for embedding generation. This design choice ensures that the architecture receives high-fidelity input, supporting robust semantic representation and contributing to the model's strong generalization performance [12].

Data splitting

To promote an equitable and unbiased assessment, the dataset was divided into training (70%), validation (15%), and test (15%) partitions via stratified sampling. Notably, this procedure, by design, preserves the class priors (fake versus real) within each subset. On the training split, parameter estimation

is performed; on the validation split, hyperparameters are tuned to temper overfitting; and on the holdout test set, out-of-sample generalization is assessed. Following the split, all documents are routed through the Contextual Encoder Module (CEM), which maps tokenized text to contextualized embeddings for downstream modules.

Contextual Encoder Module (CEM)

Within the CEM, as illustrated in Figure 2, we employ the RoBERTa-base model [13, 14], a bidirectional transformer, to generate context-conditioned semantic embeddings. In practical terms, both titles and the full body text are fed into the encoder, which maps them to dense vectors whose values are modulated by neighboring tokens and broader discourse cues. By contrast with static embeddings, these contextualized representations update with the local context, encoding syntactic dependencies and semantic relations in ways that arguably enable more fine-grained discrimination between authentic and fabricated content.

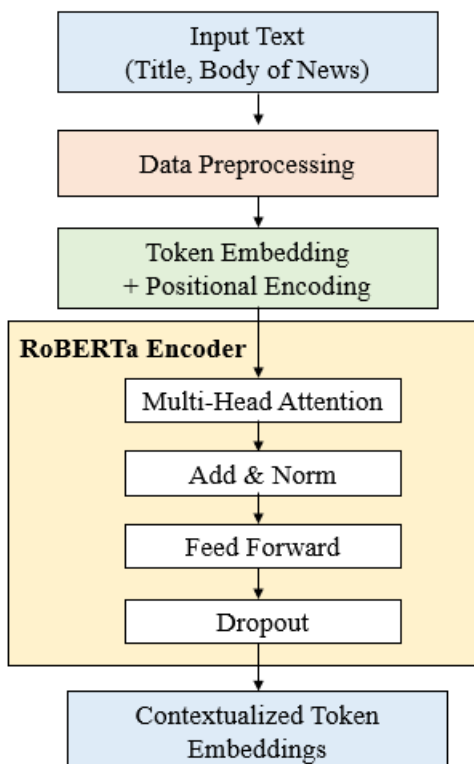


Figure 2 Contextual Encoder Module (CEM) flowchart.

At this stage, after cleaning and tokenization, the sequence is mapped to token embeddings and augmented with positional encodings, which together effectively encode lexical semantics while preserving token order. Notably, the core module is a RoBERTa encoder composed of twelve Transformer blocks that interleave multi-head self-attention with residual add-and-normalize steps, position-wise feed-forward sublayers, and dropout regularization. Accordingly, for downstream processing, these components produce high-dimensional token-level embeddings that, in

effect, preserve both syntactic structure and semantic content.

Among the hallmark contributions of the transformer is, arguably, the multi-head self-attention mechanism [8]. Rather than computing a single attention map, queries, keys, and values are first linearly projected into several learned subspaces, and within each head, attention is computed independently. This setup allows the model to attend, in parallel, to the task of distinguishing linguistic dependencies. The outputs from all heads are subsequently concatenated and passed through a final linear layer, thereby producing a composite attention representation.

Formally, the multi-head attention is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^o \quad (1)$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

Here, W_i^Q , W_i^K , and W_i^V are the learned projection matrices for the queries, keys, and values, respectively, with dimensions $R^{d_{\text{model}} \times d_k}$, $R^{d_{\text{model}} \times d_k}$, $R^{d_{\text{model}} \times d_v}$. The final output projection matrix W^o has dimensions $R^{hd_v \times d_{\text{model}}}$.

Within this framework, self-attention serves as the core of the architecture. By design, it allows the model to assign relative salience to tokens anywhere in the sequence, regardless of distance. The mechanism proceeds by projecting the input into query (Q), key (K), and value (V) matrices through learned linear maps. Notably, each projection fulfills a distinct function: the query encodes what a given token seeks from its context, the key summarizes features of all tokens for matching, and the value carries the content to be pooled. Relevance across tokens is computed as the dot product between Q and K^T , yielding pairwise compatibility scores. To temper their magnitude and consequently stabilize gradient dynamics, the scores are rescaled by the square root of the key dimension (d_k). At this stage, the scaled similarity scores are fed into a softmax, which normalizes them into attention weights that quantify each token's influence on the current representation. Applied thereafter to the value matrix, these weights yield embeddings conditioned on context across the sequence. In this formulation, arguably the mechanism's fulcrum, the model can foreground the most pertinent portions of the input irrespective of position, thereby capturing long-range dependencies and refining semantic interpretation. Formally, the self-attention operation can be written as:

$$\text{Attention}(K, Q, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V_i \quad (3)$$

Through this mechanism, the model assigns differential salience to tokens across the entire sequence, regardless of how far apart they are, enabling it to model long-

range dependencies alongside fine-grained semantic nuances. In turn, the Contextual Encoding Module (CEM) emits a stream of context-enriched embeddings that, at this point, serve as input to the Bidirectional Sequence Learner (BSL).

Bidirectional Sequence Learner (BSL)

Once the CEM has produced contextualized token representations, those embeddings are, at this point, fed into a Bidirectional Sequence Learner (BSL) tasked with modeling the text's temporal structure and other sequential regularities. Scanning the sequence in both directions (left-to-right and right-to-left), the BSL exploits prior and forthcoming context alike, thereby capturing dependencies that hinge not only on preceding tokens but, notably, on tokens still to come, including long-range interactions. Consequently, this two-way treatment, arguably indispensable in practice, affords a more complete account of discourse-level dynamics, helping preserve narrative flow, maintain structural coherence, and uphold semantic intent. Notably, and this is especially pertinent to our use case, the bidirectional pass tends to surface faint inconsistencies and recurrent structural motifs that can be read as hallmarks of sophisticated disinformation.

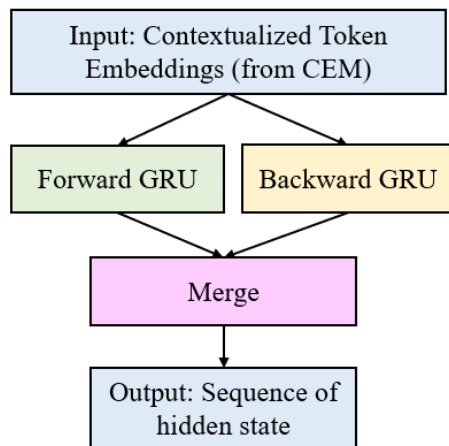


Figure 3 Bidirectional Sequence Learner (BSL) flowchart.

As shown in Figure 3, the BSL ingests the contextualized token embeddings produced by the CEM as its initial input. Instead of a single sweep, the module executes two passes in parallel: a forward pass that traverses the sequence from start to end, and a backward pass that traverses it in reverse. The activations from these two, stemming from distinct yet complementary subnetworks, are then concatenated to integrate context gathered in both temporal directions. Consequently, the BSL emits a sequence of hidden states, each state encoding a fused representation that blends forward and backward information at its corresponding position. Taken together, this output furnishes an arguably comprehensive, context-aware view of the entire sequence and is then handed off to the Context Refinement Layer (CRL).

Context Refinement Layer (CRL)

The Context Refinement Layer (CRL), positioned at the final stage of our architecture, is an attention-based module that focuses on the most informative cues for the downstream classifier. Into this layer, we pass the sequence of hidden states emitted by the Bidirectional Sequence Learner (BSL) without further transformation. Rather than treating all tokens uniformly, the CRL computes, for each hidden state, an attention coefficient that functions as an important score for its corresponding word in the sequence. Notably, words that are especially pertinent to disinformation detection, strong affective markers, concrete yet uncorroborated assertions, or propagandistic turns of phrase tend, consequently, to receive higher weights. Applying these weights yields a single, refined context vector, obtained by summing the hidden states. This vector, which can be inferred to provide a compact and task-tuned synopsis of the entire input, is optimized for the detection objective. In turn, the resulting representation serves as the final feature set and is forwarded to the classification layer to produce the prediction. The operational flow of the CRL is depicted in Figure 4.

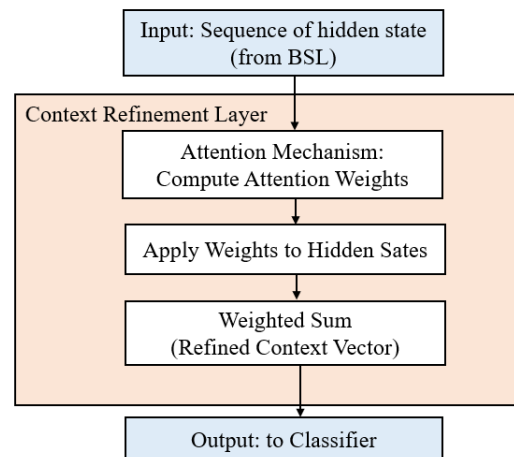


Figure 4 Context Refinement Layer (CRL) flowchart.

Classification

As depicted in Figure 5, the refined context vector is first subjected to dropout, a regularization technique intended to curb overfitting, and then routed to a dense (fully connected) layer that performs dimensionality compression and, arguably, feature consolidation. The resulting scalar is subsequently transformed by a sigmoid activation function [15]. A sigmoid activation function is defined as:

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (4)$$

where z denotes the weighted sum of features and bias, e is Euler's number (the base of the natural logarithm). The sigmoid function maps outputs to

the range [0,1], representing the probability that an input belongs to the “fake news” class (1). A threshold of 0.5 is used for binary classification: values above 0.5 are labeled as fake, and those below are labeled as real.

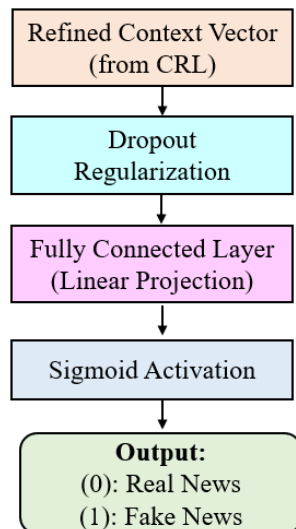


Figure 5 Classification flowchart.

Training and optimization

In practice, we optimized the Binary Cross-Entropy with Logits loss using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 64. Training was conducted on an NVIDIA GeForce RTX 4060 over 10 epochs. To ensure robust generalization, we monitored performance on a held-out validation set throughout training. The final model selected for evaluation is the checkpoint with the highest validation performance.

RESULTS AND DISCUSSION

Dataset

In this study, the WELFake dataset serves as the primary benchmark for evaluating fake news detection performance [16]. Publicly accessible and now widely used in the misinformation-detection literature, WELFake was curated specifically to support machine-learning models. Offering a balanced and comprehensive corpus of news articles, it is, consequently, well-suited to robust model training and evaluation. The dataset comprises 72,134 articles, 37,106 of which are labeled as fake and 35,028 as real. Drawn from Kaggle, the collection provides linguistic diversity alongside credible provenance. Notably, each entry contains three fields: Title, the article's headline; Text, the full body; and Label, a binary indicator in which 1 marks fake and 0 marks real news.

Figure 6 visualizes the distribution of article length for real versus fabricated news. Word count is plotted on the abscissa, and the ordinate records the number of articles in each bin. Fake items, by

and large, cluster at shorter lengths, whereas real news spans a broader, often longer, range. Even so, the two curves overlap appreciably, suggesting that length, taken in isolation, has limited diagnostic utility. Consequently, dependable classification requires deeper contextual and linguistically informed representations.

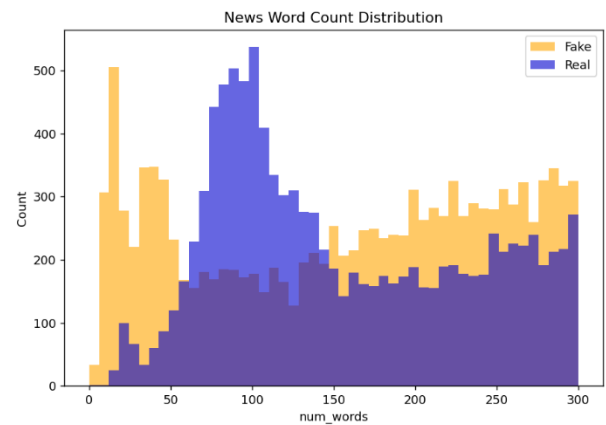


Figure 6 Word Count Distribution for Real and Fake News.

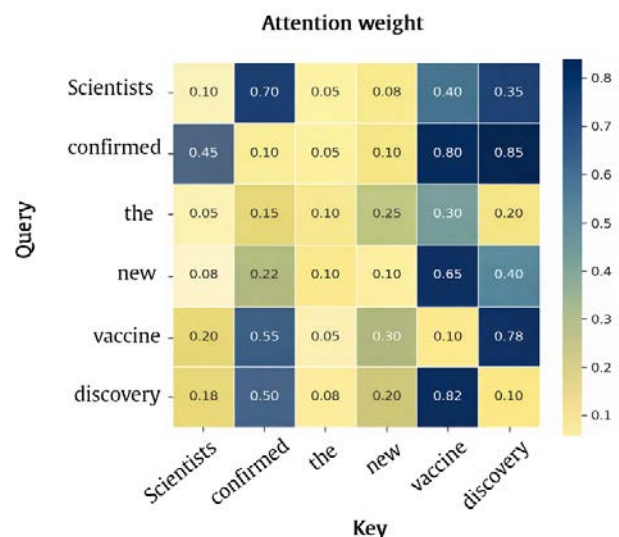


Figure 7 Attention weight heatmap.

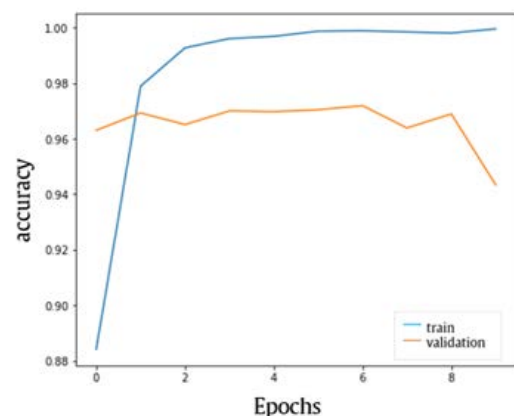


Figure 8 Training and validation accuracy curves.

Figure 7 visualizes attention weights from the Context Refinement Layer (CRL) to illustrate the

model's decision-making process. Darker blue shades indicate higher attention weights, highlighting tokens the model identifies as critical for classification. In contrast, lighter shades represent common function words or less meaningful terms that contribute less to the final representation. This visualization demonstrates how the CRL isolates contextually salient cues, such as affective markers or uncorroborated assertions, to refine the context used to detect fake news.

Figure 8 illustrates the classification performance of the proposed model throughout training, tracking its accuracy on both the training and validation sets. The curve monitors the overall proportion of correctly identified articles, both real and fake, as the model learns. A steady increase in these curves indicates that the hybrid architecture is effectively capturing the necessary contextual and sequential patterns to achieve its high final accuracy. Figure 9 displays the training and validation losses, which represent the error rates the model aims to minimize during optimization. A consistent decline in loss across both datasets demonstrates that the model is generalizing well and is not overfitting.

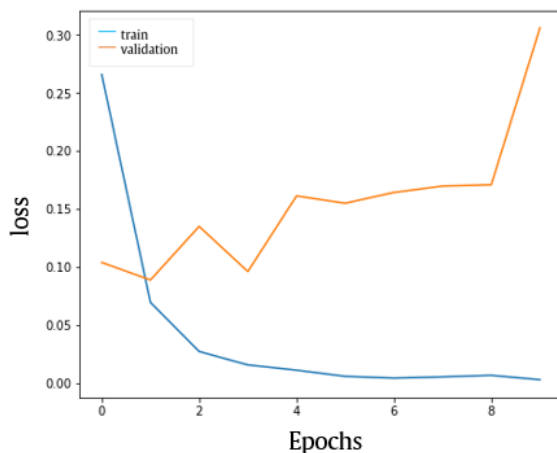


Figure 9 Training and validation loss curves.

Evaluation metrics

To rigorously assess model performance, four standard classification metrics were employed: accuracy, precision, recall, and F1-score [17].

Accuracy measures the overall proportion of correctly classified articles (both real and fake news). The formula is shown in (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

As shown in (6), precision, the fraction of truly fake articles within all items the model labels as fake (i.e., the positive predictive value), effectively gauges how accurately the model flags fake news.

$$Precision = \frac{TP}{FP+TP} \quad (6)$$

Recall measures the proportion of actual fake news articles that the model successfully detects, reflecting its ability to capture fake news cases, as shown in (7).

$$Recall = \frac{TP}{FN+TP} \quad (7)$$

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of accuracy and detection capability, as shown in (8).

$$F1 = 2 * \frac{Precision * Recall}{Precision+Recall} \quad (8)$$

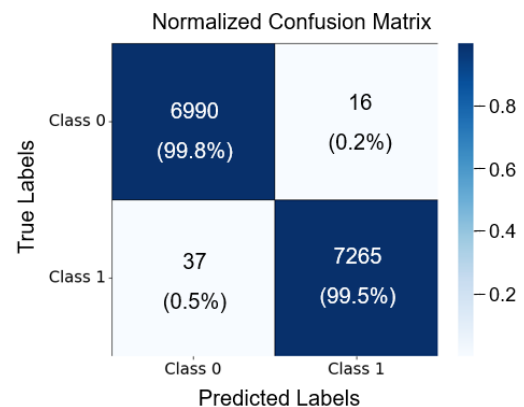


Figure 10 Normalized Confusion Matrix.

Figure 10 depicts the normalized confusion matrix for the proposed model. Considered together, the True Positive (TP) and True Negative (TN) entries are high; in aggregate, this points to robust discriminative performance. With the False Positive (FP) rate at only 0.2%, the system seldom tags legitimate news as fake, an aspect that, arguably, sustains public confidence in automated detection tools. By contrast, the False Negative (FN) rate is 0.5%, slightly higher yet still small, suggesting that only a handful of fake items slip through. Even so, such misses are nontrivial from a risk perspective, as undetected misinformation can propagate across digital networks. At this point, incorporating credibility scores or metadata-informed features could, arguably, further curtail this residual error.

On the WELFake corpus, Table 1 consolidates the side-by-side performance of the proposed framework against a suite of established machine- and deep-learning baselines. As an initial yardstick, the classical learners, CATBoost, Naïve Bayes, and SVM, appear to yield accuracies in the 89-92% range, which can be taken as a pragmatic baseline for automated fake news detection. Notably, logistic regression edges higher to 96%, whereas plain artificial neural networks (ANNs) hover around 95%.

Table 1 Result comparison of models on the WELFake dataset.

Algorithm	Accuracy	F1-Score	Precision	Recall	Reference
RoBERTa	0.86	0.85	0.85	0.86	[18]
CATBoost	0.89	0.89	0.89	0.89	[19]
GRU	0.90	0.90	0.90	0.90	[19]
SVM	0.92	0.92	0.88	0.92	[19]
BERT-Transformer-based Learning	0.95	0.95	0.95	0.95	[20]
Artificial Neural Network	0.95	0.95	0.94	0.95	[21]
Logistic regression	0.96	0.96	0.97	0.95	[21]
Naïve Bayes	0.89	0.89	-	-	[22]
LSTM	0.96	0.96	-	-	[22]
BiLSTM	0.97	0.97	0.98	0.98	[23]
CDLR	0.98	0.98	0.98	0.98	[24]
Proposed method (Our)	0.99	0.99	0.99	0.99	-

Further gains emerge with sequential architecture. GRU, LSTM, and BiLSTM improve accuracy by modeling temporal dependencies in the text; among them, BiLSTM reaches 97%, which arguably underscores the added value of bidirectional context.

Turning to transformer-based systems, their capacity for rich contextual inference is evident, though not uniformly realized. BERT posts 95% accuracy, while RoBERTa, despite its more advanced pretraining, registers a lower 86% on this dataset, likely reflecting limited task-specific fine-tuning or insufficient domain adaptation. These patterns, consequently, reinforce the need to tailor transformer models to the quirks of particular corpora and objectives.

Hybrid designs that blend contextual and sequential cues (e.g., CNN, BiLSTM, or CDLR) achieve up to 98% accuracy, attesting to the advantages of architectural fusion. At this point, the proposed hybrid framework exceeds all baselines, achieving 99% across all evaluation metrics: accuracy, precision, recall, and F1-score. This outperformance seems to arise from the synergistic integration of three components: CEM, which provides deep contextual embeddings through RoBERTa; BSL, operating bidirectionally, models the temporal and sequential dependencies; CRL, notably, recalibrates contextual relevance through attention-weighted scoring. Consequently, this layered design arguably allows the model to recover semantic, syntactic, and temporal dependencies in a unified manner.

Across all metrics, reaching 99% arguably signals not only the model's high accuracy but also its capacity to generalize and to be interpreted,

attributes that, notably, constitute core prerequisites for the operational deployment of misinformation detection systems.

Our results align with recent findings in the literature suggesting that hybrid models outperform standalone transformers. While previous studies using BERT achieved accuracies around 95%, our model's 99% performance on the WELFake dataset is due to the BSL and CRL modules, which capture nuances that standard encoders often overlook. This suggests that, for large-scale datasets, integrating sequential narrative tracking is as critical as capturing the text's deep meaning.

Ablation study

To evaluate the individual contribution of each module within our proposed architecture, we conducted an ablation study using the WELFake and FakeNewsNet datasets. We assessed three configurations: The Contextual Encoder Module (CEM) in isolation, the combination of CEM and the Bidirectional Sequence Learner (CEM+BSL), and our full hybrid model (CEM+BSL+CRL). The inclusion of the Context Refinement Layer (CRL), forming the full model, further elevated performance to 99% across all metrics. The CRL's attention mechanism identifies and weights contextually salient tokens, effectively filtering noise and sharpening the classifier's focus on deceptive markers. This progression demonstrates that while each module is beneficial on its own, their synergistic integration is critical for achieving near-perfect detection rates, as shown in Table 2.

Table 2 Ablation study results on the WELFake dataset.

Algorithm	Accuracy	F1-Score	Precision	Recall
CEM	0.86	0.88	0.82	0.95
CEM+BSL	0.97	0.97	0.97	0.98
CEM+BSL+CRL (Our)	0.99	0.99	0.99	0.99

To demonstrate the generalizability of our framework beyond the primary benchmark, we evaluated the model on the FakeNewsNet dataset. FakeNewsNet presents a distinct challenge, as it comprises 23,921 news items from Twitter, with diverse social context features and shorter text lengths than in standard news articles. As shown in Table 3, the proposed model maintained high performance

on this additional dataset, achieving an accuracy and F1-score of 0.95. Consistent with the ablation results, the full model outperformed the standalone CEM (85% accuracy) and the CEM+BSL configuration (92% accuracy). The architecture's ability to maintain high precision (0.95) and recall (0.95) on a social media dataset confirms its robustness across varying linguistic styles and data distributions.

Table 3 Comparison of model results on the FakeNewsNet dataset.

Algorithm	Accuracy	F1-Score	Precision	Recall
CEM	0.85	0.90	0.89	0.92
CEM+BSL	0.92	0.92	0.92	0.92
CEM+BSL+CRL (Our)	0.95	0.95	0.95	0.95

The 99% accuracy achieved on the WELFake dataset is attributed to the synergistic integration of the CEM, BSL, and CRL modules. While the standalone CEM model achieves 86% accuracy, our multi-level architecture sequentially captures deep semantics and temporal dependencies, thereby eliminating residual errors. Furthermore, the WELFake dataset is a large, balanced benchmark optimized for machine learning. To ensure these results were not due to overfitting, we validated the model on the FakeNewsNet dataset, where it maintained a robust accuracy of 95%, confirming the architecture's strong generalizability across different data distributions.

CONCLUSIONS

This study presents a new hybrid deep learning architecture for fake news detection that weaves together three components: a Contextual Encoder Module (CEM), built on RoBERTa to derive deep contextual embeddings; a Bidirectional Sequence Learner (BSL), which models temporal dependencies in both forward and backward directions; and a Context Refinement Layer (CRL) that applies fine-grained attention to surface semantically salient cues. Taken as a whole, these modules jointly remedy the shortcomings of standard machine learning and earlier neural models, approaches that, in practice, often fail to capture contextual nuance and sequential structure in tandem. On the WELFake dataset, the model attains 99% accuracy and surpasses all benchmark baselines. Notably, this performance appears to validate the multi-level integration strategy: contextual encoding fosters richer semantic understanding; bidirectional sequence modeling captures temporal relations that unidirectional pipelines tend to miss; and attention-based refinement sharpens interpretability while improving precision. The combination, consequently, suggests a deeper link between layered representation learning and robust detection in this domain.

Even so, several caveats deserve mention. First, although WELFake provides a widely used benchmark, it may not reflect the full linguistic heterogeneity or the evolving nature of misinformation observed in dynamic online settings (where concept drift and domain shift are common). Many social media items are multimodal, including images, videos, and memes, but the present model does not analyze them. Second, computational burden remains non-trivial; fine-tuning transformer encoders such as RoBERTa is resource-intensive and may impede real-time or large-scale deployment. Finally, the model currently assumes a binary label space (real vs. fake), whereas misinformation in the wild spans a continuum, from satire and propaganda to partially true content, that future work should more faithfully represent.

Taken together, the present results open several plausible extensions. To capture more nuanced patterns of deception, an obvious next step is a multimodal detector that fuses textual, visual, and metadata cues, potentially via co-attention or contrastive alignment, to leverage correlations among language, imagery, and provenance. Equally promising, though for a different reason, is domain adaptation and transfer learning aimed at robust cross-platform, cross-topic generalization with minimal retraining; under rapid topic drift and distributional shift, techniques such as unsupervised adaptation, parameter-efficient tuning (e.g., adapters or LoRA), or continual learning could be especially useful. At this point, efficiency also matters. Exploring lightweight architectures and model compression, pruning, distillation, quantization, and related low-rank or sparsity-inducing methods would improve throughput and latency for real-time inference, including in resource-constrained settings. Finally, integrating explainable AI tooling, whether through post hoc methods (saliency, concept activation, counterfactuals) or intrinsically interpretable modules, seems likely to deepen insight into model reasoning and to aid transparency, calibration, and user trust.

In summary, this study advances automated fake news detection by proposing an integrated deep learning architecture that, arguably, balances predictive performance with interpretability. While limitations remain, the evidence presented here lays a durable foundation for subsequent inquiry and, consequently, indicates strong potential for practical deployment against the accelerating spread of online misinformation.

DECLARATION OF AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors used ChatGPT to improve the clarity and readability of the manuscript during its preparation. All generated text was carefully reviewed, revised, and validated by the authors after using this tool/service. The authors take full responsibility for the final content of the publication.

REFERENCES

1. Paliwal S, Bharti R. Reviewing the landscape: Fake news detection methods and social media user behavior. *Indian J Tech Educ.* 2025;48(2): 207-16.
2. Nuri Y, Senyurek E. Research Abstracts Similarity Implementation By Using TF-IDF Algorithm. *IOSR J Comput Eng.* 2025;27(1):4-10.
3. Shah S, Patel S. A Comprehensive Survey on Fake News Detection Using Machine Learning. *J Comput Sci.* 2025;21(4):982-90.
4. Padalko H, Chomko V, Chumachenko D. A novel approach to fake news classification using LSTM-based deep learning models. *Front Big Data.* 2023;6:1-18.
5. Haqi Al-Tai M, Nema BM, Al-Sherbaz A. Deep Learning for Fake News Detection: Literature Review. *Al-Mustansiriyah J Sci.* 2023;34(2):70-81.
6. Alghamdi J, Lin Y, Luo S. Towards COVID-19 fake news detection using transformer-based models. *Knowl Based Syst.* 2023;274:110642.
7. Aremu T, Akinwehinmi O, Nwagu C, Ahmed SI, Orji R, Amo PAD, et al. On the reliability of Large Language Models to misinformed and demographically informed prompts. *AI Mag.* 2025;46(1):1-15.
8. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. *Neural Inf Process Syst.* 2017:6000-10.
9. Boukouvalas Z, Shafer A. Role of Statistics in Detecting Misinformation: A Review of the State of the Art, Open Issues, and Future Research Directions. *Annu Rev Stat Appl.* 2024;11(1):27-50.
10. Alghamdi J, Lin Y, Luo S, editors. Modeling Fake News Detection Using BERT-CNN-BiLSTM Architecture. 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR). 2022; 354-357
11. Xiaoyan L, Raga RC, Xuemei S, Liguori R. GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews. *J Sens.* 2022;1-12.
12. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *J Healthc Eng.* 2018;4302425.
13. Shuhai Zhang, Zeng You , Yaofo Chen , Zhiquan Wen, Qian Yue Wang , Zhijie Qiu , et al., editors. Curse of High Dimensionality Issue in Transformer for Long Context Modeling. Forty-second International Conference on Machine Learning; 2025; Vancouver, Canada: PMLR.
14. Tan KL, Lee CP, Anbananthan KSM, Lim KM. RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access.* 2022;10:21517-25.
15. Xu S, Li X, Xie C, Chen H, Chen C, Song Z. A High-Precision Implementation of the Sigmoid Activation Function for Computing-in-Memory Architecture. *Micromachines.* 2021;12(10).
16. Verma PK, Agrawal P, Amorim I, Prodan R. WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Trans Computl Soc Syst.* 2021;8(4):881-93.
17. Salman T, Ghubaish A, Unal D, Jain R. Safety Score as an Evaluation Metric for Machine Learning Models of Security Applications. *IEEE Netw Lett.* 2020;2(4):207-11.
18. Rama Moorthy H, Avinash NJ, Krishnaraj Rao NS, Raghunandan KR, Dodmane R, Blum JJ, et al. Dual stream graph augmented transformer model integrating BERT and GNNs for context aware fake news detection. *Sci Rep.* 2025;15(1):25436.
19. Hashmi E, Yayilgan SY, Yamin MM, Ali S, Abomhara M. Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI. *IEEE Access.* 2024;12:44462-80.
20. Raza N, Abdulkadir SJ, Abid YA, Albouq SS, Alwadain A, Rehman AU, et al. Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach. *PLoS One.* 2025;20(9):e0330954.

21. Trisna NJ, Sohan MFAA, Joy RM, Joy RD, editors. A Novel Machine Learning Approach for Detecting Fake News in WELFake Dataset. Proceedings of the International Conference on Industrial Engineering and Operations Management. 2024; p. 994-1002.
22. Al-Quayed F, Javed D, Jhanjhi NZ, Humayun M, Alnusairi TS. A Hybrid Transformer-Based Model for Optimizing Fake News Detection. IEEE Access. 2024;12:160822-160834.
23. Padalko H, Chomko V, Chumachenko D. A novel approach to fake news classification using LSTM-based deep learning models. Front Big Data. 2024;6(1):1320800.
24. Abbas F, Taeihagh A. A multi-level fusion-based framework for multimodal fake news classification using semantic feature extraction. Int J Mach Learn Cybern. 2025;16(1):6531–60.