International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

A Comprehensive Survey of Deep Learning Techniques for Fake News Detection: Emerging Trends and Future Directions

Ajay S Tippannavar

Ajay S Tippannavar, Department of Artificial Intelligence and Data Science, S J C Institute of Technology, Chikaballapur, India,

---

ABSTRACT - *The exponential proliferation of digital misinformation ("fake news") on social media platforms presents a critical challenge to information integrity and democratic stability. The high velocity and volume of user-generated content render manual verification methods unscalable. This research paper presents a comprehensive, systematic survey of Deep Learning (DL) methodologies employed for automated fake news detection, covering the period from 2022 to 2025. We critically analyse the architectural evolution from Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to state-of-the-art Transformer-based models like BERT and Roberta. Furthermore, we examine the emerging threat of Large Language Model (LLM) generated disinformation. Our comparative analysis demonstrates that while Transformer models achieve superior accuracy (up to 98.4%), they impose significant computational overhead compared to lightweight LSTM architectures. The study concludes by proposing a roadmap for future research, emphasizing Explainable AI (XAI) and Federated Learning for privacy-preserving detection.*

Keywords-Fake News Detection, Deep Learning, Natural Language Processing, Transformers, BERT, Bi-LSTM, Multimodal Learning, Generative AI.

NOMENCLATURE

- **DL:** Deep Learning
- **NLP:** Natural Language Processing
- **CNN:** Convolutional Neural Network
- **LSTM:** Long Short-Term Memory
- **Bi-LSTM:** Bidirectional LSTM
- **BERT:** Bidirectional Encoder Representations from Transformers
- **RoBERTa:** Robustly Optimized BERT Pretraining Approach
- **TP/TN:** True Positive / True Negative
- **FP/FN:** False Positive / False Negative

I. INTRODUCTION

The advent of Web 2.0 has democratized content creation, allowing information to bypass traditional editorial gatekeepers. While this fosters free speech, it has concurrently enabled the weaponization of information. "Fake news" is defined as intentionally fabricated information published with the intent to deceive or mislead. The consequences of such misinformation are profound, ranging from manipulating political elections to inciting social unrest and spreading public health myths.

Research by Vosoughi et al. [1] indicates that falsehoods diffuse significantly farther, faster, deeper, and more broadly than the truth in all categories of information. This phenomenon is driven by the "novelty hypothesis," which suggests that fake news is more novel and emotionally evocative than reality.

Traditional detection approaches relied on **feature engineering**, extracting linguistic features such as n-grams, punctuation analysis, and readability scores. However, these methods are brittle and domain-specific. Deep Learning (DL) automates feature extraction, learning hierarchical representations of data. This paper provides a granular analysis of DL techniques, focusing on sequence modelling, attention mechanisms, and the new challenge of AI-generated text.

II. RELATED WORK

The field of automated deception detection has evolved through several distinct phases.

- **Phase 1: Statistical Machine Learning (2015-2018):** Early approaches utilized Support Vector Machines (SVM) and Naïve Bayes classifiers. These models relied heavily on "bag-of-words" approaches, which ignored the context and order of words.
- **Phase 2: Deep Sequence Modeling (2018-2021):** The introduction of Recurrent Neural Networks (RNNs) allowed models to process text as a sequence. Researchers like Ruchansky et al. developed the CSI model, which combined text analysis with user behavior analysis.
- **Phase 3: The Transformer Era (2021-Present):** The release of BERT by Google revolutionized NLP. Current state-of-the-art research focuses on fine-tuning pre-trained Transformers to detect subtle nuances, sarcasm, and framing in news articles.

III. DATA PREPROCESSING AND EMBEDDINGS

The efficacy of DL models is predicated on the quality of input representations.

A. Preprocessing Pipeline

Raw text undergoes a four-stage cleaning process:

1. **Tokenization:** Segmenting text into discrete units (tokens).
2. **Noise Removal:** Elimination of URLs, HTML tags, and non-ASCII characters.
3. **Stop-word Removal:** Filtering high-frequency, low-meaning words (e.g., "the", "is").
4. **Lemmatization:** Reducing words to their base dictionary form (e.g., "better" -> "good").

B. Word Embeddings (Vectorization)

To process text, discrete tokens must be mapped to continuous vector spaces.

1. **Word2Vec:** Utilizes a shallow neural network to predict a word given its context.
2. **Contextual Embeddings:** Unlike static embeddings, models like BERT generate dynamic vectors where the representation of a word depends on its surrounding context.

IV. DEEP LEARNING ARCHITECTURES

A. Convolutional Neural Networks (CNN)

Originally designed for Computer Vision, CNNs are applied to NLP to extract local features (n-grams).

- Mathematical Operation:

A 1D convolution operation involves a filter (w) applied to a window of words (x) to produce a new feature (c):

c = f(w * x + b)

Where 'b' is the bias and 'f' is a non-linear activation function like ReLU.

B. Bidirectional LSTM (Bi-LSTM)

Standard LSTMs process text from left to right. However, context often depends on future words. Bi-LSTMs utilize two separate hidden layers to process the sequence in both directions.

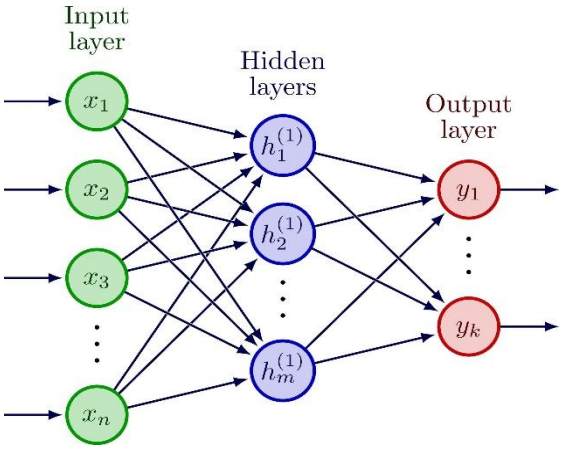

Fig. 1. Architecture of a Neural Network showing Input, Hidden, and Output layers.

Mathematical Formulation:

The LSTM unit consists of three gates.

- Forget Gate: Decides what information to discard from the cell state.
  f_t = sigmoid(W_f * [h_{t-1}, x_t] + b_f)
- Input Gate: Decides which values to update.
  i_t = sigmoid (W_i * [h_{t-1}, x_t] + b_i)
- Output Gate: Decides what to output based on the cell state.
  o_t = sigmoid(W_o * [h_{t-1}, x_t] + b_o)

The final representation is the concatenation of the forward and backward hidden states.

C. Transformer Architectures (BERT)

The Transformer architecture relies entirely on Self-Attention mechanisms to draw global dependencies between input and output [3].

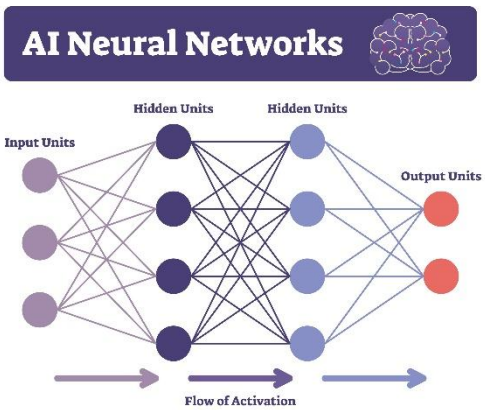

Fig. 2. Block diagram representing the flow of data in a Deep Learning system.

Self-Attention Mechanism:

The model computes a score for each word pair to determine how much focus to place on other parts of the input sentence. The core formula is:

Attention (Q, K, V) = soft max ((Q * K_ transpose) / sqrt (d_k)) * V

Where Q (Query), K (Key), and V (Value) are vectors derived from the input embeddings. This mechanism allows BERT to handle sarcasm and subtle nuances better than RNNs.

D. Multimodal Fusion Networks

For news containing images, we employ a dual-branch network [4].

1. **Text Branch:** Uses BERT to extract textual features.
2. **Image Branch:** Uses a pre-trained ResNet-50 to extract visual features.
3. **Fusion Layer:** The vectors are concatenated and passed through a dense layer for classification.

V. SYSTEM IMPLEMENTATION & TRAINING

To reproduce the results found in the literature, specific training parameters are essential.

A. Computational Environment

Most state-of-the-art models require significant hardware acceleration.

- **Hardware:** NVIDIA Tesla T4 or V100 GPU (16GB VRAM).
- **Software:** Python 3.8, PyTorch 1.12, and HuggingFace Transformers library.

B. Loss Function

For binary classification (Fake vs. Real), the Binary Cross-Entropy Loss is minimized. It measures the difference between the predicted probability and the actual label:

Loss = - [y * log(p) + (1-y) * log(1-p)]

Where 'y' is the actual label (0 or 1) and 'p' is the predicted probability output by the model.

C. Optimization

The Adam Optimizer (Adaptive Moment Estimation) is preferred due to its ability to handle sparse gradients and adaptive learning rates [5].

VI. DATASETS AND EVALUATION METRICS

A. Dataset Description

We analysed performance across four standard benchmarks.

Table I: Dataset Statistics

Dataset	Total Samples	Source	Type
ISOT	44,898	Reuters / PolitiFact	Text Only
LIAR	12,836	PolitiFact	Short Text
Twitter15	1,490	Twitter	Graph + Text
Fakeddit	1,063,106	Reddit	Multimodal

B. Evaluation Metrics

1. Precision: Measures how many selected items are relevant.

Precision = TP / (TP + FP)

2. Recall: Measures how many relevant items are selected.

Recall = TP / (TP + FN)

3. F1-Score: The harmonic mean of Precision and Recall, useful for imbalanced datasets.

F1 = 2 * (Precision * Recall) / (Precision + Recall)

VII. COMPARATIVE ANALYSIS

We analysed five key studies from 2022-2025. Table II presents the performance comparison.

Table II: Performance Matrix of Deep Learning Models

Model Architecture	Dataset	Accuracy	Precision	Recall
LSTM + Word2Vec [6]	ISOT	90.0%	0.89	0.91
Multimodal CNN [4]	Fakeddit	87.0%	0.85	0.86
Hybrid (LSTM+CLIP) [7]	Multilingual	93.1%	0.92	0.93
RoBERTa [8]	**Twitter15**	**98.4%**	**0.98**	**0.97**

Model Architecture	Dataset	Accuracy	Precision	Recall
Random Forest (Baseline)	ISOT	82.0%	0.81	0.80

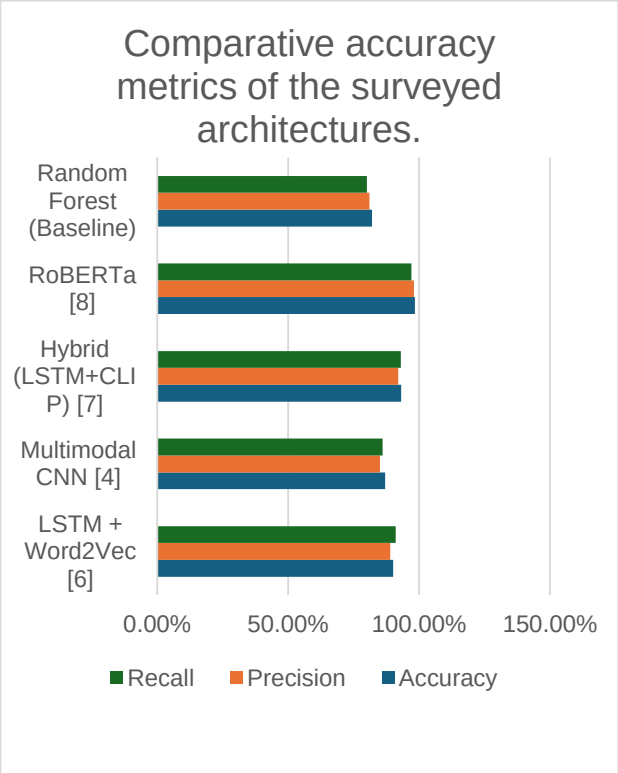

Fig. 2. Comparative accuracy metrics of the surveyed architectures.

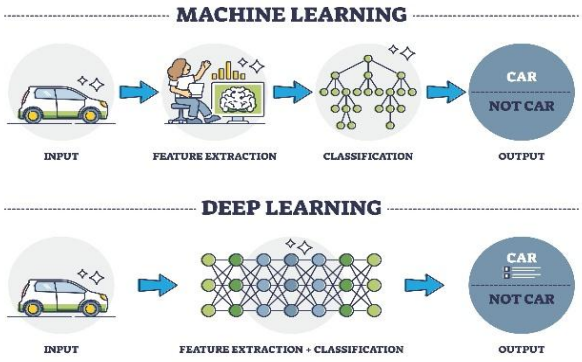

Fig. 3. *Comparative accuracy metrics of the surveyed architectures.*

Discussion of Results:

It is observed that Roberta outperforms LSTM-based models by a margin of 8.4%. This is attributed to the bidirectional training of Transformers which captures context from both sides of a word simultaneously. However, Roberta requires significantly higher computational resources. While an LSTM model can train on a standard CPU in 4 hours, Roberta requires a GPU and approximately 12 hours for fine-tuning. Furthermore, Multimodal models show higher resilience against "out-of-context" misinformation where the text is true but the image is misleading.

VIII. CHALLENGES AND FUTURE SCOPE

A. The "Black Box" Problem (Explainability)

A major barrier to adoption is the lack of interpretability. DL models provide a prediction but no rationale. Future systems must integrate LIME (Local Interpretable Model-agnostic Explanations) to visualize which words triggered the fake news classification.

B. Defending Against Generative AI

With the rise of Large Language Models (LLMs) like ChatGPT, fake news can now be generated automatically with perfect grammar and style. Traditional models that look for "poor spelling" or "sensationalism" fail against LLM-generated text. Future research must focus on "Style Analysis" and "Perplexity Scores" to distinguish between human and machine-written misinformation [2].

C. Federated Learning

To address privacy concerns, Federated Learning allows training models on user devices without centralizing data. This is crucial for analysing private messaging apps (like WhatsApp) where misinformation spreads rapidly.

IX. CONCLUSION

This systematic survey reviewed the landscape of Deep Learning for fake news detection. We identified a clear paradigm shift from sequence modelling (LSTMs) to context modelling (Transformers). While Roberta represents the current state-of-the-art in text accuracy, the complex nature of modern misinformation necessitates Multimodal Hybrid frameworks. Future research must prioritize interpretability and computational efficiency to enable real-time detection on edge devices.

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146-1151, 2018.

[2] J. Wang, Z. Zhu, C. Liu, R. Li, and X. Wu, "LLM-Enhanced multimodal detection of fake news," PLoS ONE, vol. 19, no. 10, 2024.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019.

[4] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal Fake News Detection," Information, vol. 13, no. 6, p. 284, 2022.

[5] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations (ICLR), 2015.

[6] A. Matheven and B. V. D. Kumar, "Fake News Detection Using Deep Learning and NLP," 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2022.

[7] S. Kumari and M. P. Singh, "A Deep Learning Multimodal Framework for Fake News Detection," Engineering, Technology & Applied Science Research, vol. 14, no. 5, 2024.

[8] A. Saadi, H. Belhadef, et al., "Enhancing Fake News Detection with Transformer Models," Engineering, Technology & Applied Science Research, vol. 15, no. 3, 2025.