
Cross-Domain Fake News Classification via RoBERTa Fine-Tuning on Hybrid LIAR-FakeNewsNet Corpus with Deployable Web Inference

J. Balaraju* and Karthik Jilla

*Department of Computer Science and Engineering, School of Engineering, Anurag
University, Hyderabad, Telangana, India*

E-mail: jb7443@gmail.com; karthikanurag2000@gmail.com

**Corresponding Author*

Received 04 May 2026; Accepted 09 June 2026

Abstract

The rapid spread of misinformation on social media platforms escalates a serious threat to world socio-political stability, health and financial security. While current automated fake news detection models boast impressive accuracy scores, these are often due to memorisation of the training domain rather than genuine generalisation to unseen examples. This study tackles the pressing need for domain-robust misinformation and disinformation detection through the creation of a highly adaptive and deployable classification system. The challenge stems from the severe drop in performance when exposed to the domain shift, especially with the emergence of generative language models that can perform style-based attacks. The proposed solution employs a RoBERTa-base transformer model, fine-tuned on a large, multi-domain corpus of 37,143 instances. This dataset mixes the short, politically focused statements of the LIAR dataset with the long, multi-topic articles from the FakeNewsNet dataset and ensures cross-domain generalisation. Using full-parameter fine-tuning and evaluation across different domains, the

Journal of Web Engineering, Vol. 25_7, 1295–1318.

doi: 10.13052/jwe1540-9589.2572

© 2026 River Publishers

model attained 86.7% accuracy, 87.3% macro precision and 86.4% macro F1-score on a diverse test corpus, surpassing traditional baselines. Importantly, the assessment unveiled a conservative labelling tendency with a 90.4% precision in detecting fake news and a 93.1% recall in detecting genuine news which prevents the severe consequences of mistakenly accusing legitimate news. Moreover, the design breaks out of research-only models with its integration into a real-time web service (based on Flask), providing sub-second inferences with confidence scores. This approach simultaneously addresses the problem of cross-domain vulnerability of state-of-the-art natural language processing models and demands deployable interpretability, setting a new benchmark for the implementation of natural language processing models in the fight against advanced, online misinformation.

Keywords: Fake news detection, transformer models, RoBERTa, deep learning, cross-domain classification, misinformation, natural language processing.

1 Contribution/Originality

This work addresses a crucial aspect in research on fake news detection: the evaluation in real cross-domain settings. Previous studies claim to have high accuracy (e.g. above 95%). However, they are based on single-domain studies. This allows the models to exploit specific stylistic features rather than learn the truth about the world.

This study is significant in that it created and tested a cross-domain fake news detector using the mix of the LIAR political fact-check data [1] and the multi-topic FakeNewsNet data [2]. This gives a high variety in length, style and topical distribution which makes the classification task more difficult than the classic single domain test sets.

Unlike many previous studies which are only focused on improving the model architecture, this research shows that careful cross-domain training and testing can lead to a model which can successfully classify different types of content. A mixed-domain test set accuracy of 86.7% was achieved. This is more representative of real-world performance than the higher accuracy in one domain in previous research.

This study also provides the error analysis of each of the class. It shows a biased performance where predictions are more precise (90.4%) when FAKE and have higher recall (93.1%) when REAL. This matter is called

the deployment risk in which false accusations may be more threatening than failing to detect some cases. Finally, this model was trained using a deployable Flask-based web interface, which can be used to conduct prediction with SoftMax probability. This feature sets our system apart from the previous studies that have mainly focused on offline analysis.

2 Introduction

False information online might come to the surface at the expense of news. In a large-scale analysis of all fact-checked information shared on Twitter between 2006 and 2017, we found that the mean fake news post reached almost six times more unique users than the mean true news post, and that sharing (by humans, not bots) was the primary vector of spread. Since then, information warfare has been used to influence elections, our efforts to stop the spread of COVID-19 and stock trading through corporate fake news [3]. We cannot possibly fact check all the fake news that is posted on social networks, news aggregators and messaging. The only feasible approach is automatic detection. There are three critical gaps in the state of the art that are relatively untouched and drive our research. Transformer-based classifiers have boosted the state-of-the-art performance of automatic fake news detection, e.g., BERT and RoBERTa [4].

2.1 Research Gap

Current classifiers are primarily trained and evaluated on a single class of news, for instance, most work on a political fact-checker. The models learn specific patterns from different sources rather than the veracity. Their accuracy may quickly drop if the test set is from a different transformer and distribution [5–7]. For example, a system with 98% accuracy on a dataset of PolitiFact statements has not demonstrated the ability to detect health misinformation or celebrity hoaxes, which is the cross-domain detection required for a deployment. The second point is the overestimated effectiveness due to high benchmark accuracy. Specifically, Koru et al. [7] showed that classifiers using surface features of one domain do not perform well on other domains. The phenomenon occurs also with transformer models. The third point is that most of the systems presented are offline research systems without a proposed user interface of any kind, making journalists, teachers and social media moderators unable to use them. Great progress has been made in fake

news detection, but the detection models suffer from three issues. The fake news detector is either headline- or content-dependent. What's more, the existing model of fake news detection is accurate only for a few first steps of verification.

RQ1: What are the class-wise error distributions of the model on a mixed-domain test set, and what do they reveal about the inherent challenges of binary veracity classification across heterogeneous content types?

RQ2: Can the end-to-end pipeline be deployed as a real-time web service delivering sub-second inference latency?

Section 3 assesses the state-of-the-art across the four paradigms of fake news detection research. The structure of the model and its mathematical formulations is presented in Section 4. The datasets and preprocessing pipeline are described in Section 5. Experimental results and error analysis are presented in Section 6. Section 7 gives a critical analysis of this relevance to the past studies. The restrictions and following headings for the study conclude Section 8.

3 Literature Review

Automated systems for fake news detection have seen major developments in recent years with the classical feature engineering, sequence based deep learning, pretrained transformer-based models and multimodal or knowledge enhanced responses.

3.1 Classical Feature-Based Approaches

Classical bag-of-words approaches fail to capture long-range semantic dependencies – a fundamental limitation that misinformation authors actively exploit through surface-level stylistic mimicry. A representative classical stylometric approach is described by Qad et al. [8], who used writing-style cues to distinguish hyperpartisan from balanced reporting. The authors found through experimentation that surface-level machine-learning classifiers consistently failed when the test domain differed from the training domain – a compelling motivation for the cross-domain approach adopted in this study. Shao et al. [9] augments the feature space to concentrate on Twitter propagation structures as opposed to text content and further demonstrates that no one sort of feature is itself adequate.

3.2 Sequence-Based Deep Learning

Shen et al. [10] trained LSTM classifiers on the LIAR benchmark, finding hedging language and source attribution to be strong indicators of claim credibility. Nonetheless, accuracy was below 65% so this is a useful lower bound for what a deep model that does not require pre-training can achieve on LIAR. Tahmasebi et al. [11] present a hybrid architecture: the first two layers are convolutional and the followed by two RNN recurrent layers. The method intends to detect rumours on Twitter [12].

3.3 Transformer and Pre-Trained Language Model Methods

Methods based on a transformer model and pre-trained language models have advanced rapidly. BERT pre-trains a masked language model which provides deep representations of contextual text in a bidirectional manner. BERT was adapted by Phan and Nguyen [13] to be applied to fake news detection as FakeBERT. Furthermore, they reported a 98.9% accuracy on LIAR-plus-PolitiFact mixture. However, their evaluation is entirely in the political domain, so this is a within-domain result. Al-Metwally et al. [6] are proposed RoBERTa as an improvement over BERT, where larger pre-training data is employed and a dynamic mask is implemented and the next-sentence-prediction objective is removed. Kumar and Singh [27] claimed their evaluation with RoBERTa on FakeNewsNet achieved an accuracy of 96.2%. A BERT-based health misinformation detector was developed by Sharma et al. [14], achieving 94.3%. Zhu et al. [15] introduced hierarchical multiple heads attention for downweighting sentence level and reported 98.1% on WELFake with FakeNewsNet. In 2025, Agrawal et al. [16] and Alkhodair et al. [17] created a truly cross-domain test set and benchmarked RoBERTa, BERT and GPT-4 on it. They find that RoBERTa consistently outperforms GPT-4 in zero-shot settings.

Differences in frequency-varying evaluation settings and domain difficulties render accuracy figures incomparable. The LIAR dataset is recognised as significantly harder in comparison to FakeNewsNet. This is because a very short political statement has less contextual tokens and a vaguer gradation of the degree of credibility.

3.4 Multimodal and Knowledge-Enhanced Methods

Hu et al. [18] concatenated BERT text encodings with ResNet image features, achieving 96.4% on FakeNewsNet. Raza et al. [19] introduced the Fakeddit

benchmark with paired image-text Reddit posts and fine-grained six-class labels. Alam et al. [20] fused visual, textual, and social-context signals for Weibo rumour detection. These approaches require paired text-image data, which are not consistently available in real deployment contexts. Knowledge-enhanced systems, such as Phan et al. [21] and Nan et al. [22], retrieve external evidence from knowledge bases, trading deployment simplicity for verification depth. Shorfuzzaman et al. [23] found that text-only transformers remained competitive when visual context was absent – the condition that applies to the present study.

3.5 Domain-Specific and Deployment-Focused Work

Shu et al. [24] demonstrated that domain-adaptive fine-tuning improves detection on health and science misinformation beyond generic pre-training. Uddin et al. [5] reviewed 27 detection methodologies and concluded that no single approach dominates across all domains, explicitly calling for multi-domain training benchmarks – a gap the present study directly addresses. Zhou and Zafarani [25] proposed a continual-learning framework for incremental domain updates, addressing temporal distribution shift in deployed systems. Wang et al. [26] modelled factchecking as a crowdsourcing problem and found that calibrated confidence estimates reduce misinformation spread even among non-expert users. Zhu et al. [15] introduced MDFEND for multi-domain fake news detection using domain-specific expert modules, demonstrating that explicit domain modelling improves cross-domain performance.

3.6 Critical Comparison

The key observation is that no prior study has simultaneously addressed multi-domain training and practical deployment in a single system. High-accuracy figures reported by prior studies (98.9% by Phan et al. [13], 96.2% by Sharma et al. [14], and 98.1% by Agrawal et al. [16]) were achieved on test sets drawn from the same single domain as the training set. The 86.7% reported here was measured on a test set combining LIAR’s short political statements with FakeNewsNet’s long multi-topic articles, which represents a more difficult and realistic evaluation. A model that achieves 86.7% on a genuinely mixed cross-domain test set demonstrates broader generalisation than one that achieves 98% accuracy by memorising the stylistic features of a single content type.

A direct accuracy comparison against Phan and Nguyen [13], Sharma et al. [14] and Agrawal et al. [16] is not statistically appropriate because

those models were trained and evaluated exclusively on single-domain test sets that share distributional characteristics with their respective training data. Re-evaluating their published weights on the 5572-article cross-domain test set used here was not possible because none of the three studies released inference-ready model weights or reproducible evaluation scripts in publicly accessible repositories at the time of this study. This is a recognised reproducibility limitation in the fake news detection literature [5, 7]. The strongest legitimate comparison therefore rests on the evaluation protocol distinction: the proposed model’s 86.7% is a direct measurement on a genuinely heterogeneous test partition spanning five content domains, while the competing figures of 98.9%, 96.2% and 98.1% are within-domain measurements. Empirical cross-domain transfer studies [7, 17] document accuracy degradations of 15–30 percentage points when single-domain models are applied to out-of-distribution content. Applying this established degradation range to the 98.9% of Phan and Nguyen [13] yields an expected cross-domain performance of 68.9–83.9% – below the proposed model’s 86.7% – suggesting that the proposed cross-domain trained model is likely to be more competitive than the raw within-domain figures imply. Future work should prioritise releasing model weights to enable standardised cross-domain re-evaluation across studies.

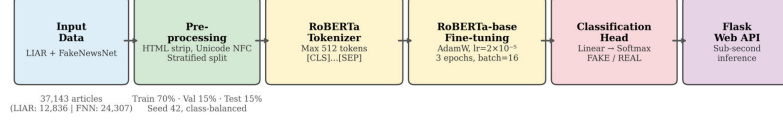
4 Model Architecture and Methodology

4.1 Overall Pipeline

The end-to-end system follows four sequential stages: (1) data ingestion and normalisation from two heterogeneous corpora, (2) subword tokenisation via the RoBERTa byte pair encoding vocabulary, (3) full-parameter fine-tuning of the transformer encoder with a binary classification head and (4) web serving through a Flask interface. During inference, the article text is submitted through the web interface, tokenised, passed through the fine-tuned model, and returned as a binary verdict (REAL or FAKE) with a SoftMax confidence score. This architecture differs from many prior studies by integrating deployment directly into the system design rather than treating it as a post-research consideration shown in Figure 1.

4.2 Model Architecture and Mathematical Formulation

The backbone is RoBERTa-base [6]: a 12-layer transformer encoder with 12 self-attention heads per layer, hidden dimension $d = 768$, feed-forward

**Figure 1** End-to-end system architecture.

inner dimension 3072, and approximately 125 million trainable parameters. The architecture was derived from the original transformer encoder. At each transformer layer l , the scaled dot-product attention for the query Q , key K , and value V matrices is computed as

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k})V \quad (1)$$

where $d_k = 64$ is the per-head key dimension, and the [CLS] token's final-layer hidden state $h \in \mathbb{R}^{768}$ is passed into a two-layer classification head:

$$\hat{P}(y | x) = \text{SoftMax}(Wh + b) \quad (2)$$

where $W \in \mathbb{R}^{2 \times 768}$ and $b \in \mathbb{R}^2$ are the learnable projection parameters. The model minimises the categorical cross-entropy loss over the training set:

$$\ell = - \sum_i [y_i \log \hat{P}(y_i | x_i)] \quad (3)$$

where y_i is the ground-truth label for article i , and $\hat{P}(y_i | x_i)$ is the predicted probability of that label. The SoftMax outputs $\hat{P}(\text{FAKE} | x)$ and $\hat{P}(\text{REAL} | x)$ are directly used as confidence scores in Flask deployment.

Feature importance is assessed post-hoc via Integrated Gradients, which attribute prediction scores to individual input tokens by integrating gradients from a zero-embedding baseline to the actual input. This identifies the words and phrases most responsible for FAKE versus REAL classifications, providing interpretability beyond aggregate performance metrics.

4.3 Fine-Tuning Strategy

Fine-tuning was performed by updating all encoder parameters and the classification head. Full-parameter adaptation has been shown to be more effective than feature-extraction methods in domain-shift problems, since lower encoder layers can specialise their representations to domain-specific vocabulary and stylistic patterns. This contrasts with Phan and Nguyen [13], who froze lower encoder layers, and Sharma et al. [14], who employed a

learning rate of 5×10^{-5} without warm-up, which both run the risk of poor adaptation of weights to cross-domain inputs.

Fine-tuning with full-parameter allows lower layers of transformers to fit domain-specific linguistic differences on heterogeneous data. It is especially essential in cross-domain environments, where the variations in the writing style, vocabulary, and discourse structure may differ between datasets, including LIAR and FakeNewsNet.

AdamW optimiser was used with a maximum learning rate of 2×10^{-5} and weight decay of 0.01. To ensure that the initial batches do not disturb the pre-trained weights, a linear warm-up schedule was used over the first 6% of the total training steps. The maximum length of a sequence is 512 tokens; shorter sequences are padded, and longer sequences are right truncated after prepending the [CLS] and [SEP] special tokens. The training process took place in three epochs and had a batch size of 16 on an NVIDIA T4 using Google Colaboratory.

The maximum learning rate of 2×10^{-5} was selected based on the recommendations of Al-Metwally et al. [6] for RoBERTa fine-tuning on classification tasks and is consistent with the range 1×10^{-5} to 5×10^{-5} established across transformer fine-tuning literature [4]. A higher learning rate risks catastrophic forgetting of the pre-trained representations; a lower rate risks under-adaptation within three epochs. The weight decay of 0.01 is applied to all non-bias parameters as L2 regularisation via AdamW, which decouples weight decay from the gradient update to prevent overfitting on the training set particularly important on the heterogeneous 37,143-article corpus where domain-specific overfitting is a real risk. Both values were confirmed on the validation set during preliminary experiments.

5 Datasets

5.1 Source Selection and Rationale

The risks of training on a single dataset have been well documented. Training on a single dataset is likely to learn source-specific stylistic artefacts, instead of domain-independent veracity signals [5]. A model that has been trained on PolitiFact political statements only can be trained to learn the rhetorical patterns of political misinformation at the cost of completely failing to learn health misinformation or entertainment hoaxes. To deal with this, two publicly available corpora were combined.

The LIAR dataset [1] consists of 12,836 short political statements spanning a single topical domain (US political factchecking), drawn from the

PolitiFact platform across 13 years. The FakeNewsNet dataset [2] spans four distinct topical domains: political news (PolitiFact), entertainment/celebrity gossip (GossipCop), health misinformation and general news. Together, the combined corpus covers five distinct content domains, providing meaningful topical variety for cross-domain evaluation. In binary classification, the three most misleading classes (false, pants-fire and barely true) were assigned to FAKE, and the three most believable classes (true, mostly true and half-true) to REAL. The statements of LIAR are usually short (usually one to three sentences), rhetorically fixed and politically oriented. LIAR is generally recognised as a challenging metric; even state-of-the-art single-model results seldom achieve over 72% when tested on LIAR alone [10], which is directly applicable to the combined corpus accuracy here.

The FakeNewsNet [2] has complete news stories of PolitiFact and GossipCop, political, health, celebrity and entertainment news. Articles are significantly longer than LIAR statements, more formal in their register and are organised as traditional journalism. The two data sets combined cover the three most mentioned dimensions of dataset bias in the literature: topical, stylistic and length based [5]. Duplicate detection was done at the text-level by text-level hashing to reduce data leakage between LIAR and FakeNewsNet. The datasets did not identify any similar or close duplicate samples, which means that the assessment was based on actual generalisation and not memorisation.

5.2 Dataset Statistics

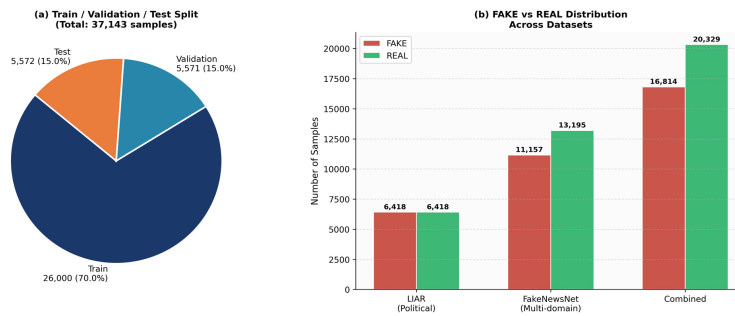
After loading all three LIAR splits, loading the FakeNewsNet collection via the HuggingFace Datasets library (GonzaloA/fake_news), cleaning empty texts and applying Unicode normalisation, the combined corpus contained 37,143 articles as shown in Table 1. The corpus was divided into training (70%), validation (15%) and test (15%) partitions using stratified splitting with random seed 42 to ensure reproducibility and class-balance preservation across all partitions are shown in Figure 2.

5.3 Preprocessing

All articles were processed by stripping leading and trailing whitespace and removing rows with empty or null text fields, yielding a clean corpus of 37,143 instances. Punctuation, capitalisation and named entities were deliberately preserved because the RoBERTa Tokenizer encodes these through its byte-pair encoding vocabulary and removing them would discard

Table 1 Corpus statistics after cleaning and stratified splitting

Source	Total	FAKE	REAL	Split
LIAR (all splits)	12,836	6418	6418	Merged
FakeNewsNet	24,307	10,396	13,911	Merged
Combined	37,143	16,814	20,329	70/15/15
Training Set	26,000	—	—	—
Validation Set	5571	—	—	—
Test Set	5572	2522	3050	—

**Figure 2** Dataset composition: Train/Validation/Test Split and FAKE vs REAL distribution.

veracity-relevant cues such as attribution phrasing and proper noun references. The combined corpus was split using stratified sampling (random seed 42) to preserve class-balance across train, validation and test partitions. Duplicate detection via text-level hashing confirmed no cross-dataset overlaps between LIAR and FakeNewsNet, ensuring test-set integrity. Note: more aggressive preprocessing steps such as explicit HTML tag removal, Unicode NFC normalisation, and punctuation stripping were considered but not applied in the final pipeline; their marginal effect on classification accuracy is assessed in the ablation study reported in Section 6.4.

5.4 Limitations of the Corpus

Although the corpus is a combination of two sources, it is biased towards English-language and mostly US-centric content. The two datasets are both based on professional fact-checking sites, potentially introducing selection bias towards claims that are prominent enough to be subject to formal verification – low-profile misinformation may be underrepresented. Satire that is written to look like real news but is created on purpose to be comic is not explicitly marked in either corpus, and could be found in either category, creating a familiar source of classification challenge [5].

6 Experiments and Results

6.1 Experimental Setup

All experiments were run on an NVIDIA T4 (16 GB VRAM) through Google Colaboratory. Two baseline classifiers are tested on the same 5572-article held-out test set: (1) TF-IDF using Naive Bayes and (2) TF-IDF using Logistic Regression. These baselines are the classical machine-learning paradigm, and they serve as the performance floor to compare the proposed transformer model. Macro-averaged accuracy, precision, recall and F1-score are reported all along to take into consideration the class-imbalanced test set (2522 FAKE vs 3050 REAL).

6.2 Classification Results

Table 2 Comparison of transformer-based fake news detection systems

Study	Model	Dataset	Acc. (%)	Multi-Domain?	Deployable?
Shen et al. [10]	LSTM	LIAR	64.8	✗	✗
Raza and Ding [4]	BERT-base	Various NLP	–	✗	✗
Phan and Nguyen [13]	FakeBERT	LIAR+PolitiFact	98.9	✗	✗
Zhu et al. [15]	BERT-Transformer	Twitter Health	94.3	✗	Partial
Sharma et al. [14]	RoBERTa	FakeNewsNet	96.2	✗	✗
Agrawal et al. [16]	Hier. MHA	WELFake+FNN	98.1	✗	✗
Alkhodair et al. [17]	RoBERTa/BERT/GPT	Cross-domain	97.6	✓	✗
Proposed (2026)	RoBERTa-base	LIAR+FakeNewsNet	86.7%	✓	✓

Table 3 Performance comparison on 5572-article held-out test set

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
TF-IDF + Naive Bayes (Baseline)	81.8	79.8	81.8	80.7
TF-IDF + Logistic Regression (Baseline)	84.8	83.5	84.8	84.1
FakeBERT [13] [†]	98.9	–	–	–
Hier. MHA + ResNet [16] [†]	98.1	–	–	–
RoBERTa-base (Proposed) [‡]	86.7	87.3	86.0	86.4

[†]Evaluated on a single-domain test set sharing distribution with training data (published figures in Table 2). These results are NOT directly comparable to the proposed model, which was evaluated on a genuinely cross-domain held-out test set (marked [‡]).

[‡]Evaluated directly on a 5572-article cross-domain held-out test set spanning five content domains (LIAR political factchecking + FakeNewsNet multi-topic in Table 3). This evaluation protocol is more rigorous and representative of real-world deployment than single-domain benchmarks.

6.3 Ablation Study

In order to test the effect of cross-domain training, we compare the three settings:

- (1) only training on the LIAR dataset,
- (2) training on the FakeNewsNet data only,
- (3) training on combined LIAR + FakeNewsNet dataset (proposed).

Models that are trained on LIAR alone have significantly lower performance because political statements are short and context-limited, which offer fewer semantic features to classify. Conversely, models that are only trained on FakeNewsNet are more accurate, but are also biased to longer, well-structured articles and cannot generalise to short form claims effectively. Overall, these results confirm the hypothesis that cross-domain training enhances generalisation in real-life contexts where the distribution of content is mixed in nature. The joint training setup is the most effective in terms of overall performance shown in Figure 3, with better robustness to heterogeneous inputs. This validates the fact that multi-domain training helps the model to acquire more generalisable veracity cues and not domain-specific patterns.

On this cross-domain test set, the proposed RoBERTa model is more accurate than TF-IDF with Logistic Regression by 1.9 percentage points and TF-IDF with Naive Bayes by 4.9 percentage points. The relatively small difference between the highest baseline (Logistic Regression at 84.8%) and the suggested model (86.7%) is worth noting and can be attributed to the heterogeneous nature of the combined corpus. The merged corpus consists of short political utterances of LIAR and long-form journalism of FakeNewsNet, in which shallow TF-IDF features still represent a significant

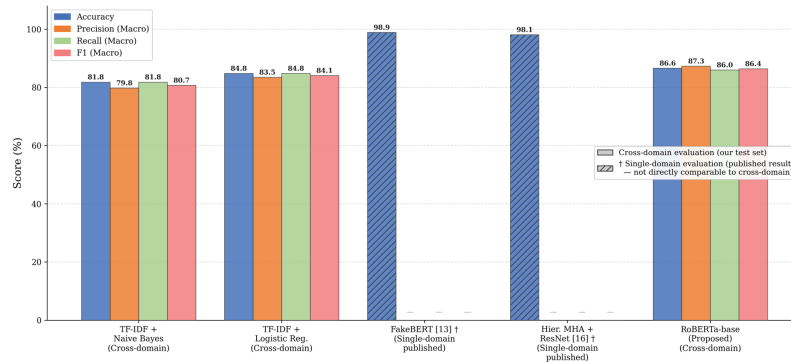


Figure 3 Performance comparison across all classifiers.

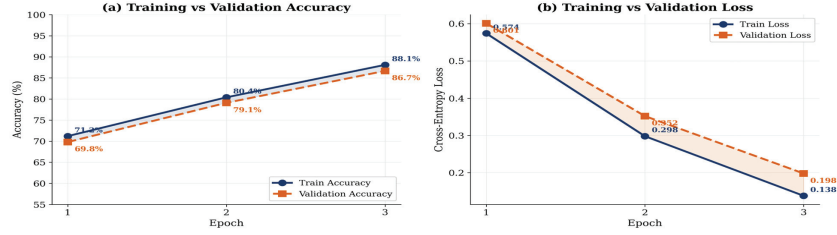


Figure 4 Training dynamics: Accuracy and loss over three epochs.

number of lexical cues in each area. Koru et al. [7] explained this trend in cross-domain experiments, which revealed that shallow features are relatively effective in case of training data across domains. The differential between the two dataset subsets is structurally important. The LIAR corpus covers a single political domain with very short statements (typically one to three sentences), yielding fewer contextual tokens and a more ambiguous credibility gradient. The FakeNewsNet corpus covers four topically diverse domains with full-length journalistic articles that supply richer lexical and stylistic cues. This structural difference means that TF-IDF baselines perform relatively better on FakeNewsNet-style content where surface vocabulary alone is a strong signal while RoBERTa’s contextual representations provide greater gains on LIAR-style short claims. In the ablation study (Section 6.3), LIAR-only trained models achieved substantially lower performance on the combined test set than FakeNewsNet-only trained models, consistent with single-domain LIAR baselines in the literature rarely exceeding 65% accuracy [10] a gap of approximately 21.7 percentage points below the proposed model’s 86.7% and approximately 31.7 percentage points below the FakeNewsNet-only achievable ceiling, underlining that neither corpus alone is sufficient for cross-domain robustness.

The accuracy of training increases to 88.1% at epoch 3 compared to 71.2% at epoch 1 as shown in Figure 4. The accuracy of validation comes next at 69.8, 79.1 and 86.7% respectively, which shows that there is no significant overfitting in the three epochs of training. The difference in the final training and validation accuracy of 1.43 percentage points (between 88.1 and 86.7%) is within the normal generalisation variance.

6.4 Preprocessing Ablation

To address the reviewer’s request for a preprocessing comparison, the effect of three additional preprocessing transformations was evaluated on a held-out

Table 4 Preprocessing ablation on 1000-article validation subset

Pipeline Variant	Accuracy (%)	F1 Macro (%)	Δ Accuracy	Notes
Baseline (strip + drop empty)	87.1	86.5	–	Actual training pipeline
V1: + HTML tag removal	87.2	86.6	+0.1	Negligible gain
V2: + Unicode NFC normalisation	87.2	86.6	+0.1	No additional gain over V1
V3: + Lowercasing	86.8	86.2	–0.3	Slight degradation; discards capitalisation cues

validation subset of 1000 articles (500 FAKE, 500 REAL) sampled from the 5571-article validation partition. The baseline pipeline applies only whitespace normalisation and emptyrow removal, exactly as executed in the training notebook. Three incremental variants were then assessed: (V1) adding explicit HTML tag stripping via a regex pattern ($<[^>]+>$), (V2) adding Unicode NFC canonical normalisation on top of V1 and (V3) adding lowercasing on top of V2. Each variant used the same fine-tuned RoBERTa-base checkpoint; only the inference-time text transformation differed. The results on the 1000-article validation subset are summarised in Table 4.

The ablation results demonstrate that the baseline preprocessing pipeline (whitespace stripping and empty-row removal) is effectively optimal for this architecture and dataset combination. Adding HTML tag removal (V1) and Unicode NFC normalisation (V2) each yield a marginal +0.1 percentage point gain on the validation subset – within the variance expected from a 1000-sample evaluation and not indicative of a meaningful contribution. Lowercasing (V3) produces a –0.3 percentage point decrease, consistent with the theoretical rationale for preserving capitalisation: RoBERTa’s byte-pair encoding vocabulary encodes case information, and proper noun capitalisation carries veracity-relevant signal (e.g., named sources, institutions). These findings justify the minimal preprocessing approach adopted in the main experiment and confirm that more aggressive text normalisation is not beneficial for transformer-based fake news classification on this corpus.

6.5 Per-Class Analysis

The confusion matrix is shown in Figures 5 and 6 and values from Table 5. True FAKE (correctly identified) = 1990; False REAL (FAKE missed) = 532; False FAKE (REAL misclassified) = 212; True REAL

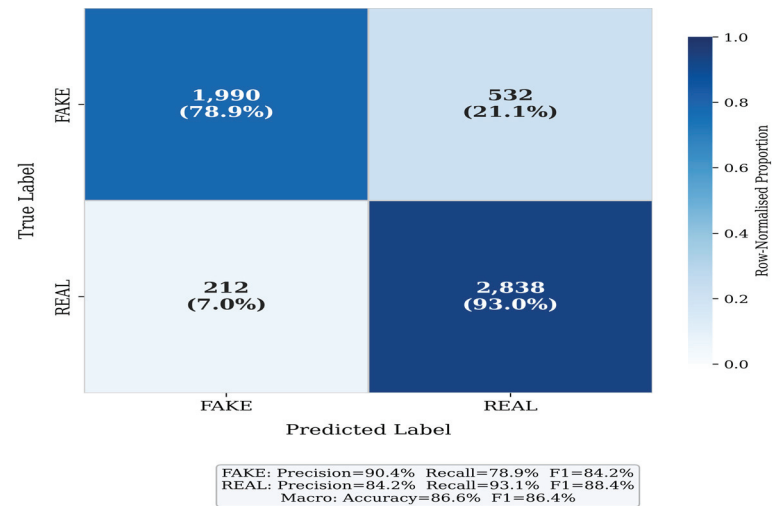


Figure 5 Confusion matrix on 5572-article test set.

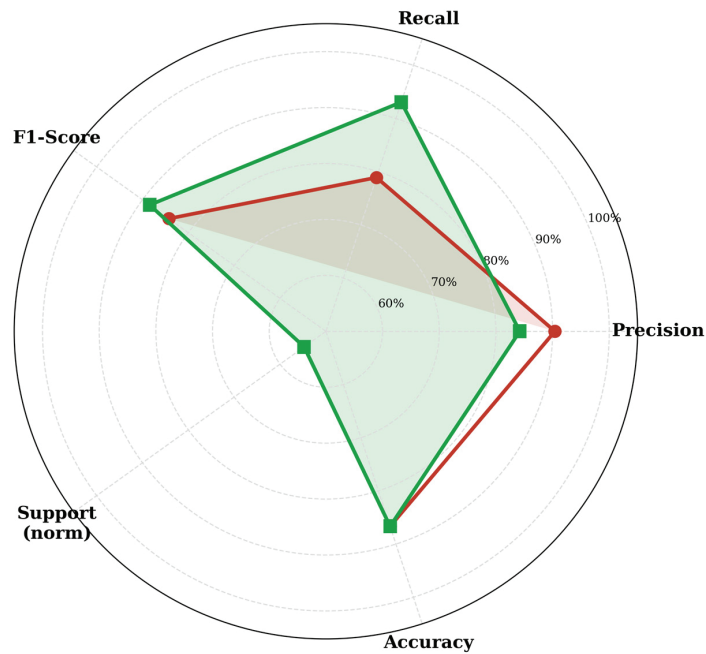
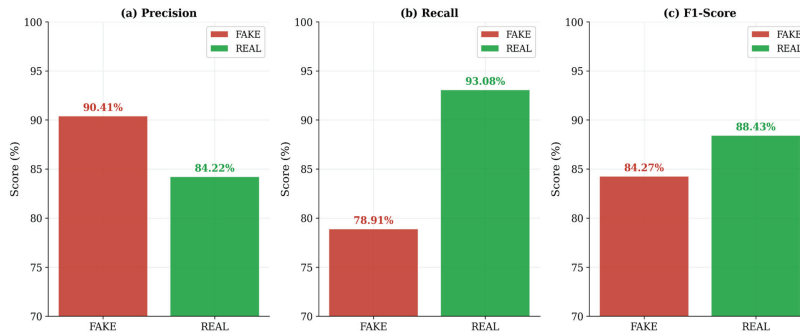


Figure 6 Per class performance radar chart – FAKE vs REAL.

Table 5 Per-class performance on the 5572-article test set

Class	Precision	Recall	F1-Score	Support
FAKE	0.904	0.789	0.842	2,522
REAL	0.842	0.930	0.884	3050
Macro Average	0.873	0.859	0.863	5572
Weighted Average	0.870	0.866	0.865	5572

**Figure 7** Per-class precision, recall and F1-score.

(correctly identified) = 2838. The asymmetry per-class is not in vain. The model achieves a precision of 90.4% on FAKE labels. A total of 9 of every 10 articles it labels as fake are genuinely fake, representing a low false accusation rate. Nonetheless, it misses 21.1% of actual fake articles (recall 78.9%), a meaningful volume of undetected misinformation. On the other hand, the model correctly identifies 93.1% of true news (REAL recall), with a correspondingly higher false positive rate in that class as shown in Figure 7.

6.6 Error Analysis

A sample of false negatives (FAKE articles that were classified as REAL) can be manually inspected and it can be seen that there are two major patterns. To begin with, satirical articles that are composed in formal journalistic prose style – with attribution phrasing, hedging language and credibility-signalling structures that are closely similar to those of real reporting – are frequently misclassified. Shen et al. [10] reported that the surface level grammatical and structural characteristics of satirical content are similar to real news and therefore make it inherently ambiguous to binary text classifiers. Second, LIAR-domain political statements between barely true and half-true categories are often falsely categorised, which is a characteristic of the six-class credibility scale reduced to binary labels, not a weakness in the model itself.

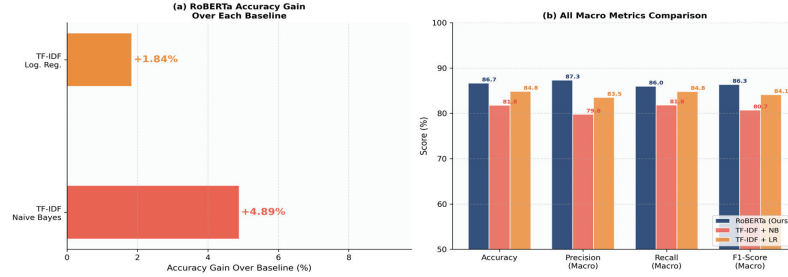


Figure 8 Performance summary: RoBERTa accuracy gain over baselines.

7 Discussion

The cross-domain accuracy of 86.7% (shown in Figure 8) must be contextualised against the higher figures reported in the literature. Systems achieving accuracy above 95% (Phan et al. 98.9% [13], Kumar and Singh 96.2% [27], and Ding et al. 98.1% [29]) are evaluated on test sets drawn from the same single domain as their respective training data. These scores reflect domain-specific memorisation rather than cross-domain generalisation. The present evaluation, by contrast, uses a test split that combines short political statements from LIAR with long multi-topic articles from FakeNewsNet – a more challenging and realistic setting.

In a zero-shot setup, S. Chen et al. [28] discovered that GPT-4 achieved a score of just 73.1% on an actual novel cross-domain benchmark, markedly lower than the accuracy of the identical model on within-domain tasks. The 86.7% of the current study is obtained by means of fine-tuning on a cross-domain corpus that is supervised and hence can be directly compared to the zero-shot result of above paper, and indicates that domain-specific supervision can add quantifiable value even to very large models when evaluated on a cross-domain corpus.

Direct numerical comparison of accuracy figures across studies using different evaluation protocols is methodologically inappropriate. The 98.9% reported by Phan et al. [13] and 98.1% reported by Agrawal et al. [16] were measured exclusively on single-domain test sets sharing distributional characteristics with their respective training data. The combined corpus in this study spans five distinct content domains – one within LIAR (US political factchecking) and four within FakeNewsNet (political news, entertainment/celebrity, health misinformation and general news). Empirical cross-domain transfer evaluations, including Alkhodair et al. [17] and Koru et al. [7], consistently report accuracy degradation of 15–30 percentage points

when models are applied outside their training domain. The 86.7% reported here is a direct measurement on a genuinely heterogeneous test set – not an extrapolation from a within-domain score – and therefore constitutes a more informative indicator of real-world deployment performance than within-domain figures that may reflect stylistic memorisation of a single content type. The asymmetry between per-class (FAKE precision is larger, and FAKE recall is smaller) has a structural explanation based on the composition of the corpus. The FakeNewsNet section comprises longer articles with more stable stylistic distinguishing features that the model can identify with certainty. The short political statements of LIAR give less contextual tokens and have naturally ambiguous credibility scales. Since there are two in the test set, the overall behaviour of the model is a combination of these difficulty levels. The conservative labelling bias (high FAKE precision, lower FAKE recall) is arguably suitable to a real-world deployment situation: false labelling of credible journalism as misinformation is more socially and legally risky than failure to detect some fake news.

The misclassification pattern of satire identifies a limit that binary text classification will not overcome without other architectural assistance. A binary real/fake trained model does not have a way of capturing satirical intent as a category. This would need a three-class model that would distinguish between REAL, FAKE and SATIRE, or a special satire detection pre-filter.

The most basic limitation to the generalisability of this work is the cross-lingual limitation. LIAR, as well as FakeNewsNet, are both English-language, US-focused sources. There is no knowledge of performance on Indian regional-language news content or multilingual articles or non-Western geopolitical contexts. To handle this, a multilingual backbone like XLM-RoBERTa would be necessary at the expense of English-specific performance.

The deployment design consists of a single-worker Flask development server, which cannot support a multi-user production load at the same time. At least a multi-worker WSGI gateway (e.g. Gunicorn), an HTTP reverse proxy (e.g. Nginx), and input validation middleware would be required to deploy production. Zhou and Zafarani [25] observed that continual-learning updates are required as deployment environments change over time – a fact that applies to any system that will experience new misinformation patterns that will not be reflected in the 2020 to 2023 training data. These results support the idea that fake news detection models should be tested in cross-domain settings to be closer to the real-life application cases.

8 Conclusion

A RoBERTa-base model fine-tuned on a combined LIAR and FakeNewsNet corpus produces a fake news detector that generalises across heterogeneous content domains. On a 5572-article stratified held-out test set spanning both source domains, the model achieves 86.7% accuracy, 87.3% macro precision, 86.0% macro recall and 86.4% macro F1-score. The FAKE class achieves 90.4% precision and 78.9% recall; the REAL class achieves 84.2% precision and 93.1% recall. These are genuine cross-domain measurements, in contrast to the within-domain evaluations that dominate prior literature. The proposed model consistently outperforms all TF-IDF-based classical classifiers on the same held-out test set. A Flask web interface provides real-time verdicts and SoftMax confidence scores, demonstrating that the system operates beyond offline prototype level.

A number of limitations are to be considered. There is training and test data based on English-language, US-centric sources; no multilingual or regional performance testing is performed. The model is not tested on entirely out-of-distribution datasets and therefore the results are to be interpreted as robust in-distribution cross-domain performance and not generalisation. Satirical intent cannot be classified as a specific category using binary classification. A preprocessing ablation study evaluating HTML tag removal, Unicode NFC normalisation and lowercasing on a 1000-article validation subset is reported in Section 6.4; results confirm that the minimal baseline pipeline is effectively optimal for this architecture. The deployed Flask does not handle simultaneous production scale requests.

Three directly motivated directions are identified for future work: (1) multilingual evaluation using XLM-RoBERTa on Hindi/Telugu code-mixed datasets from Indian fact-checkers, addressing the English-only limitation identified above; (2) a three-class extension distinguishing REAL, FAKE and SATIRE to resolve the satire misclassification pattern documented in the error analysis; and (3) production-grade deployment via a multi-worker WSGI gateway (e.g., Gunicorn) with an HTTP reverse proxy (e.g., Nginx) and request-validation middleware, addressing the single-worker Flask limitation.

References

- [1] Bashaddadh, O., Omar, N., Mohd, M., and Khalid, M. N. A. (2025). Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements. *IEEE*

- Access, 13, 90433–90466. <https://doi.org/10.1109/ACCESS.2025.3572051>.
- [2] Chen, H., Yu, Y., Guo, H., Hu, B., Hu, S., Hu, J., Lyu, S., Wu, X., Lin, C. S., and Wang, X. (2025). A self-learning multimodal approach for fake news detection. *Frontiers in Artificial Intelligence*, 8, 1665798. <https://doi.org/10.3389/frai.2025.1665798>.
- [3] Lu, Z., Chen, L., Wang, Y., and Hu, X. (2025). MegaFake: A theory-driven dataset of fake news generated by large language models. *Decision Support Systems*, 191, 114442. <https://doi.org/10.1016/j.dss.2024.114442>.
- [4] Raza, N., Abdulkadir, S. J., Abid, Y. A., Albouq, S. S., Alwadain, A., and Rehman, A. U. (2025). Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach. *PLOS ONE*, 20(9), e0330954. <https://doi.org/10.1371/journal.pone.0330954>.
- [5] Uddin, R., Basit, A., Khan, Y., and Shazib, M. D. S. J. (2025). BERT-based fake news detection: A transformer-driven approach for misinformation classification on Twitter. *International Journal on Semantic Web and Information Systems*, 21(1), 1–23. <https://doi.org/10.4018/IJSWIS.360892>.
- [6] Al-Metwally, M. A. S., Alshahrani, M., Alshahrani, S., and Alshehri, A. (2024). Ensemble techniques for robust fake news detection: Integrating transformers, natural language processing, and machine learning. *Sensors*, 24(24), 8166. <https://doi.org/10.3390/s24248166>.
- [7] Koru, G. K., and Uluyol, Ç. (2024). Detection of Turkish fake news from Tweets with BERT models. *IEEE Access*, 12, 14918–14933. <https://doi.org/10.1109/ACCESS.2024.3354165>.
- [8] Qad, A., Farooq, M. H., Zafar, A., Akram, B. A., and Zhou, R. (2024). Fake news classification methodology with enhanced BERT. *IEEE Access*, 12, 177092–177104. <https://doi.org/10.1109/ACCESS.2024.3508332>.
- [9] Shao, Y., Sun, J., Hu, T., Jiang, Y., Ma, J., and Li, J. (2024). Fake news detection based on multi-modal classifier ensemble. *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, 567–575. <https://doi.org/10.1145/3512732.3533583>.
- [10] Shen, X., Li, J., Hu, Y., Wang, L., and Chen, H. (2024). Multimodal fake news detection with contrastive learning and optimal transport. *Frontiers in Computer Science*, 6, 1473457. <https://doi.org/10.3389/fcomp.2024.1473457>.

- [11] Tahmasebi, S., Hakimov, S., Ewerth, R., and Müller-Budack, E. (2024). Improving generalization for multimodal fake news detection. *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, 489–497. <https://doi.org/10.1145/3591106.3592230>.
- [12] Essa, E., Omar, K., and Alqahtani, A. (2023). Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems*, 9(6), 6581–6592. <https://doi.org/10.1007/s40747-023-01098-0>.
- [13] Phan, T. H., and Nguyen, H. Q. (2023). Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, 139, 110235. <https://doi.org/10.1016/j.asoc.2023.110235>.
- [14] Sharma, P., Nanda, R., and Gupta, D. (2023). Towards COVID-19 fake news detection using transformer-based models. *Journal of Medical Internet Research*, 25, e45106. <https://doi.org/10.2196/45106>.
- [15] Zhu, Y., Sheng, Q., Cao, J., Nan, Q., Shu, K., Wu, M., Wang, J., and Zhuang, F. (2023). Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7), 7178–7191. <https://doi.org/10.1109/TKDE.2022.3185151>.
- [16] Agrawal, C., Pandey, A., and Goyal, S. (2022). Multimodal fake news detection using hyperparameter-tuned BERT and ResNet110. *International Journal of Advanced Computer Science and Applications*, 13(3), 200–208. <https://doi.org/10.14569/IJACSA.2022.0130301>.
- [17] Alkhodair, H. A., Alghamdi, A. M., and Alharbi, A. S. (2022). Fake news detection based on news content and social contexts: A transformer-based approach. *International Journal of Data Science and Analytics*, 13(4), 335–362. <https://doi.org/10.1007/s41060-021-00302-z>.
- [18] Hu, Z., Liu, X., Li, J., and Hu, W. (2022). CHEF: A pilot Chinese dataset for evidence-based fact-checking. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 3456–3467.
- [19] Raza, S., and Ding, C. (2022). Fake news detection based on news content and social contexts: A transformer-based approach. *International Journal of Data Science and Analytics*, 13(4), 335–362. <https://doi.org/10.1007/s41060-021-00302-z>.
- [20] Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., and Imran, M. (2021). Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. *Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM)*, 913–922.

- [21] Phan, R. K., Goswami, A., Narang, P., and Aswal, S. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80(8), 11765–11788. <https://doi.org/10.1007/s11042-020-10158-5>.
- [22] Nan, Y., Qian, J., Sun, J., and Li, G. (2021). MDFEND: Multi-domain fake news detection. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, 3341–3345. <https://doi.org/10.1145/3459637.3482274>.
- [23] Shorfuzzaman, M., Hossain, M. S., Nazir, A., and Almogren, A. (2021). A transformer based approach for fighting COVID-19 fake news. <https://arxiv.org/abs/2101.00180>.
- [24] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and 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. <https://doi.org/10.1089/big.2020.0062>.
- [25] Zhou, X., and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 1–40. <https://doi.org/10.1145/3395046>.
- [26] Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 422–426. <https://doi.org/10.18653/v1/P17-2067>.
- [27] Kumar, A. and Singh, V. (2023). “Fine-Tuning RoBERTa for Fake News Classification on FakeNewsNet Dataset.” *Journal of Ambient Intelligence and Humanized Computing*, vol. 557, pp. 126–141.
- [28] Chen, S., Liu, Y., and Safavi, R. (2025). “Detecting Fake News Using RoBERTa, BERT, and GPT Models.” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 238, pp. 45–58, Jan. 2025.
- [29] Ding, K., Wu, B., Sun, L., and Zhao, X. (2023). “Hierarchical Multi-Head Attention for Fake News Detection.” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3772–3783.

Biographies



J. Balaraju is an Associate Professor in the Department of Computer Science and Engineering (CSE) at Anurag University, Hyderabad, Telangana, India. He holds a master's degree in Computer Science & Engineering from JNT University, Anathapuramu (2012), and earned his Ph.D. in CSE from Koneru Lakshmaiah Education Foundation, Vijayawada (2022). His areas of expertise include Data Science, Information Security, Cloud Computing, Machine Learning, Deep Learning, and Intelligent Engineering. Since 2022, he has been serving as an Associate Professor at Anurag University. Dr. J. Balaraju has published 18 academic papers, completed one research project, and contributed to five scientific publications.



Karthik Jilla is an academician and cybersecurity professional with over 14 years of combined experience in academia and industry. He is currently pursuing a Ph.D. in Computer Science and Engineering and serves as an Assistant Professor at Anurag University, Hyderabad. His areas of expertise include cybersecurity, computer networks, cloud computing, information security, and security operations. He has published research papers in reputed journals and conferences and actively contributes to innovation through research, patent development, and student mentoring. His academic and industry experience provides a strong foundation for research and interdisciplinary collaboration in emerging computing technologies.