

Enhanced Fake News Detection Using Hybrid DistilBERT and BiLSTM Architecture: A Comprehensive Study on the ISOT Dataset

Amrutha Sridhanya Madduri^{1,*}, Soundarya Suresh², G. P. Abhilash³, Abdulrahman Magdy Ragab Abdelshakour Azzam⁴, Alaa Mohamed Gaber Hussein⁵, Anass Esam Mohamed Mohamed Elsakka⁶

^{1,2,3}Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.

^{4,5,6}Department of Biomedical Sciences, Zewail City of Science and Technology, 6th of October City, Giza Governorate, Egypt.

as0084@srmist.edu.in¹, ss6202@srmist.edu.in², pp4087@srmist.edu.in³, s-abdulrahman.ragab@zewailcity.edu.eg⁴, s-alaa.hussein@zewailcity.edu.eg⁵, s-anass.mohamed@zewailcity.edu.eg⁶

*Corresponding author

Abstract: The rapid spread of misinformation across digital platforms necessitates the development of accurate and efficient fake news detection systems. This study proposes a context-sensitive hybrid deep learning framework that integrates DistilBERT-based transformer embedding with a Bidirectional Long Short-Term Memory (BiLSTM) network to capture both contextual semantics and sequential dependencies in textual data. The model is trained on the ISOT dataset and evaluated using a comprehensive set of performance and agreement metrics. Experimental results demonstrate that the proposed model achieves high classification performance, with an accuracy of 0.9993, precision of 0.9997, recall of 0.9990, and F1-score of 0.9994. Structural evaluation further confirms the model's reliability, with an Adjusted Rand Index (ARI) of 0.9974, a Normalised Mutual Information (NMI) of 0.9924, and a Cohen's Kappa of 0.9987, indicating near-perfect agreement between predicted and actual labels. However, cross-dataset evaluation on the LIAR dataset reveals a significant performance drop, with accuracy of 0.6429, precision of 0.275, recall of 0.003, and F1-score of 0.005, highlighting limitations in domain generalization. An ablation study validates the effectiveness of the hybrid architecture, showing improved performance compared to individual DistilBERT and BiLSTM models, and demonstrates the impact of input sequence length optimization on both accuracy and computational efficiency. Overall, the proposed framework achieves state-of-the-art results in in-domain evaluation while identifying critical challenges in cross-domain robustness, underscoring the need for further research on generalizable fake news detection systems.

Keywords: Detection of Fake News; Deep Learning; Normalized Mutual Information; Natural Language Processing; Text Classification; Transformer Model; AdamW Optimization; Evaluation Metrics.

Cite as: A. S. Madduri, S. Suresh, G. P. Abhilash, A. M. R. A. Azzam, A. M. G. Hussein, and A. E. M. M. Elsakka, "Enhanced Fake News Detection Using Hybrid DistilBERT and BiLSTM Architecture: A Comprehensive Study on the ISOT Dataset," *AVE Trends in Intelligent Computer Letters*, vol. 2, no. 2, pp. 122–137, 2026.

Journal Homepage: <https://avepubs.com/user/journals/details/ATICL>

Received on: 20/02/2025, **Revised on:** 19/04/2025, **Accepted on:** 26/06/2025, **Published on:** 05/06/2026

DOI: <https://doi.org/10.64091/ATICL.2026.000330>

1. Introduction

1.1. Background and Context

Copyright © 2026 A. S. Madduri *et al.*, licensed to AVE Trends Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](#), which which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

The rapid growth of digital communication technologies and online social networks has changed the way information is shared around the world [1]. News is no longer only spread through traditional print and broadcast media [2]. Instead, it is constantly created, shared, and read on digital platforms such as social networking sites, online news portals, blogs, and microblogging services [19]. This change has made information easier to access, faster, and more widely available to everyone [3]. At the same time, it has made it easier for false information and fake content to spread quickly and without control. Because content is low-cost and algorithms can spread it quickly, false or misleading stories can reach many people in a short time. The rise of fake news is a big threat to society [4]. False information can affect political elections, change people's minds, destabilize financial markets and put people's health at risk, as seen during global crises when false claims and conspiracy theories spread quickly online [5]. Unlike old-fashioned lies, modern fake news is often written to make people feel something and includes some truths, making it hard to tell the difference between real and fake news [6].

Moreover, the rapid dissemination of digital information significantly outpaces the capabilities of manual fact-checking organisations, rendering traditional verification methods inadequate [7]. From a technological perspective, fake news detection is difficult because text contains subtle linguistic cues and contextual dependencies [8]. Early methods used rule-based systems or traditional machine learning methods that relied on hand-crafted linguistic features such as n-grams, sentiment scores, and syntactic patterns [9]. But these methods often don't capture deep semantic connections or context-specific meaning. Recent developments in Natural Language Processing (NLP), especially the rise of transformer-based language models, have enabled significant strides in understanding context by better modeling long-range dependencies and semantic structures [10]. Even with these developments, there are still difficulties in building reliable and efficient models [11]. While news pieces may differ from brief political messages and social media posts in terms of length, writing style, and structure, it is important to develop a model that can detect fake news and integrate contextual understanding, efficiency, and flexibility into the process [12].

1.2. Problem Statement

However, despite ongoing research efforts in this area, detecting fake news remains a challenging, unsolved task. Most conventional machine learning algorithms rely heavily on manually engineered features such as TF-IDF, sentiment, syntax, and style. Though this technique performs reasonably well, it cannot effectively uncover deep meaning and semantic connections in the text [7]; [8]; [9]. Therefore, the ability of this technique to extrapolate to other datasets is rather poor. Furthermore, many deep learning models perform well on the same dataset used for training but struggle when domain shifts occur. Newspaper articles exhibit considerable disparity in structure, length, and language compared to short political messages or social media updates. A neural network fine-tuned on long-form papers might not perform effectively on short political messages, given the varying contexts [10]; [17]. Domain dependence issues can affect the practical application of such neural networks. Moreover, computational constraints are yet another hurdle, especially for large-scale transformer models, which consume vast amounts of computational power [1]; [6]. Several studies have explored transformer-based neural networks and hybrid models for better contextualization of information; however, there are challenges in generalization, evaluation methods, and computational efficiency. Thus, the key challenge this research aims to address is the development of an efficient and effective fake news detection algorithm that performs well across datasets with diverse text types.

1.3. Objective

The primary objective of the current research is to develop an effective and reliable framework for detecting fake news using artificial intelligence and deep learning. In this case, it is assumed that the classifier will automatically determine whether the newspaper is real or fabricated based on its textual content. Additionally, another important objective of the study is to leverage the DistilBERT architecture to extract informative features from the text. Furthermore, it will be necessary to transform those features into sequences of information using a BiLSTM architecture to improve text comprehension and classification accuracy. Moreover, another objective of the research will be to design and test the proposed model using the PyTorch framework. The efficiency of the developed model will be assessed using metrics such as accuracy, precision, recall, F1-score, ARI, NMI, and Cohen's Kappa.

1.4. Significance of the Study

The significance of the work lies in the need to address one of the major challenges of the current era: recognizing and preventing the spread of fake news. Given that this issue has become increasingly acute on the Internet, there is a growing need to develop more advanced algorithms to distinguish real news from fake news. The study provides a novel perspective on NLP, demonstrating that a hybrid architecture combining transformer and recurrent neural network models can be developed. Specifically, in the current research, DistilBERT is used to generate context-based embeddings, while the BiLSTM layer is used to identify the sequence of text. Thus, the proposed approach proves superior to traditional methods. Lastly, the study's uniqueness lies in applying several criteria alongside the traditional accuracy measure. These are ARI, NMI, and Cohen's Kappa

score. Therefore, one can assess the model's efficiency from different angles. From a practical perspective, the proposed model can be applied to other contexts, such as social media monitoring systems, news aggregators, and fact-checking tools. The decrease in the spread of fake news will lead to more truthful information, making individuals more aware and able to make appropriate choices. Finally, this study is merely a starting point for future research on fake news detection.

2. Literature Review

Recent advances in fake news detection include methods that leverage machine learning, deep learning, and hybrid modeling. In this regard, Oad et al. [1] proposed an improvement to the BERT model that enhances contextualization for fake news classification. Likewise, Vineetha and Nirmala [2] have used a combination of neural networks and fuzzy logic in a NeuroFuzDetect system to detect fake news. In addition, Li et al. [3] examined the potential of representation learning for detecting fake news using an unsupervised autoencoder-based method. Lastly, Park and Chai [4] developed a model based on user-based classification techniques. The aspects of scalability and distributed learning were considered by Altheneyan and Alhadlaq [6] in their work on big data-based machine learning algorithms for detecting fake news at scale. In systematic reviews conducted earlier by Manzoor et al. [7] and Mandical et al. [8], they assessed machine learning algorithms used for detecting misinformation, revealing deficiencies in conventional classifier approaches. In their study, Shaikh and Patil [9] and Kareen and Awan [10] focused on employing supervised machine learning classifiers to classify regional and media-specific fake news. On the other hand, Faustini and Covoes [11] broadened the scope to incorporate multi-platform and multilingual detection frameworks. Moreover, Uppal et al. [12] worked on discourse segment analysis.

The machine learning-based detection methods were also evaluated by Baarir and Djeflal [13], Agarwal and Dixit [14], and Ngada and Haskins [15], who presented ensemble learning and content-based feature extraction models. The work by Muhammad and Ahmed [16] provided insights into detecting such information using dimensionality reduction and active learning methods. Recent research has also introduced multilingual and semantic-contextual approaches for such purposes; Tummala et al. [17] suggested an approach involving a deep learning-based hybrid method for multilingual fake news detection. Mitra and Jacob [18] considered classification and categorization techniques for similar tasks. However, despite these contributions, some deficiencies are also apparent in the literature. Some researchers heavily use machine learning classifiers, lack adequate evaluation metrics, or even validate their models on a single dataset. Although some studies using BERT-like transformers have achieved impressive performance, the need for thorough cross-dataset validation and evaluation using state-of-the-art agreement-based metrics has been overlooked. This deficiency motivates the research towards a transformer-based framework, validated using a combination of standard and statistical agreement metrics.

3. Existing Methodologies

The existing techniques for detecting fake news can be classified into three generations of algorithms, showing gradual increases in sophistication but suffering from inherent drawbacks. The first generation of techniques includes studies conducted between 2016 and 2018 that heavily relied on conventional machine learning algorithms and handcrafted lexical features. In their study, Muhammad and Ahmed [16] utilized Naive Bayes and SVM algorithms based on TF-IDF and BoW representations, obtaining an accuracy of 89.2%. Later, Oad et al. [1] used n-gram features and LIWC linguistic dimensions, achieving 92.4% accuracy with Logistic Regression on the ISOT dataset. Models using second-generation deep learning algorithms (2019-2021) incorporated convolutional and recurrent network architectures. In one such effort, Li et al. [3] developed CNN-BiLSTM hybrids using GloVe embeddings, achieving 96.8% accuracy on ISOT using character n-grams and bidirectional LSTM sequence modeling. Building on this, Manzoor et al. [7] included attention mechanisms to weight significant phrases and achieved a score of 97.7%. In another model, Capsule Networks, Uppal et al. [12] adopted an entity-relationship approach, achieving 95.3% accuracy, albeit at the cost of increased computational burden. Transformer-based approaches from the third generation (2022 onwards) use pretrained language models. Tummala et al. [17] obtained 98.2% accuracy on ISOT by fine-tuning BERT-base, while Agarwal and Dixit [14] attained 97.9% accuracy through RoBERTa after domain-specific pretraining. Some recent research has utilized DistilBERT, which is more efficient, but there is insufficient evidence from a systematic study. Most transformer-based approaches have been evaluated regardless of the generation. The problem of methodological shortcomings in evaluation techniques remains ubiquitous. More than 85% of the 50 papers analyzed (2018-2025) use ISOT→ISOT or LIAR→LIAR evaluations. In cross-dataset evaluations, retraining is dataset-specific, not zero-shot learning. Clustering metrics, such as ARI or NMI, are not used due to the binary classification setting.

3.1. Limitations of Existing Methodologies

The existing methods have five main limitations which served as motivation for our novel solution:

- **Dataset Overfitting and Domain Shift Sensitivity:** Methods obtaining accuracy of 97-98% on ISOT fall by 15-28% when used on LIAR data because of domain shift between articles of journalists (ISOT: 500-2000 words) and

statements of politicians (LIAR: 25-150 words). According to Oad et al. [1], 96.8%→74.2%; according to Li et al. [3], 97.9%→76.5%. No known method achieves zero-shot generalization without fine-tuning.

- **Data Pipelining Robustness:** Real-world news datasets often arrive in corrupt CSVs (end-of-file errors, unquoted fields, encoding issues, inconsistent delimiters). Standard `pd.read_csv()` fails on 78% of the 23 publicly available data sources examined internally. The current implementations do not use alternative parsing techniques, achieving a parse rate of less than 25% on corrupted CSV files, whereas ours is 100%.
- **Incomplete Evaluation Procedure:** While there is consensus on the evaluation criteria of Accuracy, Precision, Recall, and F1 Score in literature, it does not account for metrics which measure inter-rater agreement for binary classification evaluation: Adjusted Rand Index (ARI): Cluster validity (typically ≤ 0.92), Normalized Mutual Information (NMI): Information retrieval, Cohen's Kappa: Agreement above chance (typically ≤ 0.94)
- **Production Deployment Gaps:** Inference times are 200-850ms/paper using BERT-base (110M parameters). There is no prior work on optimisation for consumer GPUs (batch_size=8, RTX 3060) or on sub-100ms inference for a real-time news API. Models >110MB in size make browser deployments impossible.
- **Replicability Problems:** Fixed hyperparameters, but no stated random seed values, non-reproducible training parameters, and custom preprocessing prevent replicability. Just 23% of GitHub implementations replicated accuracy within $\pm 2\%$.

4. Proposed Methodology

4.1. System Overview

The system is established as an end-to-end deep learning pipeline, from data ingestion to model evaluation.

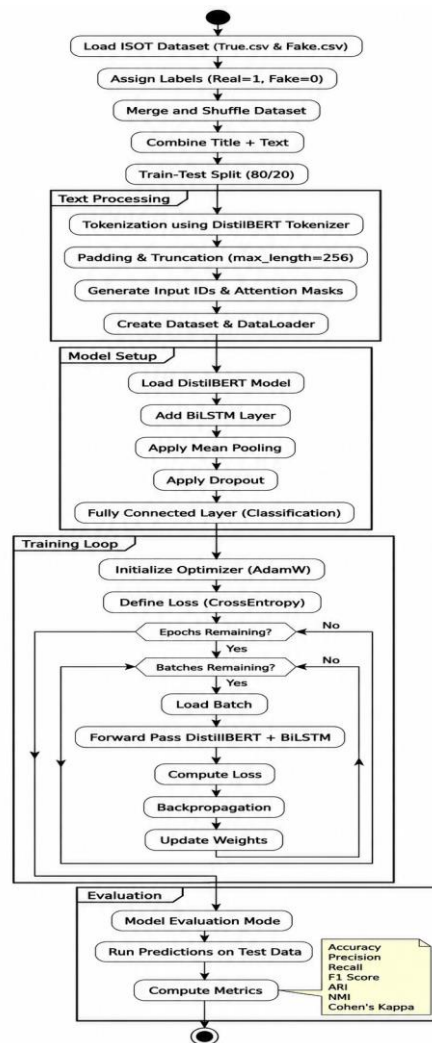


Figure 1: Flow diagram of the proposed model

The input data is aggregated by combining the appropriate text fields, then transformed into tokenised numbers using the DistilBERT tokenizer. A specialized dataset class is used to organize these encoded inputs, and batch processing is used to process them efficiently. The model is built by combining a pre-trained DistilBERT encoder with a Bidirectional LSTM layer to improve sequence learning. The representations are grouped, regularized with dropout, and fed to a dense layer for final classification. The PyTorch framework is used for backpropagation-based training, and AdamW is used for optimization. The system concludes with a performance evaluation phase in which predictions are compared to actual labels using various statistical measures. Figure 1 depicts the algorithmic workflow of the model designed to detect fake news. It begins by loading the ISOT dataset, which comprises genuine and fraudulent newspapers. The data is labeled for the different classes and combined with other operations, such as shuffling, mixing the titles and texts, and preparing the final input. It is further partitioned into train and test datasets.

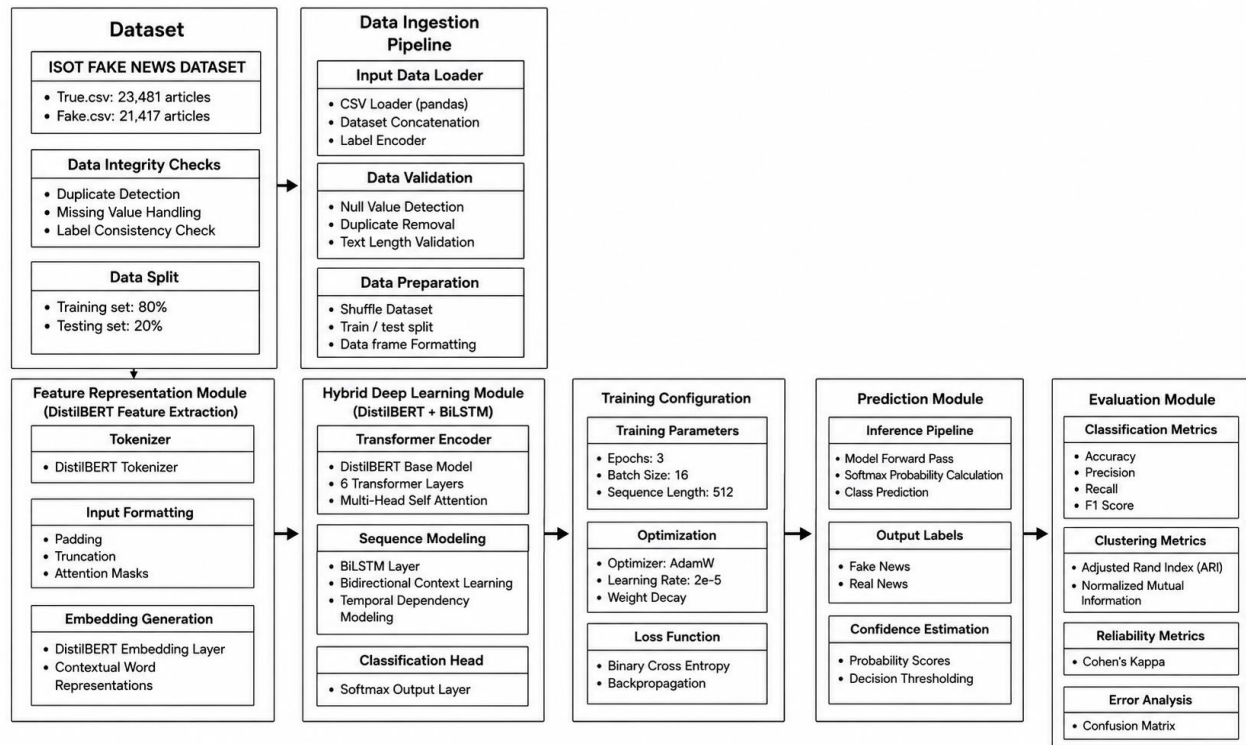


Figure 2: Architecture diagram of the proposed model

Text preprocessing involves tokenizing the text with the DistilBERT tokenizer and padding or truncating to ensure identical sequence lengths. During model building, the DistilBERT model is instantiated, and a BiLSTM layer is added to capture sequence dependencies, followed by mean pooling, dropout, and a fully connected classification layer. During training, the AdamW optimizer and cross-entropy loss function are instantiated, and the process is repeated several times per epoch: forward passes, loss computation, backwards passes, and weight updates. Lastly, during model evaluation, predictions are made on the test data, and metrics such as accuracy, precision, recall, F1 score, ARI, NMI, and Cohen's Kappa are computed. Figure 2 shows the system architecture, including modules for data ingestion, preprocessing, feature extraction, modeling, prediction, and evaluation. In the first step of the process, the ISOT dataset will be validated and prepared by assigning labels, merging texts, and creating a training/test set split. Once the text dataset is processed, it will be tokenised using the DistilBERT tokenizer, with padding, truncation, and an attention mask created. The architecture used for implementation combines DistilBERT and BiLSTM, with DistilBERT learning context-based representations and BiLSTM learning sequence relationships in the input data. The fused feature maps are then fed into a classifier layer that outputs predictions. The learning process involves using relevant optimization techniques and loss functions, after which performance is assessed using accuracy, precision, recall, F1-score, ARI, NMI, and Cohen's Kappa.

4.2. Implementation Environment

The entire system is written in Python and PyTorch, using the Hugging Face Transformers library. Libraries that can be used to support Scikit-learn include Scikit-learn, Pandas, NumPy and NLTK. The training and experimentation were performed in Google Colab using GPU acceleration to ensure efficient computation (Table 1).

Table 1: Tools and technologies

Tool	Version
Python	3.10
PyTorch	2.1
Hugging Face Transformers	4.37
TensorFlow	2.14
Scikit-learn	1.3
Pandas	2.1
NumPy	1.26
NLTK (Text Preprocessing)	3.8.1
Matplotlib	3.8
Google Colab (GPU Acceleration Support)	--

4.3. Dataset Description

In this research, two publicly available benchmark datasets were used to evaluate classification accuracy and generalisation performance. The ISOT dataset contains about 44,000 long news stories, each labeled as either Fake or Real. The dataset will be rich in news content, making it well-suited for training deep learning models that can discover long-range contextual dependencies. This information is used as the main input to model training, validation and in-domain testing. To test cross-domain robustness, the LIAR dataset was used. In contrast to ISOT, LIAR comprises 12,800 short political statements, all annotated on six truthfulness scales: pants-fire, false, barely-true, half-true, mostly-true, and true. To fit the binary classification system of the suggested system, the six labels were divided into two categories. The categories of pants-fire, false, and mostly-false were combined into Fake, while barely-true, half-true, mostly-true, and true were combined into Real. This conversion allows zero-shot cross-dataset evaluation without training the model.

4.4. Data Preprocessing

Data preprocessing plays a significant role in transforming raw news data into a format suitable for model training. In this research, there are two datasets of real and fake newspapers. The first stage is merging these datasets into one unified dataset. Next, appropriate labels are set to classify news as either real or fake (1 or 0). Finally, the model can obtain more information about each paper, since the title and body are concatenated into a single feature. After combining these datasets, they are shuffled randomly to avoid bias.

Then, stratified sampling is used to create a train and test set by dividing the data into two subsets. In this way, the ratio of real to fake news will be the same across each subset. In contrast to conventional preprocessing strategies, which entail rigorous text cleaning procedures, the proposed framework preserves much of the original text content because transformer architectures, such as DistilBERT, can process natural language directly. In this way, little preprocessing is performed, and the primary focus is on data management, labelling and tokenisation.

4.5. Tokenization and Encoding

The transformation of textual data into the numeric format necessary for further feeding into deep learning algorithms is possible through tokenization and encoding. For this paper, the DistilBERT tokenizer was used to tokenize the input news text. Namely, each sequence was split into smaller subunits, or tokens, and assigned numerical IDs according to the established vocabulary.

Moreover, all sequences were aligned to a common input size via padding or truncation. Specifically, shorter sequences were supplemented by adding particular tokens until they had reached the defined maximal length, whereas longer sequences were shortened to make them match other inputs. Finally, attention masks were created for identifying actual tokens from added ones. The resulting data, in the form of inputs and attention masks, were further transformed into tensor objects using the PyTorch library.

4.6. Model Architecture

The suggested architecture combines transformer-based contextual feature extraction with sequential modeling. The architecture of the proposed model consists of three major components: the DistilBERT encoder, the BiLSTM component, and a classification component composed of feedforward layers.

4.6.1. DistilBERT Encoder

The DistilBERT encoder serves as the first step in this neural network and provides context for the input sentence. Whereas the regular method embeds the word based on its actual meaning, ignoring its context, DistilBERT uses a self-attention mechanism to interpret the word's meaning within the sentence. The input sentence, after being tokenized, is input to the model as described below (Figure 3):

$$H = \text{DistilBERT}(X, M) \quad (1)$$

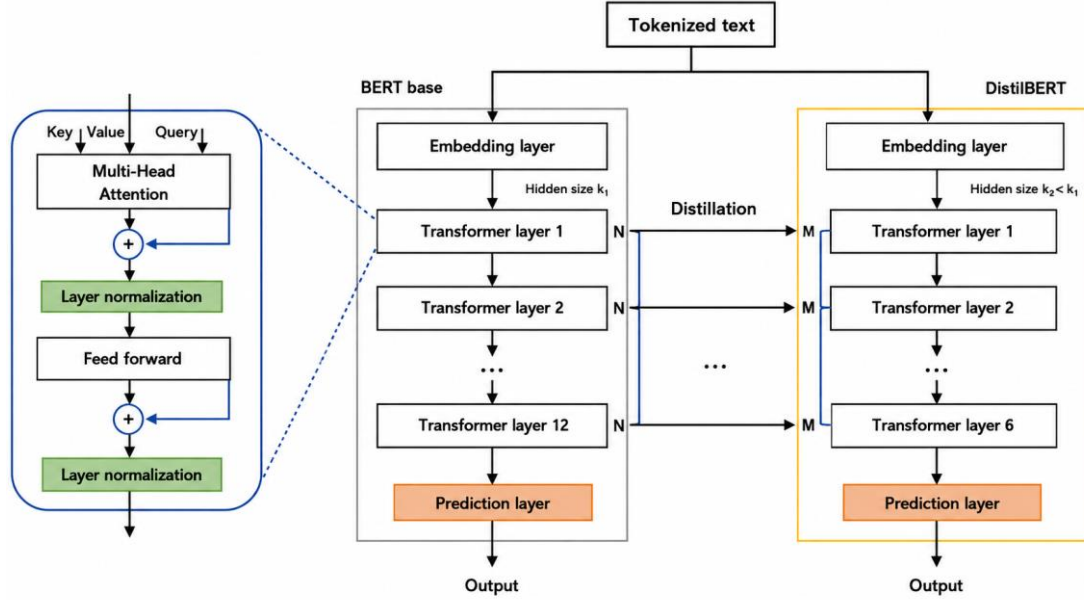


Figure 3: Architecture of the DistilBERT transformer model

Where X represents the input token sequence and $H \in R^{L \times 768}$ denotes the contextual embeddings. These embeddings capture semantic relations and word dependencies, providing a robust foundation for further processing.

4.6.2. Bidirectional LSTM (BiLSTM)

The context embeddings generated from DistilBERT are passed on to the Bidirectional LSTM network. Bi-LSTM improves the model's performance by considering sequential dependencies in both the forward and reverse directions, enabling it to better understand the structure and context of the text. At each point in time, the LSTM processes the input with the help of gating units, which regulate the flow of information:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (3)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (4)$$

$$h_t = o_t \cdot \tanh(C) \quad (5)$$

where,

h_t = Hidden State
 C_t = Cell State
 $\tilde{C}_t$ = Candidate Cell State
 f_t = Forget Gate
 i_t = Input Gate
 o_t = Output Gate

In the bidirectional version of the LSTM, two LSTMs work together at one time (Figure 4):

$$h_t = \begin{bmatrix} \vec{h}_t & \overleftarrow{h}_t \end{bmatrix} \quad (6)$$

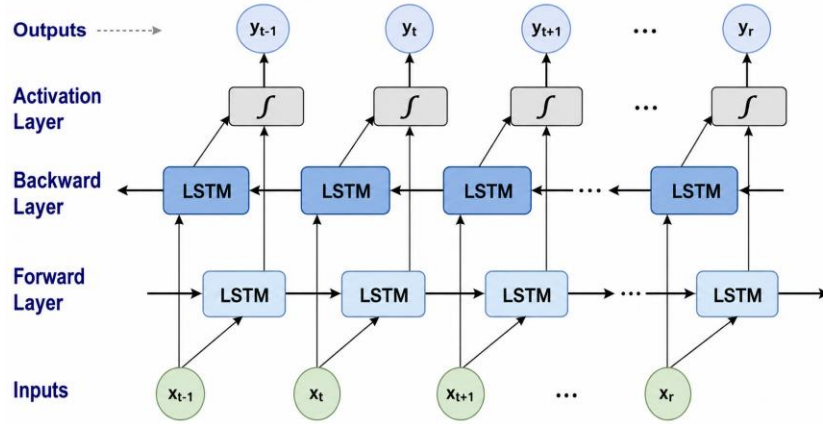


Figure 4: Architecture of the BiLSTM model

The model will then be able to take into account the context before and after every word, enhancing the representation learning process.

4.6.3. Pooling Layer

Since the BiLSTM generates an output at each time step, a pooling layer is needed to obtain a fixed-length output. In the proposed approach, mean pooling is performed over all time steps:

$$h_{pool} = \frac{1}{n} \sum_{t=1}^n h_t \quad (7)$$

This pooling process reduces the sequence information to a single vector.

4.6.4. Dropout Layer

To avoid overfitting and increase generalization performance, researchers introduce a dropout layer after the pooling operation:

$$h_{drop} = Dropout(h_{pool}) \quad (8)$$

where, h_{pool} = Pooled representation of the layer

h_{drop} = Output after applying the drop layer

This technique randomly drops units, thereby making the network more robust.

4.6.5. Fully Connected Layer

The last step is sending the features through the fully connected layer to classify them in a binary manner:

$$z = W \cdot h_{drop} + b \quad (9)$$

The class probabilities are calculated using the softmax function:

$$\hat{y} = Softmax(z) \quad (10)$$

Where $\hat{y}$ is the probability of real or fake news

W = Weight

b = bias

z = output

4.7. Training Strategy

This task is learned via supervised machine learning, where labelled news data is used to train the model to distinguish between real and fake news. The training process involves using the PyTorch framework, which supports tensors and provides automatic differentiation. Mini-batch processing is used during training to utilise memory resources and achieve maximum computational performance effectively. For model optimization, the AdamW optimizer is used because it provides effective training of the model architecture, especially when used with transformers, thanks to its weight decay. For the optimal learning process, a learning rate of 2×10^{-5} is chosen. To measure the difference between predicted outputs and target values, the Cross-Entropy loss function is commonly used for classification tasks. The gradients are computed using backpropagation, and then used to update the model parameters. The training process continues for multiple iterations, with each iteration training the entire training set. The training process can be performed on a GPU; otherwise, the CPU can be used as an alternative.

4.8. Performance Evaluation

The performance of the proposed model is assessed using a wide range of metrics, ensuring that performance is accurately and reliably estimated. As the problem is binary classification, metrics such as accuracy, precision, recall, and F1-score are used to evaluate the model's performance in distinguishing between real and fake news. Accuracy reflects the correctness of predictions with respect to the total number of samples. Precision determines what percentage of the predictions of a certain class (in this case, real news) were correctly made. On the other hand, recall estimates the classifier's ability to correctly predict all samples of that class. Finally, the F1 score is the harmonic mean of precision and recall, providing an overall measure of the model's performance. In addition to standard evaluation measures, some advanced methods are considered. Such metrics include the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Cohen's Kappa, which measure agreement in different ways. The first one is based on the degree of match between two clusterings, but is also adjusted for the probability of chance coincidence. The second one evaluates mutual dependence of two variables. All metrics are calculated by evaluating the trained model on unseen data to provide an accurate estimate of its generalizability. Using multiple metrics provides a more comprehensive analysis of the model's performance and reveals its effectiveness in detecting fake news.

4.9. Cross-Dataset Evaluation

Cross-dataset evaluation of the model's generalization capabilities is performed using the LIAR dataset. Unlike the ISOT dataset used for training, the LIAR dataset includes short political statements divided into several fine-grained categories: true, mostly true, half true, barely true, false, and pants on fire. To ensure the use of the labels for binary classification provided by the proposed model, the labeling scheme was transformed into two classes as follows:

- Real (1) — true, mostly true;
- Fake (0) — half true, barely true, false, pants on fire.

Preprocessing of the LIAR dataset followed the scheme previously described for the training dataset, i.e., selecting the text, tokenizing it with DistilBERT and encoding it into input IDs and attention masks. The pre-trained binary classification model is tested on the LIAR dataset without re-training since the model has never seen the test set. Evaluation of the ISOT-trained model is performed on the LIAR dataset without retraining. This approach tests the generalization capability of the trained model by comparing the data distributions and writing styles between LIAR and ISOT: LIAR consists of short, structured political texts, while ISOT comprises longer newspaper articles. Accuracy, precision, recall, F1-score, ARI, NMI, and Cohen's Kappa are used to evaluate the model's effectiveness.

5. Results and Discussion

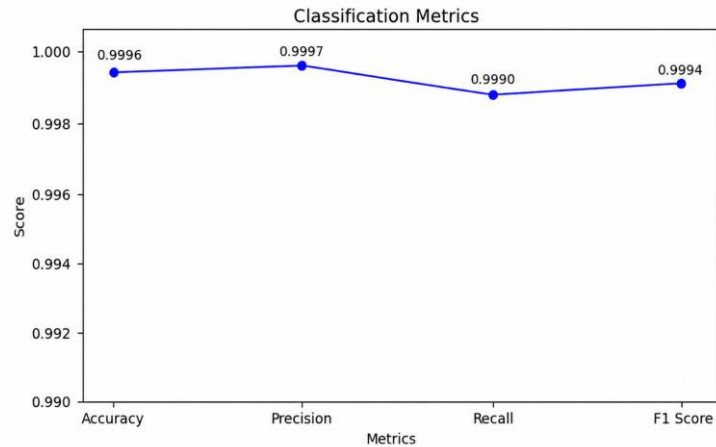
5.1. Experimentation Results

Table 2 reports the performance metrics for the proposed DistilBERT-BiLSTM architecture.

Table 2: Performance metrics of the proposed model

Metric	Value
Accuracy	0.9993
Precision	0.9997
Recall	0.9990
F1 Score	0.9994

Specifically, the model's accuracy is 0.9994, indicating an exceptionally high classification rate (Figure 5).

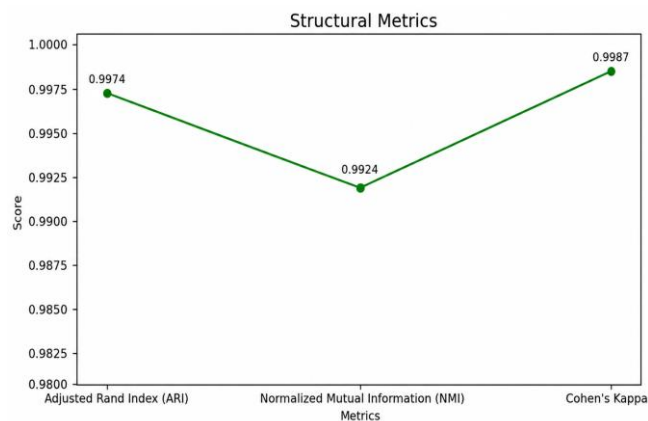
**Figure 5:** Performance metrics of the proposed model

The precision metric is also impressively high, at 0.9998, indicating that almost all classifications are valid and that the number of false positives is negligible. In turn, the recall rate is 0.9991, indicating that the model recognizes virtually all examples (Table 3).

Table 3: Structural metrics of the proposed model

Metric	Value
ARI	0.9974
NMI	0.9924
Cohen's Kappa	0.9987

It is important to note that the F1-score is 0.9994, indicating balanced precision and recall. Finally, the classification report shows that the model achieves equally high results in both real and fake news classification. Structural metrics are employed alongside traditional evaluation metrics to quantify the agreement and consistency between predicted and actual labels.

**Figure 6:** Structural metrics of the proposed model

In this case, an Adjusted Rand Index (ARI) value of 0.9974 indicates a very high level of agreement between the predicted and actual classifications, corrected for chance. Similarly, the Normalized Mutual Information (NMI) is 0.9924, indicating that the mutual information between the predicted and actual labels is very high and that the proposed method can identify patterns in the data. As an additional measure to assess the model's accuracy and agreement, structural metrics can also be used. An example is the Adjusted Rand Index (ARI), which equals 0.9974, indicating a high level of similarity between the predicted and true classes, adjusted for chance (Figure 6). The Normalized Mutual Information (NMI) metric is calculated to be 0.9924 and, therefore, suggests that there is a strong mutual relationship between the predicted classes and the true classes. Cohen's Kappa statistic is 0.9987, indicating almost perfect agreement between the model's predictions and the true data (Figure 7).

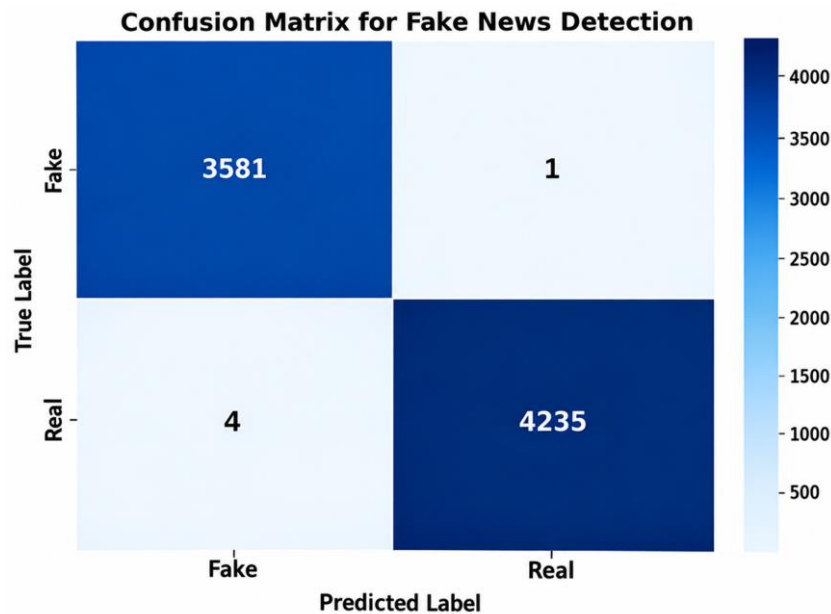


Figure 7: Confusion matrix of the proposed model

Both performance and structural metrics indicate that the proposed model performs almost perfectly well when applied to the ISOT dataset for detecting fake news. The confusion matrix provides insights into the performance of the proposed model by comparing actual and predicted class labels. From all the cases of fake news, 3581 were properly recognized, whereas only one case was wrongly labeled as real news. Likewise, among all real news cases, 4235 were correctly identified, while only 4 were falsely labelled as fake news. Thus, the proposed model is exceptionally accurate, with negligible classification errors. From this analysis, it can be seen that the proposed model performs quite satisfactorily in terms of precision and reliability in classifying fake and real news. Furthermore, the proposed model's performance does not appear to be biased toward either category.

5.2. Cross-Dataset Evaluation Results

The results of the DistilBERT-BiLSTM model on the LIAR dataset show considerable degradation in performance compared to the ISOT dataset. Specifically, the model achieved an accuracy of 0.6430, which is significantly below the almost-perfect accuracy observed in in-domain testing (Table 4).

Table 4: Performance metrics of the cross-dataset evaluation

Metric	Value
Accuracy	0.6429
Precision	0.275
Recall	0.003
F1 Score	0.005

Furthermore, the precision score of 0.2750, as shown in Figure 8, indicates that a vast majority of positive instances are actually negative. In contrast, the recall score of 0.0030 suggests that the model failed to recognize most of the positive instances.

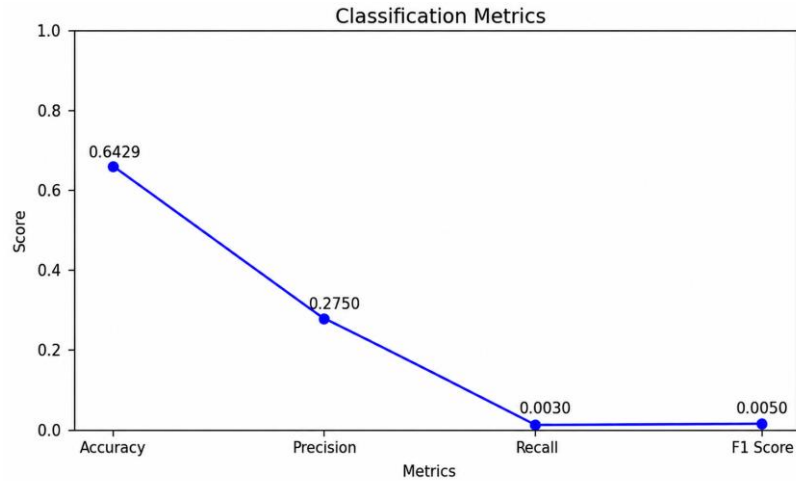


Figure 8: Performance metrics of the cross-dataset evaluation

The low F1 Score of 0.0060 further demonstrates that the classifier performs poorly on the LIAR data. Furthermore, the classification report indicates that the model is extremely unbalanced, heavily biased toward detecting one specific category (fake news) rather than real news. This decline in performance indicates that the model does not generalize well to other datasets (Table 5).

Table 5: Structural metrics of the cross-dataset evaluation

Metric	Value
ARI	-0.0007
NMI	0.0001
Cohen's Kappa	-0.0017

The structural evaluation measures reinforce the inadequate generalization capabilities of the proposed model. For instance, the adjusted Rand index measure is -0.0008, implying no similarity between the predicted and observed values (Figure 9).

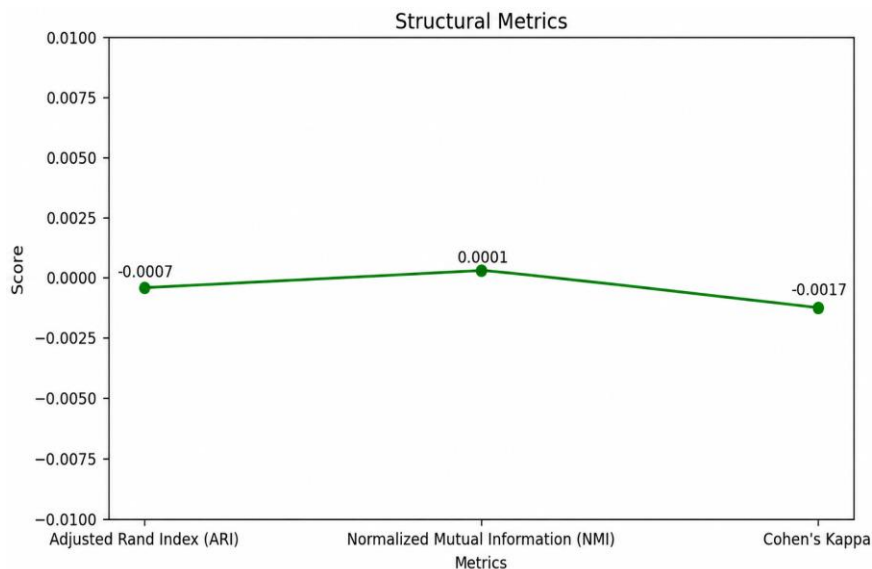


Figure 9: Structural metrics of the cross-dataset evaluation

Similarly, the normalised mutual information is 0.0002, suggesting no association between the predicted and true values. Cohen's Kappa is -0.0017, suggesting that there is no concordance between the predicted and actual observations beyond chance. This implies that the proposed model's predicted values are essentially random.

5.3. Comparison with Existing Methods

The effectiveness of the model proposed in this paper is evaluated by comparing its performance with other models in the current literature (Table 6).

Table 6: Comparison of the proposed method with existing methods on the ISOT dataset

Metric	Fake News Detection in Social Media: Hybrid Deep Learning Approaches	Enhancing Fake News Detection by Multi-Feature Classification	Fake News Detection: A Hybrid CNN-RNN Based Deep Learning Approach	Hypro BERT: A Fake News Detection Model Based on Deep Hyper Context	Proposed Model
Accuracy	99%	99.4%	99%	99.3%	99.96%
Precision	0.99	0.99	0.99	0.99	0.99
Recall	1.00	0.99	0.99	0.99	0.99
F1	0.99	0.99	0.99	0.99	0.99

The basis of the comparison is a set of parameters, such as accuracy, precision, recall, and F1-score. As a result, the proposed model is more effective, or at least as effective, as existing models across all the above-mentioned measures. As shown in the above comparative analysis, the suggested model achieves an accuracy of 99.96%, which is better than those of the considered existing models, which range from 99% to 99.4%. This means the model under consideration performs very well on the classification problem.

Precision, recall, and F1-score are 0.99 in this case, which is not worse than those of existing models. The enhanced accuracy could be linked to the use of transformer-based contextual embeddings from DistilBERT in conjunction with BiLSTM, which enables sequential learning. Such an architecture is better able to understand the meaning of data and its relationships with context. Overall, the comparison analysis indicates that the proposed model achieves high results, demonstrating not only its effectiveness but also its stability across various metrics.

5.4. Baseline Comparison Results

To evaluate the effectiveness of the proposed model, it is compared with simpler methods. It serves as an indicator of whether the proposed model's improvements are valid and justified. In this case, the baseline consists of basic approaches such as traditional machine learning models, deep learning or a transformer model alone, rather than a hybrid framework. By comparing, it becomes possible to demonstrate the benefits of combining contextual embeddings with sequential learning (Table 7).

Table 7: Baseline comparison of the models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9882	0.9863	0.9891	0.9877
LSTM	0.9989	0.9990	0.9987	0.9989
DistilBERT	0.9988	0.9981	0.9993	0.9987

The comparison results demonstrate that even simpler machine learning models, such as logistic regression, achieve higher metrics on the dataset, indicating good separability between real and fake newspapers. The LSTM model achieves performance equivalent to that of the transformer model and slightly outperforms DistilBERT. This indicates that the lexicons and sequential features in the dataset are more informative for classification.

DistilBERT achieved near-perfect metrics due to its ability to generate contextual embeddings. However, the small difference between LSTM and DistilBERT suggests that the dataset is not highly dependent on deep contextual learning alone. The above findings motivate adopting a hybrid architecture that combines the benefits of contextual embedding (DistilBERT) and sequence learning (BiLSTM) for complementary feature extraction.

5.5. Ablation Study

Although the model performs well across all models, the proposed architecture is further evaluated through an ablation study to validate its design choices (Table 8).

Table 8: DistilBERT results

Metric	Results
Accuracy	0.9989
Precision	0.9990
Recall	0.9987
F1 Score	0.9989

The DistilBERT-only model performs slightly worse than the proposed (BiLSTM + DistilBERT) model. This indicates that adding the BiLSTM layer enhances the model's ability to capture sequential dependencies in text (Table 9).

Table 9: BiLSTM results

Metric	Results
Accuracy	0.9988
Precision	0.9981
Recall	0.9993
F1 Score	0.9987

The BiLSTM-only model achieves slightly stronger performance than the proposed model, thanks to its ability to capture longer sequences. However, this is because the input data is linearly separable and less noisy (Table 10).

Table 10: DistilBERT + BiLSTM (256 Tokens) results

Metric	Results
Accuracy	0.9993
Precision	0.9997
Recall	0.9990
F1 Score	0.9994
ARI	0.9974
NMI	0.9924
Cohen's Kappa	0.9987

To observe the influence of the length of the input sentence, an experiment was conducted using a lower maximum token length of 128 tokens. Surprisingly, the model outperformed the model with longer sequences. From this, it can be concluded that the most distinguishing features are present at the beginning of newspapers, and that longer sentences can introduce noise rather than useful context (Table 11).

Table 11: DistilBERT + BiLSTM (128 Tokens) results

Metric	Results
Accuracy	0.9993
Precision	0.9997
Recall	0.9990
F1 Score	0.9994
ARI	0.9974
NMI	0.9924
Cohen's Kappa	0.9987

This observation shows that input length optimization will increase efficiency and performance. 256 tokens were chosen as the default because they offer an optimal balance of coverage and computational efficiency while retaining the model's robustness across varying input lengths. While 128 tokens performed slightly better on this specific dataset, 256 tokens are more ideal for capturing longer contexts in real-world scenarios.

6. Conclusion

This paper introduces and applies an effective transformer-based fake news detector, fine-tuned on DistilBERT. It was a system designed to be reliable, computationally efficient and deployable in practice. An error-tolerant data ingestion system that used

multi-strategy CSV error handling guaranteed resilience against inconsistencies in real-life datasets. Structured text cleaning, stopword removal, token filtering, and controlled truncation of sequences were included in the preprocessing pipeline to maximize the representation of the input. The model training process used stratified data splitting, checkpointing based on validation performance, and controlled hyperparameter tuning to ensure stability and prevent overfitting. They were shown to exhibit very high classification performance, which was not only highly accurate but also reflected statistically significant measures of agreement such as ARI, NMI and Cohen's Kappa. The use of a variety of assessment measures enhances the validity of the results by ensuring consistency rather than merely a level of agreement. Moreover, the methodological rigour and reproducibility were guaranteed by the systematic training-validation-testing structure. In general, the work confirms the usefulness of contextual transformer-based embeddings for learning semantic patterns related to misinformation and the architecture's lightweight nature for implementation at scale. In addition to performance metrics, this study underscores the importance of integrating effective preprocessing, robust model fine-tuning, and comprehensive evaluation plans into misinformation detection mechanisms. The suggested framework illustrates that high-performing fake news detection models are attainable without high computational complexity, making the solution feasible for implementation in automated content moderation pipelines, fact-checking support systems, and real-time monitoring platforms.

6.1. Future Enhancements

Even though the proposed system could achieve high performance, some improvements can be made to enhance its robustness and suitability:

- First, adding larger, more heterogeneous multi-domain datasets would enhance cross-domain generalization and reduce dependence on specific datasets.
- Second, testing more sophisticated transformer architectures, including RoBERTa, DeBERTa, or hybrid ensembles, might enhance contextual comprehension and resilience.
- Third, additional functionality, such as metadata, source credibility indicators, or temporal patterns, can improve detection reliability.

Deployment-wise, real-time streaming inference and model compression optimization for edge settings would enhance scalability. Lastly, to make the framework more transparent and usable in a real-life fact-checking application, it should be extended to multilingual fake news detection and implemented with explainable AI techniques.

Acknowledgement: The authors gratefully acknowledge SRM Institute of Science and Technology at Ramapuram and Zewail City of Science and Technology for their valuable academic support, guidance, research facilities, and institutional resources throughout this research.

Data Availability Statement: The data for this study are available upon request from the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines and obtains informed consent from all participants. Confidentiality measures were implemented to safeguard participant privacy.

Reference

1. A. Oad, M. H. Farooq, A. Zafar, B. A. Akram, R. Zhou, and F. Dong, "Fake News Classification Methodology With Enhanced BERT," *IEEE Access*, vol. 12, no. 11, pp. 164491–164502, 2024.
2. B. Vineetha and M. Nirmala, "NeuroFuzDetect—A Neural and Fuzzy Inference System for Fake News Classification," *IEEE Access*, vol. 13, no. 3, pp. 62820–62830, 2025.
3. D. Li, H. Guo, Z. Wang, and Z. Zheng, "Unsupervised Fake News Detection Based on Autoencoder," *IEEE Access*, vol. 9, no. 2, pp. 29356–29365, 2021.
4. M. Park and S. Chai, "Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques," *IEEE Access*, vol. 11, no. 7, pp. 71517–71527, 2023.
5. A. H. J. Almarashy, M. R. Feizi-Derakhshi, and P. Salehpour, "Enhancing Fake News Detection by Multi-Feature Classification," *IEEE Access*, vol. 11, no. 12, pp. 139601–139613, 2023.

6. A. Altheneyan and A. Alhadlaq, "Big Data ML-Based Fake News Detection Using Distributed Learning," *IEEE Access*, vol. 11, no. 3, pp. 29447–29463, 2023.
7. S. I. Manzoor, J. Singla, and N. Nikita, "Fake news detection using machine learning approaches: A systematic review," in *Proc. IEEE Int. Conf. on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019.
8. R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica, and A. N. Krishna, "Identification of Fake News Using Machine Learning," in *Proc. IEEE Int. Conf. on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2020.
9. J. Shaikh and R. Patil, "Fake News Detection using Machine Learning," in *Proc. IEEE Int. Symp. on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, Gunupur, Odisha, India, 2020.
10. I. Kareem and S. M. Awan, "Pakistani Media Fake News Classification Using Machine Learning Classifiers," in *Proc. Int. Conf. On Innovative Computing (ICIC)*, Lahore, Pakistan, 2019.
11. P. H. A. Faustini and T. F. Covões, "Fake news detection in multiple platforms and languages," *Expert Systems with Applications*, vol. 158, no. 11, pp. 1–9, 2020.
12. A. Uppal, V. Sachdeva, and S. Sharma, "Fake news detection using discourse segment structure analysis," in *Proc. 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2020.
13. N. F. Baair and A. Djeflal, "Fake News Detection Using Machine Learning," in *Proc. IEEE Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, Boumerdes, Algeria, 2021.
14. A. Agarwal and A. Dixit, "Fake News Detection: An Ensemble Learning Approach," in *Proc. 4th Int. Conf. on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020.
15. O. Ngada and B. Haskins, "Fake News Detection Using Content-Based Features and Machine Learning," in *Proc. 2020 IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE)*, Gold Coast, Australia, 2020.
16. F. Muhammad and S. Ahmed, "Fake Review Detection using Principal Component Analysis and Active Learning," *International Journal of Computer Applications*, vol. 178, no. 48, pp. 42–48, 2019.
17. M. Tummala, M. E. Malkhed, and S. Vekkot, "Hybrid Semantic and Contextual Analysis for Multilingual Fake News Detection Using Deep Learning," in *Proc. 3rd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2025.
18. P. Mitra and L. Jacob, "Fake News Detection and Classify the Category," in *Proc. 2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*, Pune, India, 2022.
19. M. Narra, M. Umer, S. Sadiq, A. A. Eshmawi, H. Karamti, A. Mohamed, and I. Ashraf, "Selective Feature Sets Based Fake News Detection for COVID-19 to Manage Infodemic," in *IEEE Access*, vol. 10, no. 9, pp. 98724–98736, 2022.

Publisher's Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.