International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 1071-1088
https://doi.org/10.70917/ijcisim-2026-3832
Article

Deep Learning-Based Fake News Detection in Social Media

Ali Mohammad Salloum[1], Bassel Alkhatib[1]

[1]Syrian Virtual University, Damascus Syria

Abstract: The proliferation of fake news on social media has introduced a major problem in society that is troubling public trust, political stability, health awareness and the reliability of information. Machine learning methods are inadequate in many aspects in the field of fake news classification, especially in modelling the long-range dependence and semantic meaning of text. Traditional machine learning algorithms are generally not suitable for modelling the long-range dependence and semantics of text, which is a limitation in the context of fake news classification. To overcome these drawbacks, a hybrid deep learning architecture using Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Long Short-Term Memory (BiLSTM) and Attention mechanism is proposed to ensure accurate and explainable identification of fake news. The proposed architecture is designed to leverage the contextual embedding generation, sequential dependency learning, and attention-based feature prioritization for boosting the classification performance. The experimental evaluation was performed on the two publicly available FakeNewsNet and ISOT Fake News datasets, which yielded a total of 38,994 news samples. The text quality was improved through several preprocessing techniques such as lowercasing, tokenisation, lemmatisation, stop word elimination, and noise filtering. The suggested model was verified using baseline models such as Support Vector Machine (SVM) and Logistic Regression. The experimental findings demonstrated that the suggested framework performs exceptionally well, with a 99.29% accuracy rate, a 99.74% precision rate, a 98.97% recall rate, a 99.35% F1-score, and a 99.33% ROC-AUC score. The robustness and generalization capability of the model were further confirmed by cross validation, ablation analysis, evaluation of confusion matrix, SHAP explainability and McNemar statistical testing. The results show that the embedding of transformer-based contextual learning and sequential deep learning architectures strongly improves the performance of fake news detection in social media.
Keywords: Fake News Detection, Deep Learning, BERT, BiLSTM, Attention Mechanism, Explainable AI, Natural Language Processing, Social Media Analytics, FakeNewsNet, ISOT Dataset.

1. Introduction

New media and social media communication have revolutionized the global information system. Thanks to sites like YouTube, Instagram, and Twitter, millions of individuals all around the globe have access to information. While these platforms help transmit messages

Email addresses: ali_319291@svuonline.org , slwm4767@gmail.com (Ali Mohammad Salloum), t_balkhatib@svuonline.org (Bassel Alkhatib)

and reach audiences, they've also helped to rapidly spread misinformation and fake news within online communities. Fake news is misleading or fabricated information that is intentionally spread to shape public opinion, polarize society and foster unrest within society (Sharma et al., 2019).

The consequences of fake news are social, political and economic, and this dissemination is a major concern in the world today. In many situations, like elections, the pandemic and economic crises, misinformation is disseminated by social media networks, and sometimes it reaches wide audiences before verification mechanisms can take action. As a result, fake news detection has become a crucial research field in the Natural Language Processing (NLP), Machine Learning (ML), deep learning and social media analytics (Islam et al., 2020).

Logistic Regression, Naïve Bayes, Random Forest, and SVM are some of the traditional machine learning techniques which have been widely used for fake news classification problems (Ahmad et al., 2020). While these methods are mainly based on handcrafted features and TF-IDF based representations, they fail to capture semantic relationships and contextual dependencies between text elements within textual data (Villela et al., 2023). The recent deep learning methods such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have enhanced the feature extraction and sequential learning ability (Umer et al., 2020). Moreover, BiLSTM models have been used for the forward and backward processing of text sequences to achieve better contextual understanding (Sridhar & Sanagavarapu, 2021).

The transformer-based architecture, especially that of BERT (Devlin et al., 2019), have been great leaps in natural language processing by generating embeddings and using self-attention mechanisms in language context. Over the last few years, several studies have been conducted that have demonstrated the use of BERT-based techniques for fake news detection with a positive result (Kaliyar et al., 2021; Oad et al., 2024). The significance of sequential relationships and local context in text data can be overlooked by transformer-based models. Another major issue with systems that detect fake news using deep learning is their lack of interpretability.

Therefore, in order to alleviate these shortcomings, the present study aims to innovate an efficient and interpretable fake news detection model based on stacking BERT with BiLSTM and Attention mechanism. The proposed architecture consists of three parts, contextual semantic learning, extracting sequential dependency in both directions, and attention-based feature prioritization in an end-to-end model. Experimental evaluation was carried out by using the openly available FakeNewsNet and ISOT Fake News datasets. The analysis was conducted by comparing the baseline machine learning models with the proposed framework, which showed that the proposed framework outperforms the baseline models. Furthermore, the robustness and generalization capacity of the proposed model were validated by cross validation, ablation analysis, SHAP analysis and statistical validation.

1.1 Research Objectives

The following are the main goals of the current study:

1. To use hybrid deep learning methods to get a robust structure for detecting fake news.
2. To combine the contextual embedding generation, sequential dependency learning, and attention-based feature prioritization in one architecture.
3. To evaluate how well the suggested hybrid deep learning framework performs in comparison to conventional machine learning models.
4. To assess the model's performance using common classification measures, such as ROC-AUC, F1-score, accuracy, precision, and recall.
5. To enhance model interpretability using explainable artificial intelligence techniques.

These goals are organised around the larger target of developing a system that can identify false news and improving the accuracy and reliability of disinformation classification on social media.

1.2 Research Motivation

Context, sequential dependency learning, generalisation, and interpretability continue to be problems for current false news detection systems. While some deep learning architectures struggle to describe both contextual semantics and sequential textual relations, the majority of conventional machine learning systems rely on manually created features. While transformer-based models have made significant strides in contextual representation learning, there still exists lack of effective embedding of contextual information, sequential learning, and attention-based interpretability mechanisms in the model, which is a major research issue. In this sense, the research work that is shown in the current work is motivated by the necessity of creating a powerful, accurate and transparent mixed structure that can improve the performance of the fake news classification and increase the transparency and usefulness of the model.

1.3 Major Contributions of the Study

The main contributions of this study are the following:

- A hybrid BERT–BiLSTM–Attention model is suggested for fake news detection.
- Integration of contextual semantic learning and sequential dependency extraction into a single architecture.
- An explainability analysis based on SHAP is included in order to increase the interpretability of the model.

- Comparative evaluation is conducted on a set of benchmark fake news datasets and baseline machine learning models.
- To assess the robustness and generalization capacity of the model, cross-validation, ablation study, and statistical validation are carried out.

The suggested contributions further enhance the efficacy, transparency and usability of deep learning fake news detection systems in practical social media environments.

1.4 Organization of the Paper

The remainder of the article is structured in the following manner: In Section 2, a review of the currently available literature and the research issues is presented. Detailed explanations of the approach and model architecture can be found in Section 3. Section 4 contains a discussion of the experimental data as well as an evaluation of the performance. Section 5 discusses the outcome of the suggested framework, as well as its advantages and drawbacks. In the sixth section, the findings of the study are discussed, and recommendations for ways to proceed with future research are offered.

2. Literature Review

In NLP, ML, and deep learning research, misinformation on online platforms has been the main focus. There are several approaches to detecting fake news, ranging from sophisticated transformer-based structures to conventional machine learning models. Most of the existing studies have concentrated on enhancing the classification accuracy, contextual understanding and ability of the model to generalize. Still, there are challenges of semantic representation and sequential dependency learning, interpretability, and cross-dataset robustness.

2.1 Machine Learning-Based Fake News Detection

The binary classification of fake and fake news stories in previous research on fake news detection often used logistic regression, Naïve Bayes, Random Forest, and SVM. Ensemble machine learning methods were used by Ahmad et al. (2020) to classify fake news and their predictive power was found to be better than the individual classifiers. Likewise, Sharma et al. (2019) gave an exhaustive survey of techniques to identify fake news and emphasized that linguistic and social-contextual features play an important role in fake news detection.

For feature extraction from text, the most often used machine learning models are based on TF-IDF, Bag-of-Words and n-gram representations. These methods work well on tabular data, but require manual feature engineering to grasp the semantics and long-range relationships in text data. Villela et al. (2023) conducted a systematic survey of fake news detection methods and found that traditional machine learning models tend to exhibit low scalability and low adaptability to different datasets and evolving social media environments.

2.2 Deep Learning-Based Approaches

Advances in deep learning in the last several years have greatly enhanced automatic text classification through the acquisition of hierarchical features and sequential relationships. Local textual patterns have been widely extracted using CNNs while the ability of the LSTM network in modeling sequential information in textual sequences has been shown.

To tackle the problem of fake news stance detection, Umer et al. (2020) introduced a CNN-LSTM based deep learning approach and achieved better sequential learning performance than traditional machine learning techniques. In a similar fashion, Goldani et al. (2021) applied the CNN-based framework to a margin loss objective for fake news detection with a view to better classification capability by deep feature learning. Padalko et al. (2024) also showed the ability of LSTM-based models to capture temporal and contextual relationships from the news text content.

BiLSTM models have been a significant focus as well because they can easily process textual sequences in both directions. Sridhar and Sanagavarapu (2021) have suggested a multitask learning approach that combines BiLSTM and CapsNet models for fake news detection and analysis. They found that it was beneficial to bidirectional sequence learning for contextual understanding and classification performance. However, recurrent architectures by themselves might still not be able to successfully determine which words have a high level of information that affect classification decisions.

2.3 Transformer-Based Models

In NLP, transformer architectures have revolutionized the field with the generation of contextual embeddings and self-attention mechanisms. The transformer architecture introduced by Vaswani et al. (2017) greatly enhanced the modeling of sequences and contextual representation learning in NLP applications. Transformer architectures have been leveraged by Devlin et al., (2019), who created BERT via bidirectional encoding of embeddings for textual representations.

In recent years, several studies have been conducted to perform tasks related to fake news detection using BERT-based architecture. Raza et al. (2025) performed comparative analysis of the BERT-type models and large language models for fake news classification and pointed out the increasing success of the transformer-based contextual learning method in misinformation detection. To detect social media fake news, Kaliyar et al. (2021) introduced a deep learning framework, named FakeBERT, which is based on BERT. They found significant accuracy enhancements in the classification of fake news. Angizeh and Keyvanpour (2024) evaluated and contrasted the performance of both the BERT and RoBERTa models for misinformation classification and showed that the contextual embedding-based models are more successful at capturing the semantic relationships in the news content. Oad et al. (2024) also developed the methodology of an improved BERT model combined with advanced textual representation techniques to further boost the classification effectiveness of fake news.

Transformer models offer good contextual understanding but have high computational complexity and can miss out on the sequential dependencies that recurrent models can capture well. Farokhian et al. (2022) further noted that transformer-only models may need supplementary architectures to enhance the ability to prioritize context and also the efficiency of sequential learning.

2.4 Hybrid Deep Learning Approaches

Recognizing the shortcomings of single machine learning models and transformer based models, many research efforts have investigated hybrid deep learning architectures that integrate contextual embeddings with sequential learning. To achieve better performance in fake news detection, Nasir et al. (2021) introduced CNN-RNN hybrid architecture to produce local features and sequential learning. Likewise, Tuan and Minh (2021) have combined BERT with Attention mechanism for multi-modal fake news detection, and showed the efficacy of attention-based feature prioritization. Furthermore, Yang et al. (2023) proposed a multi-modal transformer network for fake news detection, and identified that by leveraging multiple information modalities, the context understanding and fake news classification accuracy are improved.

Over the last several years, there have been many studies that integrate transformer architectures with recurrent neural networks (RNNs), aiming to enhance the robustness of the classification task and the understanding of the context. Roumeliotis et al. (2025) made a comparative study among Convolutional neural networks, Natural language processing models and large language models for fake news detection and highlighted the need for a hybrid architecture to increase the reliability of fake news classification and context understanding. Ayyasamy et al. (2025) presented a hybrid approach of deep learning, which integrates LSTM architectures and optimization strategies in the field of misinformation detection. In textual misinformation classification tasks, Cao et al. (2025) proposed a multi-stacked LSTM (MS-LSTM) framework based on DistilBERT and showed its effectiveness. Similarly, Ali et al. (2025) used transformer-based embedding to enhance the accuracy and generalization ability of fake news detection models, by incorporating metadata-related learning models.

However, there are still various architectures of hybrid solutions that are yet to overcome challenges in explainability, computational efficiency, and cross-dataset generalization. Numerous studies focus on accuracy of classification and offer little information on how salient textual features relate to the results of the predictions.

2.5 Explainable AI in Fake News Detection

An increasingly crucial element of deep learning fake news detection systems is interpretability. While more sophisticated deep learning architectures are able to classify with a high accuracy, their black-box design often decreases transparency and reliability in actual application. Explainable Artificial Intelligence (XAI) techniques hence become crucial in the enhancement of model explainability and user trust.

In recent years, attention mechanisms and explainability frameworks have been used to identify the most relevant textual features that influence the classification decisions. Abduljaleel and Ali (2025) used BERT embeddings along with attention mechanism for fake news detection and emphasized the importance of attention-based feature selection of enhancing the interpretability. Likewise, Maham et al. (2024) highlighted the need of well-designed and interpretable architectures for enhancing the reliability of misinformation classification in the presence of adversarial scenarios.

Although a number of studies have been conducted on explainable AI, there has been a small number of studies that have completely connected embedding generation for context, sequential dependency learning, attention-based feature prioritization, and explainability analysis in one system for fake news detection.

2.6 Research Gaps and Challenges

The literature review shows that current methods for detecting fake news still have some significant drawbacks. The majority of conventional machine learning models rely on superficial textual representations and manually created feature engineering, which are not effective from semantic and contextual relationships. The deep learning models enhance the ability to extract features, yet it is hard to capture the contextual semantics and the long-range sequential dependencies at the same time with a single CNN or LSTM.

While transformer-based architectures like BERT have been successful in representing contextual information, transformer-based models might fail to capture sequential dependencies at the local level or feature importance at the sentence level in text. Moreover, many deep learning models for detecting misinformation lack interpretability. Prior research also shows reliance on particular datasets which limits the generalizability to other social media settings.

Hence, a strong hybrid system able to combine contextual embedding generation, sequential dependency learning in both directions, feature prioritization based on attention, and explainable AI in a single system is still needed. This study tackle these difficulties by designing a hybrid model of BERT combined with BiLSTM and Attention networks for fake news detection that is accurate, robust and explainable.

3. Research Methodology

In the current research, a novel approach has been put forth that leverages a deep learning model for the purpose of detecting fake news on social media. This proposed system utilizes transformer network-based contextual embedding, bidirectional sequential analysis, and attention-based feature selection. It is important to mention that the framework has been devised keeping in view the strengths and weaknesses of machine learning and deep learning models separately.

The entire research process was divided into five steps: data collection, data preprocessing, generation of contextual embeddings, development of the hybrid model, training of the models, comparative analysis, explainability analysis, and statistical validation. To guarantee the reliability and repeatability of the experiments, publicly available benchmark datasets were used.

*3.1 **Dataset Strategy***

For the experimental evaluation, two publicly accessible benchmark datasets that are frequently utilised in studies on false news identification were utilised:

1. FakeNewsNet Dataset
2. ISOT Fake News Dataset

FakeNewsNet (Shu et al., 2020) is a set of news content and contextual information gathered from social media environments related to fake and real news dissemination. ISOT Fake News Dataset consists of a collection of sample textual news data that are commonly used in misinformation classification related problems. The fusion of these datasets increased the diversity of the datasets and increased the generalization ability of the presented framework. Table 1 shows information about the datasets used in the investigation.

Table 1. Summary of Benchmark Datasets Utilized for Fake News Detection

Dataset	**Description**	**Approximate Samples**
FakeNewsNet	Social media fake and real news dataset	~20,000
ISOT Fake News Dataset	Benchmark textual fake news dataset	~44,000

3.2 Data Description

After data integration and pre-processing, the final data set contained 38,994 text news samples, including fake and real news samples. The data set was split into 2 classes: label "0" (phoney news) and label "1" (authentic news). Given the data set, the final distribution of the data is shown in Table 2.

Table 2. Class-wise Distribution of Fake and Real News Samples in the Final Dataset

Label	Description	Count
0	Fake News	17,603
1	Real News	21,391

A class distribution analysis was done as a way of looking at a balance between the two classes. The same thing is depicted in figure 1.

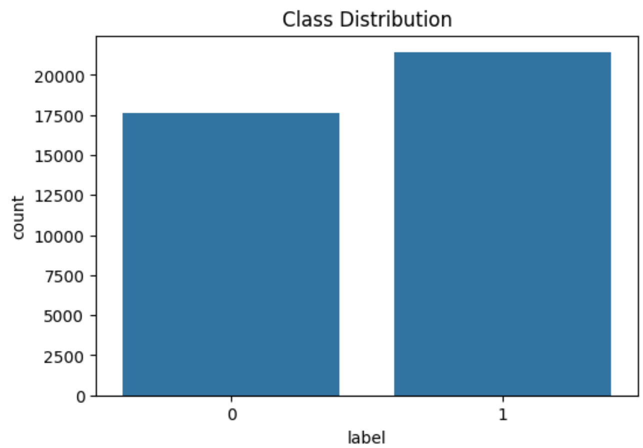

Figure 1. Distribution of Fake and Real News Samples within the Combined Dataset

The class distribution was relatively balanced, both fake news and real news, which reduced the chances of the classification bias in the process of training and testing the models.

3.3 Data Preprocessing

In the context of unstructured textual data, textual preprocessing is very important to enhance semantic consistency and noise reduction. To ensure better quality of text and representation learning in context, several preprocessing techniques were employed prior to training the model. The following steps were done in the preprocessing pipeline:

- Lowercasing
- Stopword removal
- Tokenization
- Lemmatization
- URL removal
- Punctuation removal
- Noise filtering

All the text has been lowercased so that subsequent words are textually consistent, and all text has been tokenized to decompose the sentences into meaningful lexical units. Lemmatization was used to normalize the words to their base forms to ensure semantic consistency. In addition, any URLs, punctuation marks, or noisy text patterns were stripped from the data in an effort to remove irrelevant data. The pre-processing used in misinformation detection studies has been extensively used to enhance the quality of the text features (Islam et al., 2020).

For the pre-processing operations, the following libraries have been used:

- NLTK
- SpaCy
- Pandas
- NumPy

The pre-processing stage significantly enhanced the textual normalisation and minimized unnecessary linguistic variations prior to generating contextual embedding.

3.4 Model Design

To evaluate the efficacy of the suggested framework, baseline machine learning models and hybrid deep learning architectures were used. An analysis of improvements made with the integration of contextual embedding and sequential dependency learning was performed by using comparative analysis.

❖ **Baseline Models**

The following baseline models were implemented for performance comparison:

1. Logistic Regression
2. SVM
3. CNN
4. LSTM

These conventional machine learning models—SVM and logistic regression—were selected, as they have been well used in text classification. CNN and LSTM architectures were incorporated into assess the performance of deep learning-based feature extraction and sequential learning mechanism.

❖ **Proposed Hybrid Framework**

The proposed model integrates the following major components:

- BERT
- BiLSTM
- Attention Mechanism

The hybrid architecture aimed to integrate contextual semantic comprehension and sequential dependency extraction and feature-level prioritization.

❖ **BERT-Based Contextual Embedding**

The BERT Base Uncased architecture proposed by Devlin et al., 2019 was used for the contextual embedding layer. BERT is based on transformer-based bidirectional self-attention mechanisms to help learn semantic representations in textual data. Since BERTs are able to learn contextual relationships between words (Kaliyar et al., 2021), they have been used to successfully solve fake news classification tasks. In this study, BERT embeddings were used as feature representations for a sequential feature extraction.

❖ **Bidirectional Sequential Learning**

Bidirectional Long Short-Term Memory (BiLSTM) networks were integrated into the system to simulate both forward and backward sequential relations in the text. BiLSTM models processes text information in both directions unlike traditional LSTM models, which enhances its ability to capture meaning and context.

The BiLSTM model is capable of learning the sequential information, which facilitates the extraction of temporal information and sentence-level context information from news content.

❖ **Attention Mechanism**

The Attention mechanism was used to highlight words with high information content that play an important role in the classification results. Attention-based feature prioritization enables to use influential textual features more during prediction, thus enhancing interpretability of the model.

Self-attention idea in transformer network models has proven to be effective for contextual feature learning and textual priority (Vaswani et al., 2017). The proposed framework aims to improve contextual relevance learning and classify efficiently, using the Attention layer.

3.5 System Architecture

The overall structure of the proposed framework combines the three components: contextual embedding generation, bidirectional sequential learning and attention-based feature extraction, and puts them into a classification pipeline. The flow of processing for the proposed system architecture is shown in the figure 2.

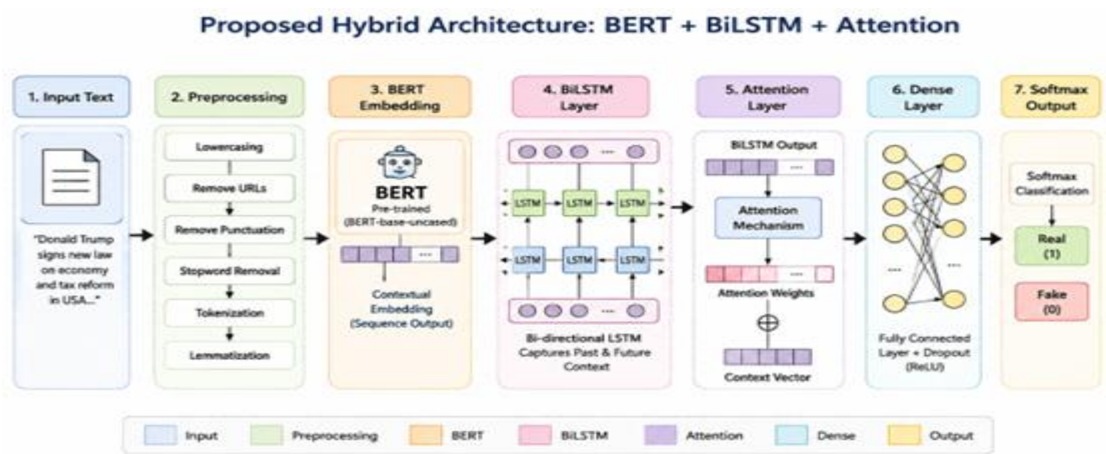

Figure 2. Architecture of the Proposed BERT–BiLSTM–Attention Fake News Detection Framework

Text data is first preprocessed and tokenized, and then fed into the BERT embedding layer. The embeddings are then passed through the BiLSTM layer, which extracts sequential dependency information, and to the final layer.The contextual embeddings are then fed into the BiLSTM layer to learn the sequential information and into the final layer. The Attention mechanism is used to learn attention weights over text features that are influential for classification. Finally, the extracted representations are fed into dense classification layers in order to classify fake and real news into binary classes.

3.6 Implementation Environment

The suggested structure was realized in Python program language and in the cloud-based and notebook-based development environment. Google Colab and Jupyter Notebook platforms were used for the implementation, as the platforms offer flexibility in computations and support deep learning libraries.

Several commonly used libraries and frameworks were used in its implementation to enable data processing, model construction, assessment and visualisation tasks. Among these were PyTorch and TensorFlow for deep learning models, Scikit-learn for machine learning processes, Pandas and NumPy for data manipulation and numerical computing, and Matplotlib and Seaborn for graphical representation and analysis. Furthermore, transformer-based natural language processing models were used with HuggingFace Transformers and SHAP was used to increase the interpretability and explainability of the models.

For generating contextual embeddings, a HuggingFace Transformers library has been used with pre-trained BERT models, while for analysis and interpretation of explainability and feature importance, SHAP has been used.

3.7 Training Strategy

To guarantee stable convergence, effective optimization and robust generalization capability, the model training process was designed. The dataset was divided into training and testing portions at an 80:20 ratio. The training configuration chosen for the proposed framework is shown in table 3.

Table 3. Training Hyperparameter Configuration of the Proposed Hybrid Framework

Parameter	Value
Train-Test Split	80:20
Batch Size	16
Epochs	3
Optimizer	AdamW
Learning Rate	2e-5
Embedding Model	BERT Base Uncased

The AdamW optimizer was used because of its efficiency in optimizing transformers and weight regularization. This is because a 2e-5 learning rate was chosen for stable fine-tuning of the pretrained BERT architecture.

For the purpose of validating the robustness and the generalization capability of the proposed framework, the following validation strategies were implemented:

- Cross-validation.
- Comparing the performance to the baseline models.
- The statistical validation of data was carried out with McNemar Test.

These evaluation methods allowed to assess the consistency of the classifications, the stability of the models and their predictive reliability in various experimental situations.

4. Performance Evaluation and Experimental Results

The performance of the proposed hybrid BERT–BiLSTM–Attention framework was assessed based on various measures, comparative analysis, explainability assessment and statistical validation techniques. The experimental analysis focused on evaluating the classification accuracy, robustness, convergence behavior, interpretability and generalization capability of the proposed architecture. A comparative evaluation between the baseline machine learning models was also conducted for assessing the improvements obtained by incorporating contextual embedding and sequential dependency learning.

4.1 Evaluation Metrics

For the purpose of thorough examination of the proposed framework, a number of standard classification metrics were used. These metrics allowed a detailed evaluation of the ability to predict, classify consistency, and model reliability. The experimental model has been tested with various performance evaluation metrics and proven to be effective and reliable. They were accuracy for overall prediction performance, precision and recall for class-specific evaluation, F1-score for balancing precision and recall, ROC-AUC for measurement of the classification capability over thresholds, and the confusion matrix for detailed analysis of prediction results.

Overall classification correctness was evaluated using accuracy while, the precision and recall metrics assess how well the algorithm can distinguish between bogus and authentic news. ROC-AUC was used to measure the discriminatory power of the classifier over a set of decision thresholds and F1-score was used to measure a balanced evaluation of precision and recall.

4.2 Baseline Model Performance

Due to their widespread use in textual classification and false news detection tasks, two popular classifiers—Logistic Regression and Support Vector Machine (SVM)—were constructed in order to create comparison benchmarks. Table 4 shows the comparative results of the basic models' performance.

Table 4. Comparative Accuracy Performance of Baseline Machine Learning Models

Model	Accuracy

Logistic Regression	97.78%
Support Vector Machine	98.17%

The experimental results indicated that slightly high classification accuracy for SVM model was achieved as compared to Logistic Regression. The improvement of SVM is due to its capability in processing high dimensional text feature space and high-curvature classification boundary. Both the baseline models were able to classify well, but not up to the level of the proposed hybrid deep learning framework. This is a good example of the fact that traditional machine learning methods are not suitable to model the complex semantics of the context and long-range dependencies of the text.

4.3 Performance Evaluation of the Proposed Hybrid Framework

The suggested BERT - BiLSTM - Attention architecture was evaluated using many classification criteria for the fake news detection problem. Overall, the combination of contextual embedding generation, sequential dependency extraction and attention-driven feature prioritization was used to great effect for overall classification. The results of the proposed framework are shown in Table 5.

Table 5. Performance Evaluation Metrics of the Proposed BERT–BiLSTM–Attention Framework

Metric	Value
Accuracy	99.29%
Precision	99.74%
Recall	98.97%
F1-Score	99.35%
ROC-AUC	99.33%

The framework proposed yielded very high performance in all evaluation parameters, which shows a very high classification ability and consistency in prediction.

High values of Accuracy, Precision, Recall and F1 score reveal the effectiveness of the proposed architecture in correctly classifying fake and real news articles with minimum classification error. Further, the proposed framework shows good discriminatory power throughout the various classification thresholds as reflected in the ROC-AUC score of 99.33%.

Confusion matrix analysis was done to assess the consistency of classification further. The confusion matrix visualization for the corresponding is presented in Figure 3.

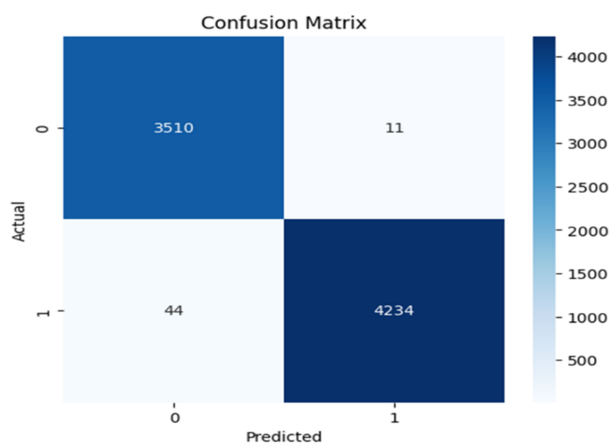

Figure 3. Confusion Matrix Representing Classification Outcomes of the Proposed Framework

The confusion matrix showed very good classification performance, accuracy of 3510 samples of fake news correctly identified and 4234 samples of real news correctly identified. Upon Prediction, 11 false positives and 44 false negatives were seen. Minimal misclassification values confirm the robustness and reliability of the proposed framework of binary fake news classification task.

The accuracy changes over epochs were analyzed during training to study the behavior of the models as they converge. The visualization is given for the corresponding in Figure 4.

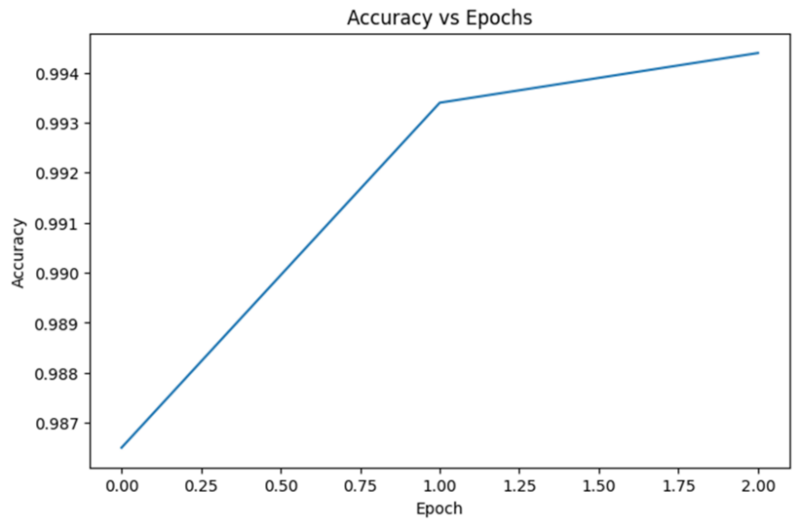

Figure 4. Training Accuracy Progression Across Successive Epochs

The training accuracy steadily improved across the epochs, suggesting a steady learning process and optimizing performance. The proposed framework was able to learn contextual and sequential textual representations with a good efficiency as the model converged quickly and with little variation. Convergence behavior of losses was also analyzed during model training, in addition to accuracy analysis. Loss visualization corresponding to the above loss is displayed in Figure 5.

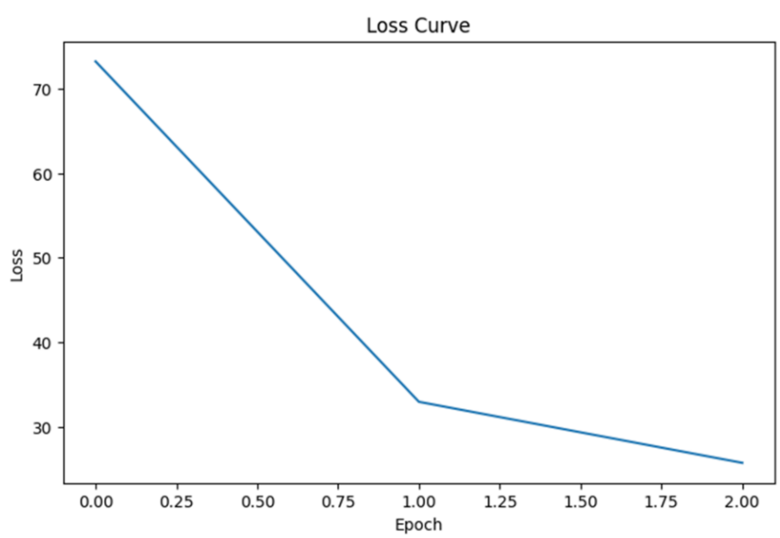

Figure 5. Training Loss Convergence Behavior During Model Optimization

The training loss gradually decreased with epochs, suggesting effective optimization and stable parameter learning. The progressive decrease in loss values shows that the proposed architecture has been effective in minimizing the prediction error during training with no unstable convergence behavior.

4.4 Explainability Analysis Using SHAP

Finding out which textual features significantly impacted the categorisation decisions was the goal of the SHAP (SHapley Additive exPlanations) analysis, which aimed to increase interpretability and transparency. One important part of systems that detect fake news is explainability analysis, since deep learning architectures are black boxes. Shown in Figure 6 is the visualisation of the SHAP feature relevance.

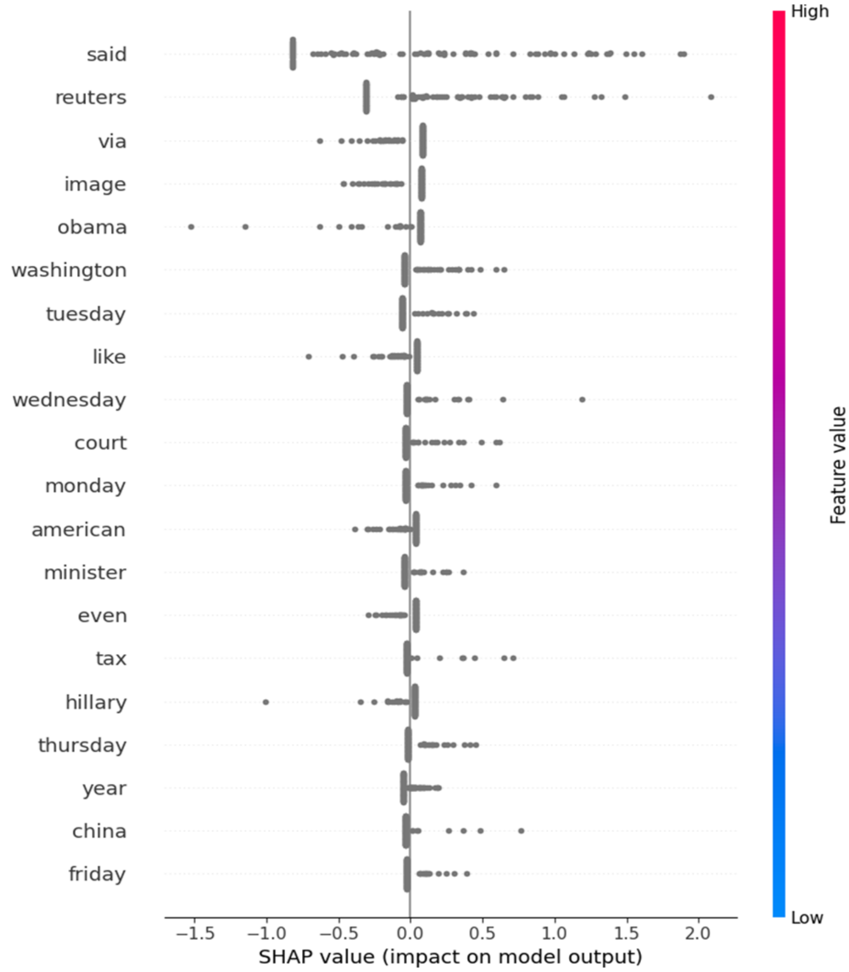

Figure 6. SHAP-Based Feature Importance Visualization for Model Explainability

The SHAP analysis revealed multiple important text characteristics, such as the words “Reuters,” “Obama,” “Washington” and “Hillary,” that strongly impacted the classification results. Positive and negative SHAP values suggested that how much words contributed to the fake or real news prediction. The explainability analysis, within the proposed framework, was used to improve the transparency of models, offering interpretable explanations of the models' feature-level decision behavior.

4.5 Cross-Validation Analysis

A k-fold cross-validation analysis was run to evaluate the stability and generalizability of the proposed framework. To evaluate the performance on several divisions of the dataset and minimize the potential bias in the evaluation, cross-validation was used. The results obtained for cross validation on the folds are provided in Table 6.

Table 6. K-Fold Cross-Validation Accuracy Results of the Proposed Framework

Fold	Accuracy
Fold 1	97.83%
Fold 2	98.09%

Fold 3	98.04%
Fold 4	97.74%
Fold 5	98.04%

The results of the cross validation showed high consistence between the different folds and accuracy values were close to each other during all cross validation runs. This average cross-validation accuracy of 97.95% further validates that the proposed hybrid architecture has a good generalization capability and predictive stability.

4.6 Ablation Study

The BERT, BiLSTM, and Attention modules were analysed separately based on different combinations in an ablation study to evaluate the contribution of each architectural component for the overall classification performance. The ablation study's findings are displayed in Table 7. The BERT, BiLSTM, and Attention modules were examined separately using different combinations. Table 7 displays the ablation study's findings.

Table 7. Ablation Analysis of Individual Components within the Proposed Hybrid Architecture

Model Variant	**Accuracy**
BERT Only	93.00%
BERT + BiLSTM	95.00%
BERT + Attention	94.00%
BERT + BiLSTM + Attention	99.29%

The results of ablation analysis clearly showed that the classification performance of the model using BiLSTM and Attention mechanisms was better than that of the model using only the BERT based model. The full hybrid structure recorded the best accuracy, which confirmed the efficiency of the hybrid method, integrating the generation of contextual embedding with the sequential dependency modeling and attention-driven feature prioritization.

4.7 Statistical Validation

The classification results were further validated with statistical evaluation using McNemar Test to ensure the reliability of the results. To analyze the consistency between the predicted and actual classification results, statistical validation was carried out. The statistical validation results can be seen in Table 8.

Table 8. Statistical Validation Results Using McNemar Test

Statistic	**Value**
McNemar Statistic	5.0
p-value	0.3017

There was consistency between the observed classification result and predicted result as evidenced by the McNemar statistical analysis. The resulting p-value showed that classification performance in both experimental evaluation settings was statistically reliable, thus validating the robustness and stability of the proposed framework.

5. Discussion

Based on the experimental results it can be seen that the proposed BERT–BiLSTM–Attention model is effective in detecting fake information on social media. This was achieved by the model's ability to learn the semantic links and context dependencies which were important to the specific task involved, through the use of transformer-based contextual embeddings, bidirectional sequential learning and attention-driven feature prioritisation.This combination of features was very good in capturing the semantic links and context-dependencies that were relevant to the specific task involved. The results demonstrate the reliability, stability and predictability of the proposed architecture.

5.1 Comparative Analysis with Existing Studies

The existing fake news detection approaches emphasized mostly hand-crafted textual features and shallow representations, neglecting the context of the news and generalization capabilities (Ahmad et al., 2020; Villela et al., 2023). There have been recent attempts on deep learning that gained sequential learning and feature extraction, but there were some works with a limited attention span in the context of either contextual representation learning or sequential dependency extraction.

The transformer-based models achieved better results in learning contextual semantic meaning (Kaliyar et al., 2021) while the CNN-LSTM models and hybrid models demonstrated better results in learning sequential dependency (Umer et al., 2020; Nasir et al., 2021). However, there were some limitations such as lack of explainability and feature prioritization in many of the existing frameworks.

To overcome these, the proposed study introduced a combination of BERT, BiLSTM, and Attention mechanism into a single model. Furthermore, SHAP explainability analysis allowed to enhance the transparency and interpretability of the model.

Table 9. Comparative Analysis of Existing Fake News Detection Studies and the Proposed Framework

Study	**Methodology**	**Observed Limitation**	**Advantage of Proposed Study**
Ahmad et al. (2020)	Ensemble Machine Learning	Handcrafted feature dependency	Improved contextual and sequential learning
Umer et al. (2020)	CNN-LSTM	Limited contextual semantic learning	Transformer-based embedding integration
Kaliyar et al. (2021)	FakeBERT	Limited sequential dependency extraction	Integration of BiLSTM and Attention
Nasir et al. (2021)	CNN-RNN Hybrid	Limited explainability	SHAP-based interpretability
Oad et al. (2024)	Enhanced BERT	Limited feature prioritization	Attention-driven feature learning
Proposed Study	BERT–BiLSTM–Attention	Computational complexity	High accuracy with improved interpretability

It shows from the comparative analysis that the proposed framework is able to effectively integrate contextual semantic learning, sequential dependency extraction, and explainability into a single framework to enhance the accuracy of the classification and the interpretability of the models.

5.2 Interpretation of Findings

The proposed framework shows an advantage on its performance, which is due to the complementary use of BERT, BiLSTM and Attention mechanisms. BERT was used to boost the representation of contextual semantics, and BiLSTM was used to boost the learning of bidirectional sequential dependency. The Attention mechanism also enhanced the classification accuracy by highlighting relevant words in the textual features that would have a strong impact on the classification results.

The SHAP explainability analysis provided insights into the importance of words that influenced classification decisions, improving transparency. Moreover, Ablation study showed that using only BERT-based network led to inferior performance and combining BiLSTM and Attention mechanisms resulted in better results. Furthermore, cross validation and McNemar statistical testing were used to validate the stability and generalization ability of the proposed framework.

5.3 Theoretical and Practical Implications

The present work makes a contribution to the field of fake news detection research by showing the effectiveness of the integrated framework of contextual embedding generation, sequential dependency learning, and attention-based features prioritization.

In practice the proposed framework could be implemented in the following ways:

- Social media monitoring systems
- The internet news verification sites.
- Automated systems for the filtering of misinformation
- Cybersecurity intelligence applications

The SHAP explainability analysis further increases transparency and reliability of automated fake news detection systems.

5.4 Strengths and Limitations

The proposed framework has the following major strengths:

- High classification accuracy
- Comprehend some context-based semantics
- Sequential dependency (effective) learning
- Attention-driven feature prioritization
- Explainable AI integration
- Strong test of validation using cross validation and statistical testing.

While the strengths are identified, there are some limitations identified as well. The framework is complex with respect to the complexity of the computation, because the framework is based on transformer-based learning, and mainly focuses on textual information (not multimodal information such as images or user interactions). Moreover, the model was shown on benchmark datasets which may not represent the dynamics of misinformation in the real-time social media setting.

6. Conclusion and Future Work

This study proposed a hybrid deep learning model that integrates attention processes with two models, BiLSTM and BERT, to detect bogus news in social media with good accuracy and explainability. The suggested architecture effectively combined attention-driven feature prioritisation, bi-directional sequential dependency extraction, and contextual semantic representation learning into a unified framework. This embedding allowed the model to understand the context relations and other relevant aspects that are needed for distinguishing between fabricated and legitimate news.

To test the proposed approach, FakeNewsNet and ISOT Fake News datasets have been utilized. The results of applying the hybrid BERT-BiLSTM-Attention method were found to be more accurate compared to baseline machine learning approaches. The model demonstrated good predictive power and classification stability with exceptional accuracy of 99.29%, very high precision of 99.74%, high recall of 98.97%, F1-score of 99.35%, and ROC-AUC score of 99.33%.

The results also showed that combination of BiLSTM and Attention mechanisms outperform the transformer-based network architectures alone. Additionally, analysis through cross-validation, ablation analysis, confusion matrices, and McNemar statistics further helped to establish the stability, reliability and generalization ability of the presented framework. Furthermore, applying SHAP explainability analysis enhanced the transparency of the model in identifying textural features that were critical in making classifications.

This paper contributes significantly to the field of fake news detection by providing a robust and interpretable hybrid deep learning architecture that enhances context understanding and sequence learning. The proposed approach will also find great practical application in social media monitoring, misinformation filtering, news verification on online platforms, and cybersecurity intelligence.

Even though the presented system performs excellently, there are still several limitations regarding the computational complexity and information employed in the model (textual information only). Directions for future research include:

- Multimodal fake news detection using textual and visual information.
- Systems that can detect disinformation on social media in real-time.
- Light weight transformer architectures for efficient deployment.
- Fake news detection frameworks (in crosslingual and multilingual settings).
- Advanced explainable AI techniques for better interpretability and transparency

In conclusion, the current study's findings demonstrate how transformer-based contextual embedding generation, sequential deep learning, and attention-based processes can greatly improve fake news detection performance in dynamic social media scenarios.

References

1. Abduljaleel, I. Q., & Ali, I. H. (2025). Detecting Fake News Using BERT Word Embedding, Attention Mechanism, Partition and Overlapping Text Techniques. TEM Journal, 14(2).
2. Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. Complexity, 2020(1), 8885861.
3. Ali, A., Haider, Z., Shahbaz, H., Masood, H., Gul, B., Rashid, A., ... & Fitriyani, N. L. (2025). Towards improved fake news detection using a hybrid RoBERTa and metadata enhanced XGBoost model. Scientific Reports.
4. Angizeh, L. B., & Keyvanpour, M. R. (2024, April). Detecting fake news using advanced language models: BERT and RoBERTa. In 2024 10th international conference on web research (ICWR) (pp. 46-52). IEEE.
5. Ayyasamy, R. K., Ponnusamy, C., Bhargavi, K. N., Cherukuvada, S., Babu, G. C., Amutha, S., & Gamu, D. T. (2025). A hybrid deep learning framework for fake news detection using LSTM-CGPNN and metaheuristic optimization. Scientific Reports, 15(1), 41522.
6. Cao, P., Shukla, P. K., Shukla, P. K., Bhatia Khan, S., Alojail, M., & Ramtiyal, B. (2025). Enhancing fake news detection for Sustainable Supply Chain Management using DistilBERT-based multi-stacked LSTM approach. Enterprise Information Systems, 19(9), 2538023.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).
8. Farokhian, M., Rafe, V., & Veisi, H. (2022). Fake news detection using parallel BERT deep neural networks. arXiv preprint arXiv:2204.04793.
9. Goldani, M. H., Safabakhsh, R., & Momtazi, S. (2021). Convolutional neural network with margin loss for fake news detection. Information Processing & Management, 58(1), 102418.
10. Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. Social Network Analysis and Mining, 10(1), 82.
11. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia tools and applications, 80(8), 11765-11788.
12. Maham, S., Tariq, A., Khan, M. U. G., Alamri, F. S., Rehman, A., & Saba, T. (2024). ANN: adversarial news net for robust fake news classification. Scientific Reports, 14(1), 7897.
13. Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International journal of information management data insights, 1(1), 100007.
14. Oad, A., Farooq, M. H., Zafar, A., Akram, B. A., Zhou, R., & Dong, F. (2024). Fake news classification methodology with enhanced bert. IEEE Access, 12, 164491-164502.
15. Padalko, H., Chomko, V., & Chumachenko, D. (2024). A novel approach to fake news classification using LSTM-based deep learning models. Frontiers in big Data, 6, 1320800.
16. Raza, S., Paulen-Patterson, D., & Ding, C. (2025). Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data. Knowledge and Information Systems, 67(4), 3267-3292.
17. Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models. Future Internet, 17(1), 28.

18. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. ACM transactions on intelligent systems and technology (TIST), 10(3), 1-42.
19. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data, 8(3), 171-188.
20. Sridhar, S., & Sanagavarapu, S. (2021, January). Fake news detection and analysis using multitask learning with BiLSTM CapsNet model. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 905-911). IEEE.
21. Tuan, N. M. D., & Minh, P. Q. N. (2021, August). Multimodal fusion with BERT and attention mechanism for fake news detection. In 2021 RIVF international conference on computing and communication technologies (RIVF) (pp. 1-6). IEEE.
22. Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B. W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). IEEE Access, 8, 156695-156706.
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
24. Villela, H. F., Corrêa, F., Ribeiro, J. S. D. A. N., Rabelo, A., & Carvalho, D. B. F. (2023). Fake news detection: a systematic literature review of machine learning algorithms and datasets. Journal on Interactive Systems, 14(1), 47-58.
25. Yang, P., Ma, J., Liu, Y., & Liu, M. (2023). Multi-modal transformer for fake news detection. Math. Biosci. Eng, 20(8), 14699-14717.