Advancements in Deep Learning for Fake News Detection: A Comprehensive Review of Techniques, Datasets, and Emerging Trends

Sunil Ramchandra Gupta[1], Shashi Kant Gupta[2]
[1] LUC Application Number: LUC/CPGS/PDF/2025140803
Lincoln Global Postdoctoral Research (LGPR) Programme
Lincoln University College Malaysia
ORCID ID: 0000-0001-8714-0281;
[2] Adjunct Professor,
Lincoln University College, Malaysia
ORCID ID: 0000-0001-6587-5607;
Email ID: [1] pdf.sunilgupta@lincoln.edu.my, [2] raj2008enator@gmail.com

---

Abstract: Fake news on social media has emerged as a major challenge, significantly affecting politics, public health, social trust, and economic stability. The rapid expansion of digital platforms has enabled the widespread circulation of misinformation, exposing the limitations of traditional fake news detection methods such as rule-based systems and conventional machine learning techniques. These approaches struggle to manage the scale, semantic diversity, structural variation, and multimodal nature of contemporary fake news.

Deep Learning (DL) has proven to be a powerful alternative by enabling automated feature learning and effective processing of large volumes of unstructured data. Advanced DL models, particularly transformer-based architectures and self-attention mechanisms, have demonstrated superior performance in capturing contextual and social information. Techniques including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Graph Neural Networks (GNNs), pruned transformer models such as BERT, and hybrid frameworks have shown promising results. This paper reviews recent advancements in deep learning-based fake news detection, highlighting key models, datasets, evaluation metrics, ethical concerns, multimodal misinformation, and the growing role of Explainable AI (XAI) to guide future research toward accurate and transparent systems.

Keywords: Fake News Detection; Deep Learning; Multimodal Learning; Misinformation; Natural Language Processing (NLP); Explainable AI (XAI).

---

1. Introduction

Fake news constitutes information that has either been made up, or is presented in a misleading manner, then portrayed as factual news in order to dupe the audience [1]. The emergence of the Internet and its usage as a news medium has only aggravated the problem of fake news, which has become a problem to be dealt with globally. The emergence of social media like Twitter, Facebook and Reddit has made the dissemination of this misinformation rampant, impacting elections, social harmony, and even pandemics [3], [6]. The repercussions of fake news on the social order is overwhelming, ranging from the interference of elections (the U.S. presidential elections of 2016), to the misinformation of health, as in with the false

claims of the COVID-19 vaccines, thus engendering confusion and mistrust, as well as public endangerment. The classic methods of detection rely on feature engineering and the use of shallow classifiers which are unable to generalize and cope with a range of languages [4]. Deep Learning (DL) using tiered networks in Neural Nets, is capable of robust contextual modeling and automatic feature extraction [5], [6]. The BERT and GPT architectures, which are built on transformers and achieve the SOTA (State of the Art) performance, have shifted fake news detection from lexical analysis to multimodal semantic reasoning discourse [7].

1.1 Challenges in Detecting Fake News

Identifying the unique attributes of fake news can be difficult for the following reasons:

- Volume and velocity: Unlike print media, the unprecedented increase in the use of digital platforms allow the instantaneous dissemination of news. The inability to monitor and fact-check every news item in real time makes this rapidly expanding digital landscape extremely difficult to navigate.
- Intangible subtleties: The nuances of fake news pass under the radar of causal scrutiny. Conventional systems and methodologies have a difficult time identifying counterfeit news pieces and distinguishing them from authentic ones.
- Linguistics: The deceitful news eludes detection of its textual forms, for journalistic styles are made to mimic them, and hence, the simpler forms of lexical analysis are not fruitful.
- Imbalanced Datasets: The number of real and fake instances in datasets captured is appallingly low, hence, a true representation is not extracted.
- Multimodality: The fake news is often the combination of a misleading text along with a manipulated image or video.
- Temporal: News taken from static sources is nearly impossible to relate, as they lose relevance and news breaks in real time.
- Deep Learning: Understanding the logic behind classification in systems which employ a black-box philosophy, like in the case of deep neural networks is difficult, if not impossible.

1.2 The Role of Deep Learning in Addressing the Challenges

These challenges are addressed using deep learning through automated feature extraction and hierarchical representation of language and media [14]. CNNs identify local n-gram structures [15]. RNNs and LSTMs describe time series [16]. BERT, RoBERTa, and DeBERTa type transformers use self-attention for global context [17-19]. BERT, RoBERTa, and DeBERTa type transformers use self-attention for the global context [17-19]. GNNs broaden the reach of detection into the sociability sphere through user–post networks [20, 21]. Models of multimodal fusion enhance accuracy by combining text, images, and surrounding context [22]. Together, these techniques have boosted detection accuracy from an estimated 80% using classical ML techniques, to over 95% on benchmark datasets [23, 24].

1.3 Objectives of the Review Paper

The primary objectives of this paper are to:

- Describe the methods of deep learning relevant to the detection of fake news and misinformation.

- Analyze the various datasets in this field and their significance in evaluating model performance.
- Describe the obstacles and gaps in the current methods of fake news detection.
- Investigate new approaches and the prospects for development in multimodal fake news detection and explainable AI.

The first section presents an introduction to fake news detection and the role of deep learning as well as the challenges associated with it. The rest of the paper is organized as follows: Section 2 describes the literature. Section 3 analyzes Deep Learning Techniques. Section 4 explains the datasets. Section 5 discusses the obstacles i.e. challenges. Section 6 assesses the prospects. Section 7 provides the summary of the paper in form of conclusion.

2. Background and Related Work

2.1 Traditional Methods of Fake News Detection

Long before deep learning techniques became innovative, the detection of fake news was done by traditional approaches which used rule-based systems and other techniques that encompassed the broader scope of machine learning. Rule-based systems were the first approach used in fake news detection. This system objectively created rules in order to determine whether or not a text contained certain patterns or words that would suggest misinformation. For instance, particular descriptors like “clickbait” or “breaking news” were marked and labeled as possible fake news indicators [51]. Nonetheless, the rule-based systems failure borders on their over reliance on handcrafted rules which, in dynamic situations like the plethora of fake news streams, have limited use [52].

The applying of machine learning techniques to fake news detection moved the methods of detection to system learning models developed on labeled datasets. Most of the machine learning techniques used in fake news cases include the following:

- **Support Vector Machines (SVMs):** These are a dominant approach used in text classification and have increased popularity over the years as more techniques are developed to handle a larger number of features due to their ability to manage high dimensions. A Support Vector Machine, given a set of features, transforms the data to some space of higher dimensions and endeavors to identify some surface that differentiates the two classes (the news and the fake news) by optimally partitioning the entire space with respect to some user-defined criteria [54].
- **Naive Bayes Classifier:** Naive Bayes Classifier is a text classifier that uses probabilistic classifiers with Bayes’ theorem. They work well with large datasets, but have major limitations, such as the assumption of conditional independence between features. This assumption is highly fallible in reality. [53].
- **Random Forest:** This is a blending technique that builds sets of decision trees and merges their predictions achieves a more accurate and stable outcome. They work well with datasets that have many features and complex interrelations among the features. [29].

Most of the initial work that sought to develop automatic detection of fake news relied on ML algorithms that worked with complex matrices of text and social features. The algorithms primarily used included Naïve Bayes classifiers, SVMs, Logistic Regression, Decision trees, and Random Forests. Feature engineering was mostly on three dimensions: [1], [6].

1. **Lexical Cues:** The most primary features of text such as term frequency–inverse document frequency (TF-IDF), text n-grams, and the Flesch index. [3]
2. **Syntactic and Semantic Patterns:** Part of speech structures, named entities, and emotional sentiment. [4]
3. **User and Network Attributes:** The reputation of a user who posted, their frequency of reposts, and follower to following ratio. [5]

Though machine learning techniques enhanced the ability to scale detection of fake news, there were still various issues with these techniques, including the following:

- **Limited Feature Engineering:** The effectiveness of these models was dependent on engineered features, which were simply unable to capture the complex, contextual relationships woven into fake news.
- **Inequitable Data:** Fake news data sets are often skewed, with a predominance of real news articles versus fake news articles, which leads to skewed predictions [17].
- **Understanding Context:** Fake news detection relies largely on the ability to understand the context of narratives, a failing with most traditional ML models.

It is also worth mentioning that these techniques had a moderate accuracy of about 70-85%, still it was clear that they lacked scalability and ability to work with various domains. Furthermore, fake news writers quickly evolved how they wrote as the rules to static lexical became redundant [6]. Contrived features also failed to capture contextual subtleties such as sarcasm or implicit bias [7].

2.2 Limitations of Traditional Methods

Traditional techniques often face the following problems:

- **Sparsity of Features:** Constructed from high-dimensional TF-IDF vectors, these often lead to over-fitting [8].
- **Dependence on Domain:** Those trained on political news often perform poorly with health-related misinformation [9].
- **Inability to Process Multimedia:** Text only features disregard the increasing text and image features of misinformation [10].
- **Temporal Drift:** Developed features lack the ability to deal with changing subjects or new entities [11].

This is when the professionals started looking at raw data to construct latent features using representation learning and deep learning techniques [12].

2.3 Overview of Deep Learning Techniques

With deep learning, ML now has non-linear transformations added to the traditional processes of the field. It is the progress brought to the field by deep learning that seems to lift the bounds placed by other techniques. Unlike other models, these are able to learn hierarchical structures from raw data without the need for defined features. Tasks with big unstructured datasets like fake news detection, with accompanying text, images, and sometimes videos, are best suited for these models. Some of the most popular deep learning models that have been used to detect fake news are:

- **Convolutional Neural Networks (CNNs):** capture the local dependencies of words and the meaning of sentences [14], [15].
- **Recurrent Neural Networks (RNNs) / Long Short-Term Memory (LSTM):** learn the order of words and the evolution of context [16].
- **Attention and Transformer Models:** Use mechanisms of self-attention to encode dependencies through both directions like in models BERT, RoBERTa, XLNet and some models of GPT [18 – 19].
- **Graph Neural Networks (GNNs):** describes the social context reasoning by user to news interactions and by the propagation graphs of user to news interaction [20], [21].
- **Multimodal Fusion Networks:** examine the textual, visual, and social modalities and find cross-modal inconsistencies [22], [23].

Each paradigm has a specific role to play, as CNNs and RNNs capture granular linguistic details, while transformers deepen the contextual grasp. GNNs extract relational information, and multimodal networks identify inter-modal discrepancies. Together, they have raised the benchmark accuracy to over 95% [24]. The major benefits of deep learning techniques compared to classical approaches include:

- **Automated Feature Learning:** Deep learning models extract relevant features from raw data without the need for manual feature engineering.
- **Contextual Understanding:** Models understand the context of news articles more than other models; hence they are able to detect more subtle forms of fake news that may not be evident in standalone pieces of content.

2.4 Prior Surveys and Their Gaps

There have been several surveys focused on fake-news detection, each covering a different aspect of the problem.

- Zhou and Zafarani, 2020 [25] classified detection approaches into four categories: knowledge-based, style-based, propagation-based, and credibility-based.
- Shu and Liu, 2019 [26] approached the problem from the data-mining perspective and concentrated on feature extraction as well as model construction.
- Varshney and Vishwakarma, 2020 [27] described the detecting pipeline, covering from data collection to the final veracity classification, but did not include developments from the multimodal domain.
- Rohera et al., 2020 [28] focused on the dualities of supervised and semi-supervised approaches.
- Abdali et al. (2024) [8] continue to grapple with cross-lingual learning and explainability while providing the first comprehensive analysis on the challenges and advancements within the field of multi-modal misinformation detection.

While these advancements are worthwhile, there has been a surge of activity around transformers, foundation models, and federated learning that would warrant a new and more integrated review. Therefore, the next section concentrates on the state of the art in deep learning and fake news in the period from 2019 to 2025.

3. Deep Learning Techniques for Fake News Detection

Deep Learning (DL) is changing fake-news detection by allowing end-to-end learning of features across various dimensions: text, image, and social [1], [3]. Different from classical ML, these models learn the intricacies of language, syntax and meaning, and multimodal ties [4], [5]. This part of the paper clusters and distinguishes the high-performing designs.

3.1 Convolutional and Recurrent Networks

Convolutional Neural Networks (CNNs) are proficient in the extraction of local linguistic and stylistic features such as n-grams or part-of-speech combinations. For instance, Mridha et al. [6] constructed FNDNet and obtained 93.5 % accuracy on the LIAR and Kaggle datasets.

Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) frameworks also capture sequential dependencies and have surpassed traditional SVM baselines on larger datasets [7]. The hybrid model also outperforms other designs as the CNN and LSTM frameworks are merged [8].

3.2 Transformer-Based Architectures

Most modern detection studies focus on the transformer models, and particularly BERT and its innovations. In model building, BERT uses self-attention layers to construct long-range dependencies [9]. Fine-tuned transformer models reached 97 % accuracy on datasets such as Fakeddit and PolitiFact, as demonstrated by Nasser et al. [5] and Emil & Remus [4].

Recent developments, such as RoBERTa, XLNet, DeBERTa and GPT-style large language models, provide even greater improvements to contextual embeddings [12]. Their core capability involves transfer learning, which permits changes across various domains with minor model re-adjustments [13].

3.3 Graph Neural Networks (GNNs)

GNNs capture and model the social media interaction- diffusion network where posts are represented as nodes and links as edges. Hu et al. [14] and Shen et al. [15] utilized GNNs on Twitter and Weibo to demonstrate that structural geometric propagation information significantly improves early-boost stage detection with over 90% accuracy. Further, this set of models enables temporal analysis of rumors as they diffuse [16].

3.4 Multimodal Fusion Networks

Recent literature indicates that fake news is often accompanied by deceptive images along with text that is misleading [8]. Multimodal DL architectures unify text CNN/LSTM encoders and visual ResNet or VGG sub-networks [18]. Abdali et al. [8] used cross-modal attention for the detection of image-text inconsistencies, and Jin et al. [20] utilized transformer- based contextual fusion layers to align disparate sections for cohesive context.

3.5 Federated and Privacy-Preserving Models

Federated Deep learning offers tool for protective decentralized model training across multiple clients, thus cushioning the impacts of data-sharing restrictions. Chandua et al. [21] employed federated CNN on distributed COVID-19 tweets and achieved 92% accuracy, all while maintaining data privacy. The remaining challenges of unsatisfactory collusion and domain heterogeneity are still worth exploring [22].

3.6 Comparison of Deep Learning Models and Techniques for Fake News Detection

Table 1. Compares Deep Learning Models and Techniques for Fake News Detection related work by the researchers

Author	**Dataset(s)**	**DL Technique / Model**	**Strengths / Contributions**	**Limitations**
Mridha et al. (2021) [6]	LIAR, Kaggle	CNN (FNDNet)	Extracts deep text features; 93.5 % accuracy	Computationally expensive
Kaliyar et al. (2021) [2]	Kaggle	CNN–LSTM Hybrid	Captures sequential + spatial patterns	Interpretability issues
Wang et al. (2017) [10]	FakeNewsNet	BiLSTM	Models' context across sentences; 92.6 % accuracy	Requires large datasets
Hiramath & Deshpande (2022) [23]	ISOT	DNN	Lightweight model with fast inference	High variance on small data
Nasser et al. (2025) [5]	Fakeddit, Twitter	BERT, RoBERTa	Achieved >97 % accuracy; context-rich	Resource intensive
Emil & Remus (2025) [4]	PolitiFact, GossipCop	GPT-4-based LLM	Handles multilingual input	Prone to training bias
Hu et al. (2024) [14]	PolitiFact, Weibo	GNN	Exploits social propagation graphs	Complex training
Shen et al. (2023) [15]	Twitter	GNN (GAMED)	Multi-expert propagation modeling	Requires detailed graph data
Abdali et al. (2024) [8]	Weibo, Twitter	Multimodal Fusion	Detects cross-modal inconsistencies	Limited temporal modeling
Jin et al. (2023) [20]	Weibo	Transformer Fusion	Text–visual alignment for multimodal data	Dataset-specific performance
Qi et al. (2023) [19]	Fakeddit	CNN + Attention	Combines semantic + sentiment signals	Sensitive to noise
Chandua et al. (2025) [21]	FL-COVID19	Federated CNN	Privacy-preserving architecture	Client heterogeneity
Tan & Bakir (2024) [11]	Twitter	Transformer + SMOTE	Balances data; 99.9 % accuracy	Overfitting risk
Alshuwaier & Alsulaiman. (2025) [3]	LIAR, ISOT	BiLSTM	Robust multilingual generalization	Weak interpretability
Goldani et al. (2022) [16]	FakeNewsNet	Capsule Network	Captures hierarchical dependencies	Complex optimization

3.7 Performance Comparison

In almost all benchmark datasets when examining the accuracy and F1-score results for BERT and RoBERTa, the neural transformer models have outperformed the CNN/LSTM models by 5-10% [9], [10]. GNNs have excelled in the early stage of detection because of prior use of propagation cues [14], while the multimodal fusion of image and associated text enhances the robustness of the system [8]. Still, computational cost and interpretability are the enduring trade-offs. [23], [24]

4. Datasets and Benchmarks

4.1 Commonly Used Datasets

In the domain of fake news detection, the construction and evaluation of models is heavily dependent on reliable clones of fake news datasets for training and assessment. In the following are some of the most widely adopted datasets in the field:

- LIAR Dataset: This database comprises 12,800 concise assertions categorized as true or false. It is one of the principal databases utilized for various tasks, including fake news and sentiment analysis [55].
- FakeNewsNet: This is an all-encompassing dataset that comprises news content and its associated social media metadata, allowing for the analysis of fake news in the social media environment [1].
- ISOT Dataset: It contains over 23,000 news articles from diverse sources that are classified as real or fake articles. This dataset is valuable for training models in a more streamlined or controlled training environment [9].

The benchmark datasets are essential in assessing the effectiveness of systems that detect fake news. They vary in terms of language, domain, modality, and annotation. Since 2017, the LIAR, FakeNewsNet, and Fakeddit corpora have provided access to datasets that enable reproducible research experiments [1], [8], [10]. Newer datasets offer cross-lingual or multimodal components that mimic real-world disinformation and misinformation [8], [19].

Early corpora were restricted to the analysis of text. The LIAR dataset [10] consists of 12.8 k short political statements that have been fact-checked and categorized on a 6-point truth scale and fact-checked by PolitiFact. ISOT Fake News [9] contains 45 k English news articles with a balanced real and fake class, and the Kaggle Fake and Real News [6] dataset consists of 40 k articles obtained from various publishers. These datasets are still widely used for baseline evaluation because of their simplicity and balanced class representation, but they are devoid of multimodal or social context.

4.2 Social and Contextual Datasets

To model patterns of social engagement, FakeNewsNet [8] integrates article text with user profile metadata, retweets, and records of the article's temporal diffusion. Sub-collections of PolitiFact and GossipCop facilitate the credibility assessment of fact-checking domains. The Weibo and Twitter datasets [8], [14] are useful for propagation-based models that utilize repost graphs and stance-shifting interactions. However, noise in the annotated data and the rapid evolution of the platforms restrict their generalization for long-term use [14].

4.3 Multimodal Datasets

The progress regarding the pair textual and image signals. Fakeddit [10] offers 200 k Reddit posts classified under six veracity categories, including memes and news headlines. Weibo-MM [19] and MultiFakeMM [10] append image-text pairs for cross-modal attention. These resources support transformer-fusion networks [20] and studies on visual-semantic consistency [8].

4.4 Multilingual and Federated Datasets

The language bias problem impacts most datasets, which are still predominantly centered around English. LIAR+ and ISOT-Extended [13] created multilingual statements, while AraNews and FakeNewsArabic [34], [41] focus on Arabic. FL-COVID19 [21] offers 1.2 M tweets spread among federated clients for privacy-preserving studies.

4.5 Comparative Overview of Datasets (From Year 2019 – 2025)

Table 2. Compares Deep Learning Models and Techniques for Fake News Detection related work by the researchers

Author	Dataset Name	Size / Language	Features and Modalities	Limitations
Wang et al. (2017) [10]	LIAR	12.8 k / English	Short political claims + 6-label truth scale	Small; limited context
Ahmed et al. (2020) [54]	ISOT Fake News	45 k / English	Full articles (text)	Single domain
Mridha et al. (2021) [6]	Kaggle Fake and Real News	40 k / English	Headline + content	Sparse metadata
Abdali et al. (2024) [8]	FakeNewsNet	23 k / English	Text + social context + propagation	Limited languages
Hu et al. (2024) [14]	Weibo	16 k / Chinese	Post text + repost graph	Platform-specific
Jin (2024) [20]	Weibo-MM	120 k / Chinese	Text + image pairs	Restricted access
Wang et al. (2017) [10]	Fakeddit	200 k / English	Multimodal Reddit posts	Label inconsistency
Alshuwaier & Alsulaiman. (2025) [3]	LIAR+ / ISOT-Extended	50 k / Multilingual	Parallel multilingual statements	Unbalanced topics
Qi et al. (2023) [19]	MultiFakeMM	963 / Multilingual	High-quality multimodal samples	Small size
Chandua et al. (2025) [21]	FL-COVID19	1.2 M / Multilingual	Federated tweets	Noisy text

Rahman et al. (2023) [38]	FakeNewsArabic	60 k / Arabic	Text + metadata	Limited domain
S. Taskin, (2022) [34]	AraNews	25 k / Arabic	Generated + real articles	Synthetic bias
Wang et al. (2017) [10]	Fakeddit (Extended)	1.1 M / English	Text + image + engagement	High redundancy
Hu et al. (2024) [14]	PolitiFact + GossipCop	60 k / English	Fact-check metadata	Domain bias
Qi et al. (2023) [19]	Weibo PropNet	100 k / Chinese	Text + propagation graphs	Temporal drift

4.6 Dataset Challenges

While there has been progress in some areas, there are still issues with the dataset, such as:

- Class Imbalance: In open collections, real articles in the vast majority are 3 or more times as many as fake articles [9].
- Annotation Quality: Labels are often assigned based on insufficient fact-checking [8].
- Cross-Lingual Coverage: There are very few datasets with aligned multilingual content [13], [34].
- Rapid topic evolution: Static corpora do not capture new events or crises after a certain point in time [14].

These challenges have spurred the ongoing need for larger, multimodal, multilingual, and constantly updated datasets [8], [19], [41].

5. Challenges and Limitations

In the realm of fake-news detection and its effective use of deep learning, there is considerable progress, but there remain many technical and ethical issues. Such issues arise from the data being used, the model itself, and any potential risks associated with the model's use. As such, these issues need to be solved in order to ensure the construction of reliable and seamlessly applicable models.

5.1 Data Imbalance and Domain Dependency

Class imbalance, where the number of real news samples overwhelmingly supersedes fake news samples is able to be identified even in the earliest research [6],[9]. A model that is trained in such data will almost indiscriminately associate news as real, predicting the real class with extreme certainty, and display low recall metric on fake class [10]. Approaches, such as Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive learning [12], while useful in such challenges, often results in highly distorted, and sometimes meaningless, distributions of certain features.

Domain dependence, akin to the problem of class imbalance, and more generally imbalance data sets, is exemplified by the LIAR and FakeNewsNet data sets which are both rather shallow in their topic coverage. One dominated by the politics of a certain country, and the other which is highly concentrated on the

entertainment industry. Models that are taught a certain topic, such as politics, typically lack the necessary skills to perform in other domains, such as health [9],[13]. The problem of cross domain generalization is known, with the necessary strategies being domain adaptation and continual learning [8].

5.2 Interpretability and Explainability

There is a lot of ignorance when it comes down to deep learning models. Even though they perform endorsements incredibly, their models are incapable of providing even a modest amount of transparency. From a professional and economical standpoint, it is of utmost importance to understand the reasons behind these models' decisions, especially in sensitive contexts [1] [27]. Even though there are many approaches to Explainable AI (XAI) such as attention and feature attribution and the use of concept activation vectors, these methods have been integrated to fake-news detection systems [49]. Unfortunately, the explanations provided are vague and misleading. They tend to deal with linguistic elements in the text and neglect the format of the message, which is of utmost importance [11]. It is a basic research problem to intercept commonsense understandable explanations of reasoning behind decisions made by models constructed with Transformer based deep learning systems [8] [49].

5.3 Multimodal Fusion and Data Alignment

The combination of various sources of signals such as text, images, videos and even propagation signals play a huge role in fake detection systems [8] [19]. It is vital for system performance, but it comes at the cost of added complexity concerning the alignment systems used. The times when the image and text pairs we used were roughly the same bumper images in fake advertisements never got the appropriate exclamations so. So, these text image disassociates will never advance elimination the text for inaccuracy [18].

The recent models of multimodal transformers are emerging which try to associate the images with their correct captions by using cross attention between disparate types of data [19][47] so. These processes consume a lot of resources and work best without distraction between elements in the images. In addition, the volume of paired text and images and videos is still much smaller and more homogenous than datasets containing only text which restrict inadequate learning capabilities such as [10] [19].

5.4 Temporal and Evolving Information

The nature of fake news is such that it develops along dynamically trending axes, with specific topics garnering attention for a few hours and then being rapidly supplanted with novel ideas. The vast majority of models under consideration still work within a static framework, failing to incorporate the dynamic nature of the language and the subsequent change propagation paths [14]. The use of Temporal Graph Neural Networks (TGNNs) and incremental learning frameworks begin to address the challenge [16], [20], but the problem of continuous retraining is expensive.

The problem of concept drift, in which the associations between the features and the labels change over a designated timeline, can greatly deteriorate the performance of a model [8]. To deploy such model in the real world means that it must have some form of adaptive online updating capabilities [15].

5.5 Ethical, Privacy, and Societal Concerns

The advent of deep-learning models brings with it profound ethical issues.

- **Bias Amplification:** Through training, models that possess data that is biased on a demographic and ideological level, may perpetuate such bias [27, 49].
- **Privacy Concerns:** The acquisition of user-level social data a and compilation of it is a threat to sensitive data [21]. Though federated learning protects a variety of datasets, it is still vulnerable to data leakage [22].
- **Dual-Use Risk:** Opponents of a system may use fake-news detectors to refine techniques of disinformation by reverse engineering the detection parameters [8].

Moreover, the governance of automated moderation remains problematic. The use of AI systems to automatically label misinformation raises issues of misinformation, the blurring of lines of responsibility, and the deterioration of civil liberties [7]. Consequently, there is a need to deploy detection models within a flexible framework that allows for the human feedback cycle to provide ethical oversight [49].

5.6 Computational and Environmental Costs

The training and use of modern transformer models is a complex and costly undertaking. The GPT-4 and RoBERTa-large models in particular are especially resource-hungry for both training and inference [11, 12]. This creates a barrier in accessibility for poorer nations and small institutions, while also adding to the carbon footprint as a result of the training of models on larger datasets. More efficient alternatives like DistilBERT, TinyBERT, and MobileBERT [8] still suffer from resource and accuracy issues [49].

To conclude, even though deep-learned systems have purportedly distinguished between genuine news and fake with never before attained virtues, such advances have considerable transparency, scalability, and ethical cost. The following section analyzes potential upcoming research areas and directions that focus on tackling these discrepancies.

6. Future Directions

Even though there is great achievement in the use of deep learning technologies for the task of detecting fake news. The progress in the area is still very fast. The recent state of the art in explainable AI (XAI), multimodal fusion, federated learning, and large language models (LLMs) is setting new research priorities in the area aimed at improving adaptability and scalability. This section outlines new research directions.

6.1 Explainable and Interpretable AI (XAI)

As fake news detection impacts society and policy, its use and the models adopted become vital [27], [49]. Explainable AI (XAI) is concerned with closing the gap between predictive performance and understanding a system.

Contemporary attempts merge attention-based methods, SHAP, and LIME within the framework of Transformer models [49]. Although these methods shed light on the contributions of tokens, they hardly provide causal explanations for deception. Coming up with more effective XAI has to:

- Exhibit reasoning chains whereby users understand what models use to flag certain content.
- Provide advanced human-bot collaboration where machine precision is coupled with expert oversight.

- Advance methods to measure and report confidence on predictions [8].

The level of transparency will improve while the issues of black-box overreliance and trust in automated moderation and system use will be reduced.

6.2 Multimodal and Cross-Lingual Learning

As shown in the work of [8], [19], [47], misinformation in the real world tends to arise as a multimodal phenomenon that incorporates text, images, sounds, and video. Therefore, the use of single-modality models will not be sufficient. Systems of the future should use cross-modal embeddings that represent the integrated visual, textual, and auditory semantic elements. Document and vision transformers, such as CLIP and ViLT, show the increasing contextual enhancement that aligned embeddings deliver [47].

Cross-lingual frameworks need to be more robust. There is a huge gap ignoring billions of people who do not speak English or the English-centric datasets that continue to persist [13], [34], [41]. Cross-lingual transformer models such as XLM-R and mBERT assist in the gap by providing shared latent semantic space across multiple languages, hence, lowering annotation expenses [8], [41]. Other more contemporary examples involve code-switching detection, which is dominant in South Asia and Africa, and where multilingual posts that combine languages are frequent. These low-resource conditions are ideal candidates for few-shot and meta-learning frameworks [35].

6.3 Federated, Privacy-Preserving, and Decentralized Learning

Research on fake news involves taking the system's infrastructure into account. Data privacy is still a major ethical and societal problem, and the publicity-driven approach to misinformation is oversimplified. By focusing on the privacy issues, the authors posit the risk of excluding useful and essential elements, for instance, the psychological and emotional complexities of the actors involved [21]. "The first step…addressed by improving local models to carry out more complex operations for part of the training and sending only the rough gradients to be aggregated by the central server, a technique called deep federated learning" [21]. Chandua et al. [22] argue violable CNN architectures and propose intervals for COVID-19 misinformation. Performance is vetted against privacy concerns. All over the world research is being conducted towards developing a framework, which combines elements of:

- Preventing leakage of confidential information through aggregate secret sharing.
- Managing client heterogeneity, or unequal non-independent and identically distributed data set problem.
- Modular Blockchain architectures for easy and reliable audits.

In addition, elements of FL can be enhanced through the engagement of classic approaches. In this instance it is the differential privacy combined with homomorphic encryption technique to safeguard sensitive data of the users during the training phase [15].

6.4 Adaptive and Temporal Modeling

With respect to fake-news topics, static detection models deteriorate over time. The evolution of fake-news themes novels the use of Temporal models, such as Temporal Graph Neural Networks (TGNNs), and streaming transformers for static model learning [14], [20]. Concept-drift detection and online adaptation should be the emphasis of future research. New models, self-updating based on the emergence of new

misinformation patterns, can be reliably used across news cycles [8], [23]. It is possible that combining these strategies with reinforcement learning could enable the system to enhance self-governance during stability periods and better respond to new, emerging misinformation campaigns [24].

6.5 Large Language Models (LLMs) and Generative Detection

As demonstrated with the recently released PaLM, LLaMA, and GPT-4 foundation models, LLMs have revolutionized the field of NLP [11]. These models are capable of zero and few-shot classifications, allowing for fake news detection in entirely new domains [12]. Of note, LLMs are capable of generating disinformation (thus creating a dual-use dilemma) [49]. Future work should focus on enhancing alignment and controllability. More LLMs need to be able to detect lies rather than propagate them [11]. Hybrid generative detection frameworks, in tandem with discriminative classifiers, have the potential to merge contextual understanding with precision [8]. Additionally, it will be necessary to construct evaluation benchmarks specifically for LLM-based detectors for their responsible deployment, including assessments on hallucination resistance, facts grounding, and bias mitigation.

6.6 Multimodal Misinformation Intervention

In addition to detection, future systems should be able to assist in active intervention and mitigation. Automated explanations with integrated fact-checking retrieval modules that cross-verify with databases of proven claims (like PolitiFact or Snopes) could be beneficial [8].

Enhanced, cross-discipline multimodal reasoning (text, images, and user actions) will help scale systems to assess credibility instantaneously [8], [47]. Future directions envision integrated systems of AI, social networks, and human fact-checkers as the next step toward more efficient, sustainable frameworks to counter misinformation.

6.7 Sustainable and Resource-Efficient AI

With the growing costs associated with deep learning models, the academic research domain places more value on the development of model and system designs with lower energy footprints, often referred to as ‘energy-efficient’ [11, 12]. Methods such as model pruning, quantization, and knowledge distillation can often greatly lower inference costs while maintaining a satisfactory level of accuracy [8]. Cloud or edge-deployed fake-news detectors will soon become essential for the scalable moderation of large-scale systems deployed globally.

7. Conclusion

The evolution of fake-news detection and classification systems has progressed from simple, hand-crafted rule-based approaches to sophisticated, deep-learning frameworks harmonizing text, images, and social media. Detection systems employing hand-crafted feature set approaches relied heavily on context, lacked scalability, and were poorly understood [3], [6]. Transformer-based and multimodal deep-learning techniques provide unrivaled performance, flexibility, and dominance in cross-lingual and multi-domain tasks. This review has constructed the deep-learning research landscape on model, dataset, and evaluation metric consolidation frameworks for fake-news detection. Text-based CNN and RNN architectures provide strong baseline performance, while transformer approaches such as BERT, RoBERTa, and the various GPTs handily dominate the field with their unrivaled contextual reasoning.

Domain-specific Graph Neural Networks (GNNs) and multimodal fusion frameworks augment the detection systems by providing rich context through user activity and cross-modal reasoning.

Pervasive issues such as ethical slant and insufficient resources, alongside dataset imbalance and insufficient algorithm interpretability, still require considerable work. Highly relevant future directions such as cross-lingual, federated, and foundation model approaches alongside explainable AI offer the deep learning landscape greater clarity, flexibility, and trust. To sum up, deep learning has made tremendous improvements in the ability to detect misinformation, but responsible and transparent detection remains to be seen. The next stage couples deep learning with rigorous interpretability and ethical frameworks to produce responsible and socially valuable deep learning systems for detecting and classifying misinformation. Integrating diverse contextual sources, multilingual capacities, adjustable interpretive frameworks, and ethical alignment will extend the reach of fake news detection systems.

References

1. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information," *Information Systems*, vol. 101, pp. 101561, 2020. https://doi.org/10.1016/j.is.2020.101561.
2. R. K. Kaliyar, A. Goswami, and P. Narang, "Fake news detection using a CNN–LSTM hybrid model on Kaggle dataset," *Procedia Computer Science*, vol. 189, pp. 417–425, 2021. https://doi.org/10.1016/j.procs.2021.05.073.
3. F. A. Alshuwaier and F. A. Alsulaiman, "Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review," *Computers*, vol. 14, no. 394, 2025. https://doi.org/10.3390/computers14090394.
4. R. S. Emil and B. Remus, "A Review of Automatic Fake News Detection: From Traditional Methods to Large Language Models," *Future Internet*, vol. 17, no. 435, 2025. https://doi.org/10.3390/fi17100435.
5. M. Nasser, M. Hassan, S. Alam, and F. Akhtar, "A systematic review of multimodal fake news detection on social media using deep learning," *Results in Engineering*, vol. 26, 2025. https://doi.org/10.1016/j.rineng.2025.104752.
6. M. F. Mridha, S. Sarkar, and M. M. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156431–156455, 2021. https://doi.org/10.1109/ACCESS.2021.3129329.
7. B. Hu, Z. Mao, and Y. Zhang, "An Overview of Fake News Detection: From a New Perspective," *Fundamental Research*, vol. 5, pp. 332–346, 2025. https://doi.org/10.1016/j.fmre.2024.01.017.
8. S. Abdali, S. Shaham, and B. Krishnamachari, "Multi-modal misinformation detection: Approaches, challenges, and opportunities," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–38, 2024, arXiv:2203.13883v7.
9. R. K. Kaliyar, A. Goswami, and P. Narang, "DeepFake: A Deep Learning Based Multi-Model Ensemble Approach for Fake News Detection," *Applied Soft Computing*, vol. 98, 2021. https://doi.org/10.1016/j.asoc.2020.106879.

10. W. Y. Wang, “Liar, liar pants on fire: A new benchmark dataset for fake news detection,” in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, vol. 2, Vancouver, BC, Canada, pp. 422–426, 2017. https://doi.org/10.18653/v1/P17-2067.
11. R. Tan and M. Bakir, “Hybrid Transformer Architecture for Fake News Detection on Twitter,” *IEEE Access*, vol. 12, pp. 91215–91227, 2024.
12. M. H. Goldani, “Detecting Fake News with Capsule Neural Networks,” *Expert Systems with Applications*, vol. 184, pp. 115462, 2021.
13. S. Verma and R. Gupta, “WELFake: Word Embedding-Based Linguistic Features for Fake News Detection,” *IEEE Access*, vol. 9, pp. 97863–97875, 2021.
14. B. Hu, Z. Mao, and Y. Zhang, “Propagation-Based Fake News Detection,” *Fundamental Research*, vol. 5, pp. 340–342, 2025.
15. Z. Shen, “GAMED: A Graph-Based Multi-Expert Discriminative Framework for Fake News Detection,” *Pattern Recognition*, vol. 144, 2023.
16. M. H. Goldani, S. Momtazi, and N. Safaei, “Capsule networks for fake news detection,” *Neurocomputing*, vol. 500, pp. 10–27, 2022. https://doi.org/10.1016/j.neucom.2022.05.023.
17. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, vol. 2, pp. 427–431, Valencia, Spain, 2017.
18. B. Al-Tarawneh, “Enhancing Fake News Detection Using TF-IDF and CNN Models,” *Pattern Recognition Letters*, vol. 165, 2023.
19. P. Qi, “Text-Image Correlation Features for Multimodal Fake News Detection,” *Expert Systems with Applications*, vol. 208, 2023.
20. Z. Jin, “Multimodal Fusion with VGG and BERT for Fake News Detection,” *IEEE Transactions on Multimedia*, vol. 26, 2024.
21. S. V. Chandua, U. S. Varria, and V. Raj, “Federated Learning in Detecting Fake News: A Survey,” *Procedia Computer Science*, vol. 260, pp. 457–467, 2025. https://doi.org/10.1016/j.procs.2025.03.223.
22. R. Tan and M. Bakir, “Balanced Transformer for Fake News Detection,” *IEEE Access*, vol. 12, 2024.
23. M. S. Hiramath and A. Deshpande, “Fake News Detection Using Deep Learning Techniques,” *Procedia Computer Science*, vol. 217, 2022.
24. P. Bahad, “Fake News Detection Using Bidirectional LSTM-RNN Models,” *Social Network Analysis and Mining*, vol. 12, 2022.
25. X. Zhou and R. Zafarani, “Network-Based Fake News Detection,” *ACM Computing Surveys*, vol. 53, 2020.
26. K. Shu and H. Liu, *Detecting Fake News on Social Media*, Morgan & Claypool, 2019.
27. D. Varshney and D. Vishwakarma, “A Survey on Fake News Detection,” *IEEE Access*, vol. 8, pp. 13292–13312, 2020.
28. S. Rohera, “Machine Learning Approaches to Fake News Detection,” *Procedia Computer Science*, vol. 171, pp. 409–416, 2020.

29. S. A. Williams, P. R. Burnap, and O. Rana, “Detecting online fake news using sentiment analysis and machine learning,” *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, San Diego, CA, USA, pp. 1–8, 2017. https://doi.org/10.1109/IRI.2017.8078893.
30. B. Pardamean and E. Pardede, “Fake News Detection Using BERT Embeddings,” *Procedia Computer Science*, vol. 176, 2020.
31. B. L. Gereme, “Fake News Detection for Low-Resource Languages Using Bi-LSTM,” *IEEE Access*, vol. 11, pp. 32412–32426, 2023.
32. K. Ivancova, “Fake News Detection for Slovak-Language Datasets,” *Expert Systems with Applications*, vol. 188, 2022.
33. E. M. Nagoudi, “AraNews: Arabic Fake News Generation and Detection Dataset,” *J. King Saud Univ. Comp. & Info. Sci.*, vol. 34, 2022.
34. S. Taskin, “RNN-Based Multilingual Fake News Classifier,” *Neural Computing and Applications*, vol. 34, 2022.
35. D. Sahoo, “Fake News Detection via Sentiment and Style Features,” *Applied Soft Computing*, vol. 120, 2022.
36. G. Singh, “Fake News Detection Using Ensemble Learning and NLP,” *Computers & Electrical Engineering*, vol. 104, 2023.
37. M. Memon and S. Ahmed, “Multimodal Fusion of Visual and Textual Data for Fake News Detection,” *IEEE Access*, vol. 10, pp. 93280–93294, 2022.
38. A. Rahman, M. Alshamrani, and M. Al-Turaiki, “FakeNewsArabic: A benchmark dataset for Arabic fake news detection,” *IEEE Access*, vol. 11, pp. 12045–12058, 2023, https://doi.org/10.1109/ACCESS.2023.3245671.
39. Khan, “Hybrid CNN-BiLSTM Framework for Fake News Classification,” *Computers in Human Behavior*, vol. 142, 2024.
40. H. Patel, “A Graph-Based Transformer Model for Fake News Detection,” *Knowledge-Based Systems*, vol. 279, 2024.
41. S. Ghosh and S. Chowdhury, “Cross-Lingual Transfer Learning for Fake News Detection,” *Information Fusion*, vol. 98, 2024.
42. P. Meel and D. K. Vishwakarma, “Fake News, Rumor, and Information Pollution: A Comprehensive Review,” *IEEE Trans. Comput. Social Systems*, vol. 8, pp. 178–195, 2021.
43. S. Choudhary, “Hybrid Deep Models for Misinformation Detection,” *Pattern Recognition Letters*, vol. 169, 2023.
44. L. Bahri, “Cross-Platform Multimodal Fake News Detection Using Deep Transformers,” *Information Fusion*, vol. 106, 2024.
45. Narang, “COVID-19 Fake News Detection Using Transfer Learning,” *J. Information Security and Applications*, vol. 68, 2023.
46. R. Kumar, “Explainable AI in Fake News Detection: A Survey,” *Artificial Intelligence Review*, vol. 58, pp. 131–160, 2024.
47. X. Xu, “Multilingual and Multimodal Fake News Detection: Trends and Future Directions,” *IEEE Access*, vol. 13, pp. 22101–22123, 2025.

48. M. Potthast, “A Stylometric Approach to Hyperpartisan News Detection,” *Information Processing & Management*, vol. 57, 2020
49. F. Al-Ahmad, “Explainable Deep Models for Fake News Detection,” *Expert Systems with Applications*, vol. 213, 2024.
50. Q. Zhu, “Stylometric and Linguistic Analysis of Fake News Writing Styles,” *IEEE Access*, vol. 9, 2021.
51. V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, pp. 3391–3401, 2018.
52. C. Buntain and J. Golbeck, “Automatically identifying fake news in popular Twitter threads,” *Proc. IEEE Int. Conf. Smart Cloud (SmartCloud)*, New York, NY, USA, pp. 208–215, 2017. https://doi.org/10.1109/SmartCloud.2017.66.
53. B. Fröhlich, F. Oberländer, M. Bick, and D. E. O’Sullivan, “Fake news detection on Twitter using stance classification: A case study on the German federal election 2017,” *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Singapore, pp. 188–193, 2018. https://doi.org/10.1109/ICDMW.2018.00034.
54. H. Ahmed, I. Traore, and S. Saad, “Detecting fake news using machine learning: A systematic literature review,” *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 11, no. 5, pp. 1–10, 2020. https://doi.org/10.14569/IJACSA.2020.0110501.
55. S. A. Williams, P. R. Burnap, and O. F. Rana, “Using the LIAR dataset for detecting fake news in political statements with machine learning,” *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, San Diego, CA, USA, pp. 1–8, 2017. https://doi.org/10.1109/IRI.2017.8078893.