Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Evolving Techniques in Fake News Detection: From Human Expertise to Large Language Models

Stefan Emil Repede*[1], Remus Brad[2]
[1,2]*Field of Computer Engineering and Information Technology, Lucian Blaga University, Romania*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

The proliferation of fake news poses significant challenges in today's digital age. This paper provides a comprehensive review of the most recent advancements in fake news detection methodologies, tracing their evolution from the earliest manual approaches to the latest state-of-the-art models, including large language models. The study identifies four key perspectives: knowledge based, style based, propagation based, and source based. Research on detection methods is classified into manual approaches and automatic approaches utilizing data science techniques like traditional machine learning, deep learning, and large language models. The review highlights the dual role of LLMs in generating and detecting fake news and discusses limitations of current methods, as the lack of datasets, emphasizing the need for multimodal analysis, interdisciplinary collaboration, and improved model transparency.

Keywords: Deep Learning, Disinformation, Fake News, Large Language Models, LLM, Machine Learning.

INTRODUCTION

In the digital era, the immense flow of information from various media channels subject's individuals to a continuous influx of data. The lack of robust verification mechanisms, coupled with the accessibility of information, contributes to the rapid spread of unverified and often false content, collectively known as "fake news" [1]. The ease of sharing and commenting on such content—often without verification—amplifying its reach, impacting public opinion and behavior at both local and global levels [2]. The consequences of fake news are particularly severe when it influences critical areas such as public health or political events. For example, during the 2016 U.S. Presidential election, it was estimated that over half of voters were exposed to misleading content, highlighting how fake news can sway public decision making [3]. Beyond politics, misinformation has undermined public health campaigns, such as those advocating for vaccination, underscoring an urgent need for reliable detection and mitigation strategies [4].

Fake news spans various forms, including misinformation, disinformation, satire, conspiracy theories, and even humorous fabrications. These diverse forms complicate the task of distinguishing truth from deception. Researchers commonly define fake news as intentionally and verifiably false information designed to mislead [5]. Examples include deep fakes, which use advanced media manipulation to portray public figures in fabricated scenarios, adding to the challenge of distinguishing reality from deception [6].

Social media platforms represent important vectors for the rapid dissemination of fake news. These platforms often foster a high level of trust among users, who may unintentionally spread false information due to lack of awareness, confirmation biases, or an inclination to trust social media sources over traditional outlets [7]. Studies reveal that a small number of false posts can quickly generate widespread traction, such as during natural disasters when manipulated images or sensational claims gain thousands of shares and reposts [8, 9].

Some research groups show that the spread of fake news involves two primary user groups: malicious actors and naive users. Malicious actors, including automated bots and pseudonymous accounts (sock puppets), intentionally craft and propagate misleading content to shape public perception [10]. Naive users, on the other hand, can share this information without verification, unknowingly amplifying its reach. Therefore, the current detection methods must account for these varied patterns of dissemination and motivations [11].

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Despite significant research, existing detection methods face challenges due to the volume, diversity, and complex nature of fake news [12]. This paper reviews current methodologies that aim to address these limitations, categorizing detection approaches as either manual or automatic. Manual detection relies on expert and crowd- sourced fact checking, though scalability remains a barrier. Automated methods leverage machine learning, deep learning, and emerging Large Language Models (LLMs) to analyze fake news on a larger scale. The rise of LLMs, in particular, offers new opportunities for detecting nuanced patterns in text and multimedia content, while multimodal approaches that integrate textual, visual, and auditory analysis show promise for capturing the complex nature of disinformation. Nonetheless, further advancements in transparency, model robustness, and interdisciplinary collaboration are needed to enhance the reliability and accessibility of these detection technologies.

METHODOLOGICAL PERSPECTIVES ON FAKE NEWS ANALYSIS

Before attempting the identification of fake news, researchers must reach a comprehensive understanding of its various forms. Thus, the study of the causality and motives behind fake news dissemination, including malicious entities, financial gains, and algorithm driven social media, becomes an important initial step. This also helps raise awareness about how fake news affects society, including eroding trust in media, challenges to political discourse, or even incitement of violence [1, 13]. Only after researchers have a detailed understanding of how disinformation works can they begin to explore methods for its detection.

A number of authors have adopted separate perspectives in studying fake news. Some investigate the clarity levels in news content [14,15], while others examine the processes behind the creation of fake news [16, 17]. In contrast, some studies focus directly on the transmission routes of fake news [18, 19], while others assess source reliability [20, 21].

As shown in Figure 1, there are currently four principal perspectives that address fake news detection: fact-checking approaches, stylistic analysis approaches, propagation dynamics approaches and source credibility approaches. The first two perspectives study fake news during its creation stage, while the latter two examine fake news after its dissemination.

Fact-Checking Approaches

Once raw data has been verified through fact-checking, researchers often analyze the stylistic elements of news content to identify deceptive writing patterns. Fact-checking approaches, otherwise referred to as knowledge based approaches, involve data gathering from different, often open, sources to build a knowledge base. These approaches may be expert oriented, crowd sourcing oriented or computational oriented [22]. The obtained data has to be further processed in order to address topics such as redundancy, invalidity, balance, or incompleteness. Five assessments are needed to make the datasets usable: data matching, time-stamping, uniformity analysis, exhaustiveness assessment, and reliability assessment. Data matching links records of related or similar facts, while time-stamping measures the temporal validity of facts. Uniformity analysis ensures factual consistency; exhaustiveness assessment verifies the inclusion of all relevant information; and reliability assessment validates fact truthfulness. By conducting these five tests, the data can be properly refined, enabling the successful implementation of fact-checking approaches [11].

Stylistic Analysis Approaches

Using such approaches, researchers aim to evaluate whether news content is proposed to mislead the public. This largely consists in observing and retaining the journalistic style of news content by categorizing styles through measurable attributes. The type of news is then categorized using different methods [23]. The data analyzed for style may be textual and visual. For example, textual patterns of fake news may involve relaxed tone, diversity, emotional bias, and visionary rhetoric. Conversely, fake images often lack diversity but offer a high clarity and coherence [24].

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

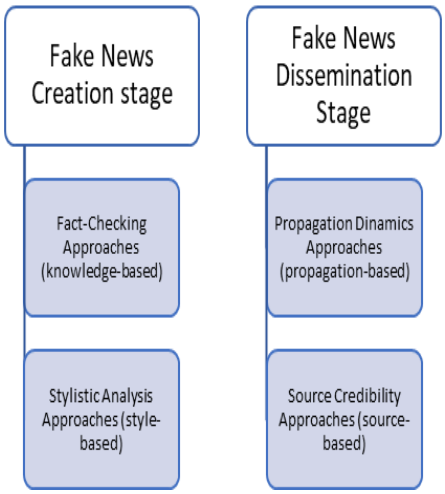

Figure 1: Perspectives in Fake News Detection. The diagram illustrates four main analytical perspectives categorized by stage: Creation Stage' involving fact-checking and stylistic analyses and 'Dissemination Stage' involving propagation dynamics and source credibility analyses

Propagation Dynamics Approaches

Using this approach, researchers investigate how information circulates across different media, predicting that news originating in anomalies is likely to be false. Propagation based studies study the ways users are involved in spreading disinformation [25]. The input may be news cascades or graphs. For the content input the propagation is directly shown while graphs serve as indirect representations that encapsulate additional insights about the propagation process. The news cascade is a tree like layout originated in the fake news source and comprising various nodes formed by users that further disseminated the news [26]. The graph is a representation of a network used for fakes news spreading and can be: homogeneous, made up of one type of node and edge; heterogeneous, containing multiple nodes and edges; and hierarchical networks composed of nodes organized into hierarchical layers, in which the root node begins the disinformation [27].

Source Credibility Approaches

Such approaches investigate and evaluate the credibility of sources. This involves assessing the trust worthiness of individuals and organizations that produce and share news. This evaluation can focus on both the content and the broader social context of the news [28]. One effective approach to determine source credibility is to analyze the roles of news creators and publishers. Malicious users on social platforms can fabricate and distribute news stories widely, while regular users might share misleading information without verifying its truthfulness. Research on sources examines user interactions with fake news and their involvement in its production, dissemination, and circulation [29].

METHODS FOR FAKE NEWS DETECTION

Although there are various methods available for classifying true from fake news, none can fully differentiate between true and false information because of various limitations. These include: the lack of comprehensive datasets, designed to address diverse contexts, modalities, and languages significantly which limit the generalizability of fake news detection models; challenges associated with preparing vast datasets for analysis; the diverse and dynamic nature of information; and insufficient exploration of multimodal data. Consequently, researchers have categorized fake news detection methods into different approaches. As illustrated in Figure 2, methods for fake news detection can be broadly classified into manual and automatic techniques, with subsequent methods that will be detailed in this paper.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

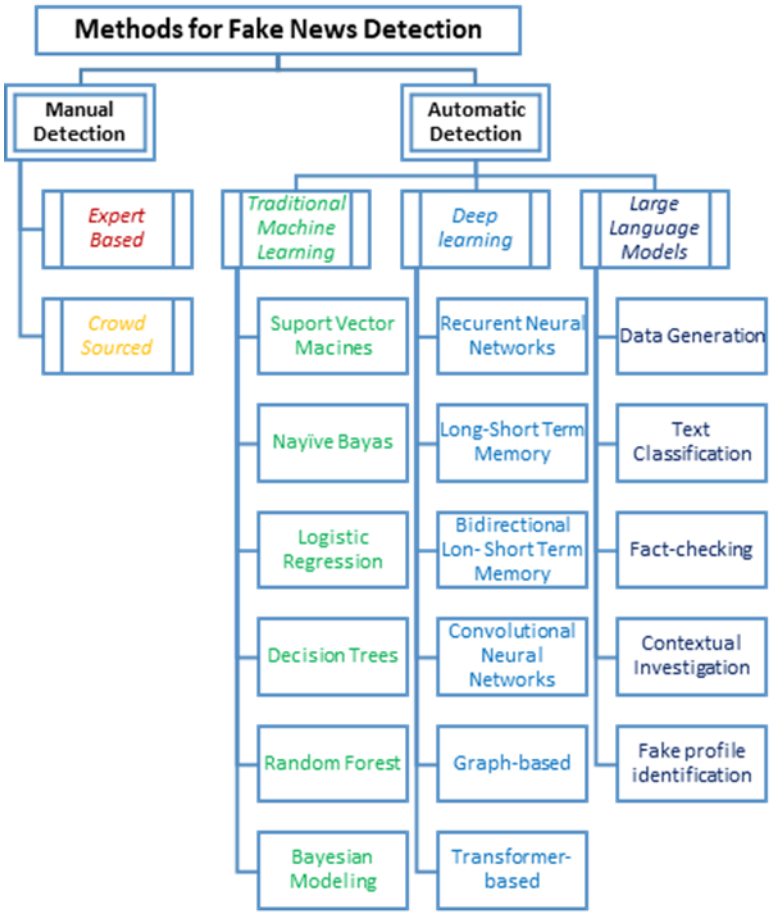

Figure 2: Methods for Fake News Detection. This figure outlines the proposed structure of detection methods, divided into Manual and Automatic methods.

Manual Detection Methods

Manual fact-checking involves obtaining factual verifications that allow readers to critically assess content for relevance and integrity. As of November 2024, the Duke Reporters' Lab [30] reports that the number of reputable fact-checking sites worldwide is 446 that are still active and 161 are marked as inactive. This figure reflects the global landscape of active fact-checking organizations dedicated to verifying information and combating misinformation. Manual fact-checking processes may be categorized into two primary groups: expert-based detection and crowd-sourced methods, as depicted in Figure 2.

Expert-Based Methods

This traditional fact-checking method involves experts, such as journalists or researchers, verifying claims against established evidence or previously reported facts. While thorough, expert-based fact- checking is time-consuming and costly. As of 2021, there were 391 active fact-checking projects world- wide, spanning 105 countries [31]. In the European Union, a comprehensive map lists numerous fact-checking organizations across member states. In Romania for example, five fact-checking initiatives are noted as active [32]. One of the main usages of expert-based reports in automatic fake news detection consists in the extraction of verified datasets, that are generally named after the fact checking organization or institution that gathered them, for example: ISOT, Kaggle, PolitiFact, Twitter or LIAR [32].

Crowd-Sourced Fact-Checking

Such a project leverages the collective intelligence of a large group of individuals acting as fact-checkers. This approach often requires filtering unreliable users and resolving conflicting verification results. Maintaining it is somewhat demanding, and it typically lacks the efficiency of expert-driven methods. These challenges gain

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

importance as the fact-checkers grow in number. Platforms like Fiskkit [33], Community Notes (formerly Birdwatch) [34], Debunk.org [35], StopFake [36], and Alt News [37] exemplify the growing reliance on crowd-sourced efforts to identify and combat misinformation online. However, studies have indicated that humans are not particularly effective at identifying false information. Accuracy rates for human detection of false information, including hoaxes and fake news, range between 53% and 78%. Even well informed readers can be deceived by well-crafted false content [38].

Automatic Detection Methods

Automated detection of fake news aims towards reducing the reliance on human effort to mitigate the rapid dissemination of false information. Various labeling and scoring strategies have been developed for this purpose. Machine learning and deep learning represent two primary subfields in data science, which is employed extensively in this domain.

Traditional Machine Learning Techniques

Such methods are still extensively utilized to tackle the issue of fake news detection, employing both supervised and unsupervised methods. As of 2024, many researchers continue to use core machine learning algorithms for detecting fake news, frequently making adjustments to improve their effectiveness in this area.

Support Vector Machines Support Vector Machines (SVMs) are utilized to classify fake news by stylistic approach, media source, and secondary data about the article. The basic approach involves inputting the text to be analyzed, extracting relevant features, and then classifying it in a binary way, as real or fake news. SVMs are capable of rapidly handling vast datasets, are accurate, and can make real-time predictions.

A 2024 study evaluated fake news detection on social media using several machine learning algorithms (Na¨ıve Bayes, Logistic Regression, SVM, Decision Tree, Random Forest, K-Nearest Neighbor) and deep learning models (CNN, LSTM) on a COVID-19 dataset containing 1,375,592 tweets. Data preprocessing involved normalization, tokenization, and TF-IDF feature extraction. The Support Vector Machine algorithm achieved the highest accuracy (98%), followed by Logistic Regression (95.2%), while LSTM yielded the lowest accuracy (65%) [39].

Another notable approach applied a SVM classifier to the ISOT Fake News dataset from Kaggle. Experimental settings involved dataset preprocessing (tokenization, stop-word removal, TF-IDF feature extraction), splitting data into 80% training and 20% testing subsets. The procedure utilized a linear SVM kernel for classification. Results demonstrated an overall accuracy of 99%, correctly classifying 4,542 fake and 4,264 true news items, with minimal misclassification (20 false positives, 26 false negatives) [40].

A hybrid method involving SVM and K-Nearest Neighbors (KNN) improved on the nuanced understanding of fake news identification, highlighting limitations of traditional Na¨ıve Bayes algorithms, primarily the assumption of conditional independence among features, which often does not hold true in real-world text data. Using the BuzzFeed News dataset, the experimental procedure involved text preprocessing, feature extraction with truncated Singular Value Decomposition (SVD), and classifier training. The enhanced Na¨ıve Bayes model achieved 99% accuracy, outperforming Random Forest (80%), standard Na¨ıve Bayes (69%), and Passive Aggressive (87%) classifiers [41].

Naive Bayes This machine learning algorithm is built upon Bayes' theorem, which calculates the probability of a hypothesis being true given some evidence. Within the domain of fake news detection, Naive Bayes (NB) is capable of processing large datasets to identify patterns that suggest the presence of misinformation. For instance, it can recognize words and phrases typically linked to fake news articles, as well as evaluate an article's writing style by comparing it to other sources. Additionally, it can analyze the overall sentiment of an article to gauge its probability of being true or false.

Such a new variant of model achieved 99% accuracy on a BuzzFeed fake news dataset by introducing an enhanced hybrid Na¨ıve Bayes model combined with truncated Singular Value Decomposition (SVD) for fake news detection. Experiments involved text preprocessing (cleaning and normalization), feature extraction through dimensionality reduction using SVD, and training classifiers. Conducted in a supervised learning environment, the new variant significantly outperformed standard classifiers in accuracy, like Random Forest (80%), standard

Na¨ıve Bayes (69%), and Passive Aggressive (87%). [42]. A similar study identified hoax or insourced news in internet media using the Na¨ıve Bayes Multinomial method. The experimental procedure involved text vectorization with Count Vectorizer to convert text into numeric vectors The model achieved an accuracy of 94.73% when using an 80% training and 20% testing data split. The confusion matrix results were: True Negative (TN) = 4,555, False Positive (FP) = 178, False Negative (FN) = 295, and True Positive (TP) = 3,952. The findings prove that the model has a high reliability in practical applications [43].

Logistic Regression this is an effective statistical method for detecting fake news by evaluating different traits of news articles. It analyzes aspects such as the origin, title, content, writer, and additional characteristics that could indicate the news is deceptive. Through this analysis, the algorithm estimates the likelihood that the news article is false.

A key benefit of Logistic Regression is its capacity to produce accurate results regardless of the size of the dataset, which is especially advantageous for large collections of news articles.

A research group obtained an accuracy of 97.92% using an optimized Logistic Regression (LR) algorithm against traditional Linear Regression for predicting fake job postings. Using a dataset of job postings, the experimental procedure involved data preprocessing, feature extraction, and model training [44].

A combined method of SVMs, LR, and Long-Short-Term-Memory (LSTM), through Stacking and Delegation, obtained decent 95.09% and 95.62% accuracy scores. The experimental procedure involved analyzing preprocessing techniques such as count vectorization and TF-IDF to assess their impact on detection effectiveness. Results demonstrated that ensemble methods, particularly probability-based stacking, achieved high Area under the Curve (AUC) scores of 0.9394 and 0.9509. Delegation strategies also performed well, with iterated delegation reaching AUCs of 0.9280 and 0.9477. This technique is currently the most accurate next to the SVMs. [45].

Decision Trees: Decision Trees (DT) are powerful machine learning algorithms for fake news classification, known for their capacity to process large datasets rapidly and precisely. They can uncover connections between data points by evaluating factors like the article's source, linguistic style, main content, and contextual framework. Thus, DTs can assist in pinpointing the individuals or groups responsible for spreading fake news.

Typically, DT methods outperform SVMs, like in this 2022 experiment that achieved higher precision and an accuracy range of 90% to 97% across approximately 10 iterations, outperforming a SVM algorithm, which had an accuracy of around 91.5%. [46]. However, 2024 studies claim higher scores. For example, a study comparing Logistic Regression, Decision Tree, and Random Forest models using a Kaggle dataset of news articles from the 2016 election period that involved dataset preprocessing, feature selection, training, and validation within a Python-based supervised learning environment shows superior results for Decision Tree. The model achieved the highest accuracy (99.64%), outperforming Random Forest (99.23%) and Logistic Regression (98.80%) [47].

A somewhat different study proposes a fake news detection method based on DT that integrates a hybrid metaheuristic optimization algorithm IBAVO-AO (African Vultures Optimization + Aquila Optimization), with an XGBoost Tree classifier. Experimental settings involved preprocessing over 44,000 ISOT news articles using GloVe embeddings and Relief feature selection. The hybrid algorithm optimizes the selection of relevant features, enhancing classification accuracy. Experimental results demonstrated superior performance compared to existing methods, achieving over 92.5% accuracy [48].

Random Forest Classifier: Random Forest (RF) is a highly flexible classifier capable of identifying fake news by building numerous decision trees during the training phase and taking the majority vote of their predictions. In the context of fake news detection, this algorithm analyzes various aspects such as the content of news articles, writing style, tone, and source. It assesses all these features collectively to determine the likelihood of an article being fake. The RF algorithm can also detect biases in the data, shows lower susceptibility to over fitting, and performs effectively on new data. A modified RF model claimed 99.32% accuracy over a Kaggle dataset in 2023 [49], while another research team proposed a three-stage approach: data preprocessing (stop-word deletion, stemming, tokenization), feature selection using the Honey Badger (HB) optimization algorithm, and classification via a lightweight convolutional random forest (LCRF) algorithm. This methodology was

applied to a dataset of COVID-19-related news articles. The LCRF-HB model achieved an accuracy of 98.7%, precision of 98.3%, specificity of 95.4%, and recall of 97.6% [50].

Bayesian Modeling: Bayesian Modeling (BM) utilizes probability theory to identify and interpret data patterns, which makes it quite effective for identifying disinformation based on its content.

For example, a news report that contains false information may feature specific expressions or terms often associated with fake news. BM can recognize these patterns to create a classification model for such stories. Additionally, the use of BM priors enables researchers to integrate prior knowledge regarding the likelihood of news being true. This capacity sets Bayesian methods apart from other modeling techniques, including deep learning algorithms.

An Indonesian team used Complements Na¨ıve Bayes (CNB) to develop a fast fake news detection system for COVID-19-related content on social networks. The team used Multinomial Na¨ıve Bayes (MNB), Complement Na¨ıve Bayes (CNB), Decision Tree (DT), and Gradient Boosting (GDBT) models. The experimental environment involved a dataset of 10,700 tweets, with preprocessing (tokenization, TF-IDF, DFT), SMOTE for class balancing, and hyperparameter tuning via GridSearchCV. CNB and MNB achieved the best performance, with 92% accuracy, precision, recall and F1-score. CNB also had the shortest runtime (0.55s), making it the most effective model for real-time fake news detection on COVID-19 content [51].

Deep Learning Techniques

Models from the Deep Learning (DL) category have been extensively studied because of their capacity for automatic learning of layered feature representations from raw data. Neural networks, especially those tailored for natural language processing (NLP), have demonstrated significant potential in identifying fake news [52]. Research teams have explored the creation of fake news, its dissemination, the challenges in detection, and various perspectives within the realm of fake news research, as illustrated in Figure 1. However, the absence of a universally accepted definition of ”fake news” presents significant challenges to research in this field, particularly regarding dataset standardization.

This conceptual ambiguity has led to the development and collection of datasets through diverse methodologies and analytical strategies, resulting in a scarcity of datasets that meet uniform standardization criteria. Consequently, deep learning techniques have proven particularly effective in addressing these limitations. Deep learning models, especially those based on neural networks, exhibit strong capabilities in managing heterogeneous and unstructured data. Their ability to learn hierarchical representations enables them to identify complex patterns within textual content, thereby facilitating the distinction between authentic and fabricated information.

Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been successfully utilized in fake news detection by capturing linguistic features and contextual nuances. These models offer substantial adaptability, allowing fine-tuning across datasets with varying structures and labeling schemes. This flexibility is particularly valuable in light of the ongoing lack of dataset standardization. Through techniques such as transfer learning and domain adaptation, deep learning models can consistently achieve high performance across disparate datasets, effectively mitigating some of the challenges posed by variability in data sources [53].

Recurrent Neural Networks (RNNs): These models excel at handling sequential data, which makes them ideal for analyzing text, speech, and video content in the context of fake news detection. By retaining a memory of earlier inputs, RNNs can uncover patterns and relationships that signify misinformation, including particular word choices and phrasing trends. Additionally, they are capable of recognizing discrepancies among various sources. In 2024, a method integrating GloVe word embeddings with a bidirectional Long Short-Term Memory (BiLSTM) network was proposed in order to enhance fake news detection accuracy. The experimental setup involved preprocessing textual data using GloVe to capture word relationships, followed by classification with deep learning models, including CNN, DNN and RNN. Utilizing the Curpos fake news dataset, the RNN with GloVe preprocessing achieved the highest accuracy of 98.974%, outperforming other models in detecting false news [54].

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Long-Short Term Memory (LSTMs): LSTM models represent a specific kind of RNN designed to learn long-term dependencies while addressing the vanishing gradient issue. They are particularly adept at identifying patterns across lengthy sequences, which makes them highly effective for spotting similarities and differences in news articles. Additionally, LSTMs are capable of analyzing writing styles and language usage.

Recent 2024 research still provides a somewhat underwhelming result when concerning classical models. A recent study utilized a Kaggle-sourced fake news dataset to evaluate a LSTM model com- pared against Support Vector Classifier, Logistic Regression, and Multinomial Naive Bayes methods. The LSTM achieved a 94% accuracy score [55].

However, a hybrid RNN-LSTM algorithm proposed in 2023 shows potential. The research proposes a fake news detection system using a Recurrent Neural Network–Long Short-Term Memory (RNN- LSTM) model with hybrid feature extraction techniques (TF-IDF, N-gram, POS tagging, etc.). The experimental environment included Java-based implementation on a standard CPU setup. The LIAR dataset was preprocessed and split (70% training, 30% testing) and classification was performed using RNN-LSTM and compared with models such as SVM, ANN, RF, NB, DT, and DNN. The RNN-LSTM with ReLU activation achieved the highest accuracy of 99.1%, outperforming other methods across precision, recall, and F1-score, especially under 15-fold cross-validation [56].

Bidirectional Long Short-Term Memory (Bi-LSTMs): Bi-LSTMs analyze data in both forward and backward directions, allowing them to capture context from both past and future states simultaneously. This bidirectional approach boosts the model's capacity to grasp contextual information and identify subtle patterns related to fake news, thereby enhancing classification performance.

Recent advancements, using a Self-Attention Bidirectional LSTM (SA-BiLSTM) model trained on the ISOT dataset containing over 44,000 labeled news articles, show superior results to traditional LSTM models. The experimental environment used in 2024 consists of Python, Keras, and TensorFlow on a CPU-GPU setup. The procedure included text preprocessing, word embedding, model training (70% data), and testing (30% data), optimized with ADAM and SGD. The SA-BiLSTM model outperformed CNN, GRU, LSTM, and other baselines, achieving 99.98% accuracy, 99.96% F1-score, and 99.98% AUC, identifying misinformation through contextual attention-based learning [57].

Convolutional Neural Networks (CNNs): Originally created for image processing, CNNs have been repurposed for text analysis in the realm of fake news detection. They are capable of identifying local features and patterns in textual data, including n-grams and phrases that suggest deceptive content. Their efficiency in managing large datasets makes them particularly useful for examining the extensive information related to fake news.

Recent contributions toward fake news detection using CNN introduced a model employing word embeddings to detect COVID-19-related fake news. The experimental setup involved hyperparameter optimization using grid search to identify the optimal CNN architecture. The model's performance was evaluated against various state-of-the-art machine learning algorithms. The proposed CNN model achieved a mean accuracy of 96.19%, a mean F1-score of 95%, and an area under the ROC curve (AUC) of 0.985, outperforming other methods in detecting COVID-19 fake news [58].

Another hybrid CNN–Bi-LSTM–SelfAttention model attained a 98.71% accuracy over a COVID-19 dataset.The model showed improved evaluation metrics such as Loss, Accuracy, F1-score, and Recall by at least 1%, outperforming baseline models. The experimental environment utilized Python with TensorFlow and Keras libraries. Procedures included data preprocessing, feature extraction via CNN, sequence learning through BiLSTM, and emphasizing relevant features using AM. [59].

Graph based models Propagation Networks Models such as the Propagation Graph Neural Network (PGNN) [60] focus on how fake news spreads through social networks, using graph struc- tures to model and detect anomalies in information dissemination. Graph-based models have been successfully applied to multimodal datasets (text-image).

For example, a 2024 study proposes a multimodal fake news detection system integrating Text Graph

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Convolutional Networks (TextGCN) and Vision Transformers (ViT), utilizing both textual and visual features from the Fakeddit dataset. The experimental environment included TensorFlow, Scikit-learn, and HuggingFace tools on a system with Intel i7, 16 GB RAM, and RTX 3060 GPU. Text data was processed into graphs with TF-IDF and PMI, while images were encoded via pretrained ViT. Feature fusion was performed using Random Kitchen Sink (RKS) mapping and classified using Artificial Neural Networks (ANN), Random Forest, and SVC. Results showed that ANN outperformed other models, achieving 94.17% accuracy (binary), 90.14% (3-class), and 75.91% (6-class), with precision and recall of 0.98. The study demonstrates the superior performance of multimodal fusion over unimodal methods and highlights future directions in handling class imbalance and applying explainable AI techniques [61].

Transformer Based Models: Transformer based models are a subclass of deep learning models designed to handle sequential data, predominantly for tasks in natural language processing (NLP). They were introduced in the foundational study ”Attention is All You Need” [62] and are renowned for their use of attention mechanisms to process and generate data efficiently. One of the most widely used classes of models for fake news classification is built on the architecture called Bidirectional Encoder Representations from Transformers (BERT) [63]. Since models based on this architecture consistently report high accuracy [64] on established datasets obtained from the work of reputable fact-checking associations, the latest developments for this family of models involve a more integrated and complex exploration of multi-class datasets, as opposed to a binary true/fake approach.

For example, a 2024 study explores multiclass fake news detection using transformer-based mod- els—mBERT, SBERT, and XLM-RoBERTa—on the CheckThat-2022 dataset, classifying news into true, false, partially false, and other categories. Conducted on Amazon Azure, experiments involved training with both original and ChatGPT-augmented data to mitigate class imbalance. Results showed mBERT achieved the best performance (accuracy: 34%; F1-score: 0.23). ChatGPT-generated data improved classification, especially for underrepresented classes, achieving a macro F1-score of 0.26 [65].

Another recent study proposes a multi-class fake news detection framework using an ensemble of advanced machine learning and deep learning models (BERT, RoBERTa, BiLSTM, and LightGBM) on the LIAR dataset. The experimental setup involved extracting contextual features using BERT and RoBERTa, capturing sequential dependencies via BiLSTM, and integrating textual and numerical features through a LightGBM stacking ensemble. The models were trained and evaluated using k-fold cross-validation and assessed by accuracy, precision, recall, and F1-score. The ensemble model achieved the highest performance with a 41% accuracy and 0.42 F1, outperforming individual models [66]. A different new perspective that involves a Graph Convolution Network and BERT combined with a Co-Attention (GBCA) model is applied on 3 datasets: Twitter15, Twitter16 (that contain 1490 and 818 claims) labeled with four labels: true news (T), fake news (F), non-fake news (NF), unverified news (U) and PHEME, that has 9 classes containing 2402 claims, which are annotated as fake news (F), true news (T) and unverified news (U). The proposed model shows improvement over the base versions when dealing with unstructured multi label data and takes into account user interaction [67].

Applying a combination of models like BERT and VGG-19 on multimodal datasets (text-image) represents another research path utilizing transformer architecture. A recent study proposes a dual- phased framework for fake news detection by integrating transformer and deep learning techniques on both textual and multimodal data. Textual experiments used the ISOT dataset with models such as Random Forest, SVM, and Logistic Regression. Random Forest achieved 99% accuracy, precision, recall, and F1-score. The multimodal approach combined BERT-based textual embeddings and ResNet-based visual features with an attention mechanism, evaluated on the MediaEval 2016 dataset. The proposed model outperformed existing baselines like SpotFake, achieving 94.4% accuracy and an F1-score of 0.892, showing a 3.1% improvement in accuracy showing robust detection across diverse data types [68].

A similar study involved locally fine-tuning pre-trained BERT and VGG-19 models. BERT was entirely retrained, while VGG-19 underwent structural modifications, including the addition of a global average pooling layer and a redesigned classifier. Experiments conducted on the Fakeddit binary dataset demonstrated that this approach achieved a high accuracy rate of 92% [69].

Other encouraging results on nuanced (multi-class) classification are shown by ELD-FN, an ensemble deep

learning model for detecting multi-modal fake news using the Fakeddit dataset (over 1 million samples). Its experimental environment includes NLP preprocessing (tokenization, lemmatization, sentiment analysis), feature extraction via V-BERT for both text and images, and a combination of bagged/boosted CNN and LSTM classifiers. A 10-fold cross-validation was applied. ELD-FN achieved 88.83% accuracy, 93.54% precision, 90.29% recall, and 91.89% F1-score, outperforming baseline models (FakeNED, MultiFND). Results also highlight the positive impact of sentiment analysis, preprocessing, and oversampling [70].

Transformer architecture was also fine-tuned for multi-language usage. Such a comprehensive framework was presented in 2024 for fake news detection in Turkish using a newly constructed dataset, TR FaRe News, consisting of 18,695 manually labeled tweets. The experimental environment involved data preprocessing with the Zemberek NLP tool and classification using both traditional machine learning models (SVM, RF, NB, LR, Voting Classifier) and advanced deep learning models (BERTurk, DistillBERTurk with CNN and Bi-LSTM). Experiments were conducted on TR FaRe News and multi- ple benchmark datasets (ISOT, LIAR, GPT-2, Twitter15 and 16, etc.). The BERTurk and CNN model achieved 94% accuracy, outperforming baselines. The framework demonstrates strong cross-dataset generalizability for BERT-based models [71].

Similarly, having a future impact in fake news research using transformer based models, large-scale empirical study analyzing factors influencing cross-lingual zero-shot transfer in multilingual language models like BERT have been started [72], further adapting such models trained on high-resource languages to low-resource languages by utilizing external dictionaries for tokenizer adaptation on Silesian and Kashubian languages [73].

Large Language Model (LLMs) Based Techniques

Recent advancements in LLMs have significantly impacted areas involving the generation and detection of fake content and phony accounts. Their advanced capability to understand and produce text that closely mirrors human writing has been leveraged not only to create intricate misleading content but also to improve the efficacy of identification methods. The methods used for detecting fake news have advanced significantly with the incorporation of LLMs.

Traditional and deep learning techniques typically depend on supplementary data such as the date, author, publication source, or subject matter, alongside the article's text. However, these methods encounter practical limitations, as such secondary data is not always readily available. Additionally, utilizing online fact-checking services for dataset collection can be time-consuming, hindering the ability to keep pace with the fast-paced creation of disinformation [74].

Recent research has demonstrated that customized pre-trained language models (PLMs) like BERT, or LLMs, including LLaMA 3 and GPT-4, can effectively detect fake news without relying on substantial supplementary data, using different approaches [17].

LLM-based Dataset Generation

To overcome current limitations, new methods have been devised which incorporate a combination of actual news and truthful data mixed with deliberately false content created by human operators. One study is called MegaFake, a large-scale, theory driven dataset of fake news generated by Large Language Models (LLMs), constructed using social psychology-based framework called LLM-Fake Theory. The experiment included natural language understanding (NLU) and natural language generation (NLG) models trained and tested on MegaFake and GossipCop datasets. Six NLU models (e.g., CT-BERT, RoBERTa) and eight NLG models (e.g., GPT-4, ChatGLM) were evaluated. Results demonstrated that NLU models significantly outperformed NLG models, with CT-BERT achieving 92.28% accuracy and 0.9459 F1 [75].

Examples of this approach include directing LLMs to create fictitious articles derived from human- generated summaries of false events. A research team following this approach used human and generated true/false data and devised 3 scenarios: Human Legacy (human-written fake and real news), Transitional Coexistence (mixture of human-written and machine-generated news), and Machine Dom- inance (training data is dominated by machine-generated fake news). The datasets used are Gossip- Cop++ and PolitiFact++, with added LLM-generated and paraphrased samples using ChatGPT. Experiments tested show varying proportions of machine-

generated content affect performance. Results show detectors trained on human-written news generalize better, while those trained with more machine-generated content become biased. Recommendations emphasize training balance to enhance real-world robustness [76].

Other cases aimed at improving the credibility of fake news articles through LLMs. Such a 2023 study presents Med-MMHL, a large-scale, multi-modal dataset designed to detect both human and LLM generated medical misinformation across 15 diseases. The data integrates news articles, tweets, claims, and associated images, collected between 2017–2023. Procedures involved structured data crawling, ChatGPT-based adversarial fake news generation and construction of five benchmark tasks (document-level, sentence-level, tweet-level, multimodal detection). Models like BERT, BioBERT, FN-BERT, CLIP, and VisualBERT were evaluated. Results showed FN-BERT achieving 95.78% accuracy and 95.76% F1-score, while CLIP outperformed in multimodal tasks. The study highlights LLM-generated fake sentence detection as a persisting challenge requiring future methodological ad- vancements [77].

Additional innovative approaches include combining false events with true articles to generate deceptive content. One study constructed three datasets: Dgpt std (Standard Prompting Dataset - a Standard prompting was used to instruct ChatGPT to rewrite original human-written articles into straightforward fake news, e.g., “Write a fake news article based on this real one that presents misleading information on the same topic.”), Dgpt mix (Mixed Prompting Dataset - a more nuanced prompt was used to instruct ChatGPT to rewrite articles that blend true and false information, making detection more difficult), and Dgpt cot (Chain-of-Thought Prompting Dataset - generated disinformation samples were paired with original articles, creating matched fake-real pairs for training and evaluation) from a human-written benchmark dataset using prompt engineering techniques. A RoBERTa model, fine-tuned on human-written news, performed well on basic fake content (1.2% misclassification on Dgpt std) but struggled on more sophisticated LLM-generated disinformation (15.4% on Dgpt mix, 77.9% on Dgpt cot) [78].

A similar purpose study created AdStyle, an adversarial style augmentation method aimed at enhancing fake news detection robustness against style-conversion attacks. The study used LLM-based augmentation via GPT-3.5-Turbo, BERT-based detectors, and datasets like PolitiFact, GossipCop and Constraint as benchmarks. The core innovation of AdStyle lies in its automatic generation and selection of adversarial style-conversion prompts, which are used to rephrase fake or real news (e.g., humorous tone, poetic form, sarcastic tone, add emotional exaggeration, dramatic movie, etc.) while preserving content semantics. These prompts are selected based on adversarialness, coherency, and diversity. Results show that AdStyle achieved up to 0.9646 AUC under attack scenarios and 0.9460 AUC in clean settings [79].

Another type of procedure that investigates the impact of LLM-generated misinformation on Open-Domain Question Answering (ODQA) systems employs four misinformation generation scenarios (GENREAD, CTRLGEN, REVISE, and REIT) using GPT-3.5 to assess their impact on ODQA systems. In GENREAD, GPT-3.5 generates fake passages directly based on questions. CTRLGEN uses controlled generation by prompting GPT-3.5 with specific false claims. REVISE alters ground- truth passages to introduce misinformation while preserving context. REIT (Reinsertion and Iterative Training) involves multiple iterations of misinformation injection and retraining to simulate persistent disinformation environments. These generated passages were injected into QA corpora (NQ-1500 and CovidNews), and the effects were evaluated using BM25 and DPR retrievers with FiD and GPT-3.5 readers. Results show a performance drop of up to 87%, highlighting the vulnerability of ODQA systems to misinformation [80]. Such approaches, especially those centered on manually created fake content, are effective in curbing the automated large-scale generation of misleading articles. However, methods that depend on fabricated summaries frequently result in content that lacks depth, and modifications to particular events or components can create problems with maintaining contextual coherence.

A 2024 research involved the enhancing multilingual fake news detection by augmenting datasets using Llama 3, a large language model (LLM). The experimental environment employed BERT-based classifiers evaluated on two real-world multilingual datasets: TALLIP and MM-COVID. The procedures included translating all samples to English, generating synthetic news samples using Llama 3 with prompt-based paraphrasing, and applying various augmentation strategies: augmentation rates, random vs. similarity-based sampling, and class specific

augmentation (only fake, only real, or both). Results showed the "only fake" augmentation strategy at rate 1 significantly improved F1 scores, with a 7.7% boost for English and 4.4% for Hindi. However, a key concern within this experiment is that fabricated summaries or paraphrased content may not preserve the full complexity, narrative nuance, or factual structure of real-world news [81].

A different, Algerian team study explores the effectiveness of translation-based data augmentation for fake news detection focusing on the Algerian dialect. The experimental environment includes GPT-4 evaluated using BLEU, Chrf++, COMET, and expert human judgments. Pre-trained models—AraBERTv02, MARBERTv2, and DziriBERT—were fine-tuned on the Khouja fake news corpus translated from Modern Standard Arabic (MSA). Experiments assessed manual vs. automatic translation and various augmentation sizes. Results showed that automatic translation enhanced recall but reduced precision due to noise. AraBERTv02 achieved the best F1-score (0.67). This confirmed limitations of augmentation in dialect sensitive languages as it may lose nuances, misinterpret or distort rhetorical structures, idiomatic expressions, or cultural relevance present in the original text [82].

Text Classification

Similar to previous models, LLMs are fine-tuned on labeled datasets that include instances of both real and fake news. This technique takes advantage of the language understanding abilities of LLMs to identify subtle indicators of misinformation. In recent years, an emphasis on fine-tuning pre-trained models on datasets containing real and fake news has become essential for improving accuracy in misinformation recognition. The classification approach utilizes the powerful capabilities of PLMs and specifically tailors them for recognizing disinformation, either between true versus false classification or within multiple categories, resulting in notable enhancements in performance metrics.

As we examine the 2024 SheepDog research, a style-agnostic fake news detection model designed to withstand adversarial style-based attacks generated by LLMs, we observe that traditional detectors suffered up to 38% performance loss under style attacks, while SheepDog significantly outperformed baselines across all datasets, achieving up to 93.04% F1 on LUN. The experimental environment involved testing on benchmark datasets like PolitiFact, GossipCop and LUN, with adversarially restyled fake news using GPT-3.5 and LLaMA-2. The procedures included LLM-based news reframing, style- agnostic training, and content-focused veracity attribution [83].

Another type of approach includes fine-tuning LLMs such as ChatGPT and Gemini with the LIAR benchmark dataset has led to accuracy scores between 89% (ChatGPT) and 91% (Gemini) [84].

In 2024, a cross-domain fake news detection framework called FakeNewsGPT4 was introduced. This framework enhances large vision-language models by incorporating forgery-tailored data to improve manipulation reasoning while also leveraging comprehensive world knowledge as a supplement. The experimental environment used ImageBind and Vicuna LLMs, trained on DGM4 and NewsCLIP- pings datasets under single-domain, multi-domain, and cross-dataset settings. The procedure involved injecting two knowledge types (semantic correlation and visual artifact traces) via dual lightweight modules, followed by instruction-tuned alignment. Results showed significant improvements over base lines, with FKA-Owl achieving up to 89.61% AUC and outperforming SOTA methods like HAMMER and PandaGPT, especially under domain shift [85].

The latest ChatGPT, Gemini, and Meta LLaMA 3 models were used for nuanced fake news detection. A 2024 study investigates the effectiveness of a fine-tuned LLaMA 3 (8B) model in multi-class fake news detection across bilingual datasets (English and Romanian). The procedure involved tokenization, normalization, and model fine-tuning, followed by comparative evaluation against ChatGPT-4, Gemini, LLaMA 2, and other baseline models. Results show the proposed model achieved a accuracy of 39% on Romanian data, outperforming all LLaMA versions [86].

A similar experiment using a fine-tuned LLaMA-3 (8B) model within the IberLEF 2024 FLARES Subtask 2 across multiple language datasets included training on the RUN-AS dataset, annotated via the 5W1H journalistic method (the classic framework used in journalism and information analysis to ensure comprehensive and structured reporting - who, what, when, where, why, how). The system achieved a Macro F1-score of 0.59658, ranking

second overall in the task [87], showing improvements over other methods.

Fact-Checking

LLMs can automate fact-checking by comparing statements within an article to trustworthy knowledge bases or databases containing verified information. This approach identifies factual inconsistencies, contradictions, or inaccuracies that may suggest falsehoods. More sophisticated models are capable of analyzing complex sentences and grasping context, which enhances the accuracy of the verification process.

A research team proposed two methods using LLMs. The first, known as Reinforcement Retrieval Leveraging Fine-grained Feedback (FFRR), seeks to enhance the accuracy and information value of evidence retrieval. [88]. It was assessed on two public datasets and demonstrated significant improvements over strong LLM-enabled and non-LLM baselines. FFRR generates sub-questions examining various aspects of a claim using prompting techniques to gather relevant data. Subsequently, FFRR gathers detailed responses from the LLM model regarding the retrieved data across both documents and question dimensions. The feedback is utilized as rewards to refine the document retrieval list and enhance the retrieval strategy for intermediate queries. The experimental environment is based on RoBERTa and evaluation is achieved using RAWFC and LIAR-RAW datasets. Results show that FFRR (document + question level) significantly outperforms all baselines, achieving up to 57% macro F1 on RAWFC and 33.5% on LIAR-RAW.

A second method for facct-checking, named Hierarchical Step-by-Step (HiSS), guides LLMs to break down a claim into multiple subclaims, then validates them step-by-step using a series of question- answering rounds [89]. This procedure involves decomposing the claim, verifying each subclaim, and then offering a definitive prediction. Initially, the LLM breaks the claim into subclaims to ensure no detail is overlooked. Next, it validates each subclaim by creating and responding to a sequence of in-depth queries, referencing external sources as needed. After verification, the LLM provides a final classification of the original data. Experiments conducted on two open-source fake and true news datasets indicate that HiSS prompting exceeds the performance of other fully supervised state-of-the- art methods and robust few-shot contextual learning benchmarks. Compared to standard prompting, vanilla Chain-of-Thought (CoT), and ReAct prompting, HiSS achieved the best performance, surpassing strong supervised models with 53.9% F1 on RAWFC and 37.5% on LIAR. HiSS also produced superior, fine-grained and explainable reasoning outputs. HiSS tackles two significant problems in news claim evaluation: omitting critical details and introducing “hallucinations” as facts.

A different methodology involves fine-tuning LLMs to generate coherent explanations that validate or critique news headlines, thereby providing transparent justifications for each classification [90]. This strategy integrates the predictive capabilities of LLMs with the ”Chain of Thought” (CoT) reasoning technique, enabling models to produce step-by-step rationales that mirror human reasoning processes. By employing model distillation, the researchers enhance detection accuracy and simplify the models’ complex decisions for better human comprehension. By fine-tuning FLAN-T5 and Llama-2, their approach demonstrated significant improvements, outperforming existing state-of-the-art models by 11.9% and enhancing the overall performance of LLMs in disinformation detection.

Another relevant study, called PASTEL (Prompted Weak Supervision with Credibility Signals), utilizes large language models to generate weak labels for various credibility indicators using prompts [91]. It starts with an open-ended prompt to extract answers from the news article, which are then classified using a generic prompt. Credibility signals are assessed by pairing an instruction prompt with the article and then applying a specific prompt for each signal in sequence. A zero-shot prompt identifies fabrication as a binary class (true versus fake news). If string-matching cannot assign a class, a task-neutral mapping prompt is used. This technique, which combines zero-shot labeling with weak supervision, surpasses the performance of other state-of-the-art classifiers tested on two fake news datasets and eliminates the need for ground-truth class labeling in the model training phase. PASTEL was evaluated on four article-level datasets: PolitiFact, GossipCop, FakeNewsAMT, and Celebrity.

Results show PASTEL significantly outperforms unsupervised baselines (↑38.3%) and achieves 86.7% of the performance of a supervised RoBERTa model, while demonstrating 63% better cross-domain

generalization.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Contextual Investigation

By examining the context in which information is delivered, LLMs evaluate the coherence, consistency, and plausibility of the content within a larger narrative framework. This process involves assessing the logical progression of information, detecting inconsistencies, and recognizing manipulative language or rhetorical techniques typically found in fake news. Additionally, semantic analysis aids in uncovering the underlying meanings and intentions behind the text.

A relevant study for this method introduced a designed network named the Adaptive Rationale Guidance (ARG) [92]. ARG combines small and large language models, enabling smaller models to select useful explanations for decisions. It encodes inputs with small models, which leverage informative rationales from larger models, considering written details, common knowledge, and factual accuracy to enhance performance. The models collaborate through news-rationale interactions, LLM judgment prediction, and rationale evaluation, facilitating deep engagement between news and rationales. Final judgments are based on aggregated interactive features. Furthermore, ARG-D, a cost-sensitive version of ARG that is rationale-free, has been proposed to operate without large model queries, demonstrating the effectiveness of both ARG and ARG-D in experimental conditions. Results show ARG outperforms all baselines, achieving macro F1 scores up to 0.784 (Chinese) and 0.790 (English), while the distilled ARG-D model retains high accuracy in cost-sensitive scenarios.

The formerly studied method, SheepDog [83], uses LLM-based news reframing to tailor articles to different styles through specific prompts. This approach ensures the detector remains resilient to stylistic variations by prioritizing content over style to maintain prediction consistency.

A different research team developed DELL, a framework for misinformation detection with three stages [93]. Initially, LLMs generate fabricated reactions toward news to reflect different viewpoints. Secondly, they provide explanations for proxy tasks, refining feature embedding. Thirdly, DELL uses three LLM strategies to combine classifications from expert models for better calibration. Experiments on various datasets show that DELL improves the macro F1-score by up to 16.8% over state-of-the-art baselines.

Fake Profile Identification

Leveraging the advanced natural language understanding (NLU) capabilities of Large Language Models (LLMs), researchers and platforms have developed methods to identify fake profiles by analyzing linguistic patterns, behavioral anomalies, and contextual inconsistencies.

A relevant example is MedGraph, a model designed for use on online dating platforms [94]. Med- Graph employs an embedding-based Graph Neural Network (GNN) framework to identify deceptive actions, specifically edges within a temporal mutual graph. This approach integrates both user at- tributes and their interaction patterns. The methodology involves several key components: a Motif- Based Graph Neural Network that identifies reciprocal user attributes using a bipartite graph alongside a motif-based GNN for neighbor sampling; a Temporal Behavior Embedding for historical interaction data analysis to reveal unusual user behaviors; a Co-Attention Mechanism that observes and distinguishes abnormal user actions; and a Prediction Layer that assesses whether a specific interaction is malicious. Together, these components create an effective system for detecting malevolent behavior in online dating environments.

Another approach, named LeRuD (LLM-empowered Rumor Detection), uses advanced prompting techniques to analyze credibility signals such as source trustworthiness and factual consistency [95]. By combining zero-shot classification with weak supervision, the model identifies misinformation without relying on labeled datasets. Task-agnostic mapping ensures alignment with predefined categories, while dataset augmentation with synthetic examples improves robustness. This method surpasses state-of- the-art benchmarks, providing a scalable and interpretable solution. The experimental environment includes Twitter15, Twitter16 and Weibo datasets, with extensive filtering to remove ethics-related and data leakage risks. Results show that LeRuD outperforms state-of-the-art rumor detection models by 3.2% to 7.7%, demonstrating strong zero-shot detection capabilities without requiring training data. A hybrid approach has also been developed to identify fake and LLM-generated profiles on the LinkedIn platform during registration, prior to users forming connections with others [96]. This method, named Section and Subsection Tag Embedding (SSTE), emphasizes the consistency of textual and metadata

information in profiles submitted during registration and utilizes sophisticated word embeddings from models such as BERT and RoBERTa. The experimental environment utilized textual data from 3600 LinkedIn profiles (1800 legitimate, 600 manually-created fake, and 1200 ChatGPT- generated). The procedure involved preprocessing, applying word embeddings (GloVe, Flair, BERT, RoBERTa), and incorporating tag embeddings to enhance discriminative features. Classifiers (LR, RF, SVM) were trained and tested on multiple scenarios. Results showed up to 96.3% accuracy (RoBERTa) for fake profile detection and 70–90% accuracy for LLM-generated profiles, demonstrating strong early-stage detection performance without dynamic user activity data.

CONCLUSION AND FUTURE OUTLOOK

This paper has traced the evolution of fake news detection techniques from early human-driven efforts to the latest AI-driven models. We provided a dual categorization of approaches into manual fact-checking methods versus automatic detection algorithms. Within these, we identified four complementary methodological perspectives: knowledge-based (making use of external facts and databases), style-based (analyzing linguistic writing cues), propagation-based (modeling how fake news spreads in networks), and source-based (assessing the credibility of news sources).

By structuring the landscape along these axes, our study highlights how each approach contributes to a holistic defense against disinformation. Notably, we also emphasize the dual role of modern large language models: they are not only powerful tools for detecting fake news but can they generate deceptive content, underscoring the nuanced challenge they present. This dual-use nature necessitates vigilant development of detection strategies that keep pace with generative capabilities.

Historically, automatic detection began with traditional machine learning classifiers like SVM, Naive Bayes, Logistic Regression, Random Forest, operating on handcrafted features and content cues. These classical models demonstrated that data-driven techniques could outperform manual scrutiny at scale, especially on structured datasets like ISOT and LIAR. However, the field rapidly progressed into deep learning as researchers harnessed neural networks to capture complex patterns in text. Recur- rent models, particularly LSTMs and BiLSTMs, improved the understanding of sequential context in news, while CNNs excelled at extracting local textual patterns. Hybrid architectures combining these strengths, for instance, a CNN with a Bidirectional LSTM and Self-Attention (CNN–BiLSTM–SA), achieved remarkable accuracy gains. One recent study reported a Self-Attention BiLSTM model

attaining 99.98% classification accuracy on the large ISOT news dataset, effectively outperforming earlier RNN and CNN baselines.

Such results illustrate that deep learning models, through contextual feature learning, can nearly saturate certain benchmarks. To broaden evaluation, researchers have validated approaches across a spectrum of datasets: from news article collections like ISOT to short claim datasets like LIAR, and from multimodal social media corpora such as Fakeddit and the MediaEval challenge to expanded social news sets like GossipCop++ and even cross-lingual corpora. The consistent pattern is that advanced models significantly improve detection performance across these diverse benchmarks. For example, the multimodal transformer-based framework integrating text and image achieved about 94.4% accuracy on the MediaEval-2016 fake news challenge dataset, surpassing prior methods by over 3%. Likewise, style-robust models have emerged to counter adversarial attempts: the 2024 SheepDog detector, which is agnostic to writing style, withstood intentionally rephrased fake news generated by LLMs and delivered up to 93% F1.

These improvements point out how the evolution from basic classifiers to deep neural networks and the inclusion of images, social context, and style invariance is still boosting accuracy and robustness in fake news identification.

The advent of transformer-based models and Large Language Models has pushed fake news detection into a new era. Pretrained transformers like BERT and RoBERTa brought a leap in language understanding and fine-tuning them for fake news tasks quickly became a standard. Building on this, recent efforts make use of the unprecedented linguistic competence of state-of-the-art LLMs (such as OpenAI's ChatGPT, Meta's LLaMA 3, and Google's Gemini) for both detecting and analyzing misinformation. Our review highlighted several key innovations introduced by LLM-based techniques. First, researchers have explored zero-shot and few-shot

detection, where LLMs can classify news credibility with little to no task-specific training by utilizing prompt-based knowledge. For instance, the PASTEL framework uses prompted LLMs to generate weak labels for various credibility signals and applies a zero-shot prompt for veracity classification. This approach, combining zero-shot reasoning with weak supervision, was shown to match or outperform supervised models on multiple datasets, all without requiring ground-truth labels for training. Such results suggest LLMs can generalize detection skills from their vast pretrained knowledge, opening avenues for low-resource and rapid deployment scenarios.

Second, LLMs have been employed to generate synthetic data and adversarial examples that improve model training. The MegaFake study constructed a large-scale dataset of fake news articles written by LLMs and revealed that detectors trained on a mix of human and AI-generated news became more robust, whereas over-reliance on AI-generated training data can introduce biases, high- lighting the need for balance. Similarly, the AdStyle approach used GPT-3.5 to automatically rephrase news content in diverse writing styles (humorous, sarcastic, poetic, etc.) as a form of adversarial augmentation. This not only tested detectors against style-transferred fake news but also improved their resilience; the augmented model maintained high performance even when confronted with stylized attacks.

Third, LLM-based methods have advanced the explain ability of fake news detection. Techniques like HiSS (Hierarchical Step-by-Step) prompting guide large models to break down claims and verify each part, yielding more transparent reasoning. HiSS was shown to exceed the accuracy of other state- of-the-art methods on challenging datasets while producing fine-grained, explainable rationales for its decisions, effectively addressing issues like omitted evidence or AI “hallucinations” in the fact-checking process. This marks a significant step toward detectors that not only output a verdict but also justify their judgment in human-understandable terms. Additionally, multimodal LLM-driven frameworks are emerging, for example, the FakeNewsGPT4 system augments a vision-language model with tailored knowledge modules to analyze visual artifacts and textual context jointly. By injecting semantic world knowledge and visual anomaly detection into a GPT-4-based architecture, this approach achieved superior results (e.g., 89.6% AUC) under cross-domain evaluation, outperforming prior multi-modal fake news detectors.

These innovations demonstrate the versatility of LLMs, from enabling more adaptive learning strategies (like few-shot prompts and synthetic data generation), to supporting interpretability and multimodal reasoning, all contributing to more robust fake news identification systems.

Fake news detection techniques have become increasingly sophisticated, transitioning from labor- intensive human fact-checking to automated machine learning classifiers, to deep neural networks, and now to powerful LLM-based systems. Each wave of innovation has broadened the capabilities of detectors: improving accuracy, generalization, and the ability to handle complex, real-world misinformation. Yet, our survey also makes it clear that the battle against disinformation is far from over. Current state-of-the-art models still face limitations, such as a scarcity of comprehensive, high-quality datasets (especially for low-resource languages and multimodal content) and the continual evolution of fake news tactics.

Going forward, the research community should prioritize hybrid approaches that combine the strengths of multiple methods to capture the multifaceted nature of fake news.

There is also a need to improve the efficiency and scalability of detection: large transformers and LLMs offer excellent performance but are computationally intensive; future work may explore model compression, distillation, or federated techniques to deploy these models more widely without sacrificing accuracy.

Another important direction is deeper multimodal analysis. Fake news is no longer confined to articles or posts. It appears as images, deepfakes, and short videos on platforms like TikTok and YouTube. Indeed, misinformation on short form video platforms contains heterogeneous content and as it increasingly includes images, video and audio, detectors must jointly analyze content across these modalities, a trend already started with multimodal datasets (like Fakeddit) and vision-language models.

Ensuring transparency and trust in AI-driven detectors is another important issue. Adopting explainable AI (XAI) techniques, as seen in the use of reasoning traces (HiSS) or rationale-guided frameworks, will help users and domain experts understand and corroborate the system’s findings.

Finally, continued interdisciplinary collaboration will be increasingly important: combining insights from journalism, social science, psychology, and computer science can enrich fake news detection strategies, leading to more holistic solutions.

By pursuing these future directions, we can develop next-generation fake news detectors that are not only highly accurate and robust, but also faster, more inclusive of different content forms, and transparent in their decision making.

Such advancements will be necessary for keeping pace with the ever-evolving threat of disinformation and safeguarding public trust in the information ecosystem.

REFERENCES

[1] B. Omar, O.D. Apuke, Z.M. Nor, The intrinsic and extrinsic factors predicting fake news sharing among social media users: the moderating role of fake news awareness, Curr Psychol, 43(2024), 1235–1247, doi:10.1007/s12144-023-04343-4.

[2] Q. Huang, S. Lei, B. Ni, Perceived Information Overload and Unverified Information Sharing on WeChat Amid the COVID-19 Pandemic: A Moderated Mediation Model of Anxiety and Perceived Herd, Front. Psychol., 13(2022), 837820, doi:10.3389/fpsyg.2022.837820.

[3] F. Olan, U. Jayawickrama, E.O. Arakpogun, Fake News on Social Media: The Impact on Society, Inf Syst Front, 26(2024), 443–458, doi:10.1007/s10796-022-10242-z.

[4] L. Enria, H. Dwyer, M. Marchant, N. Beckmann, M. Schmidt-Sane, A. Conteh, Political dimensions of misinformation, trust, and vaccine confidence in a digital age, BMJ, 385(2024), doi:10.1136/bmj-2024-079940.

[5] E. Rahmanian, Fake news: a classification proposal and a future research agenda, Spanish Journal of Marketing - ESIC, 27(2023), 60–78, doi:10.1108/SJME-09-2021-0170.

[6] DeepMind, Mapping the misuse of generative AI, Google (2024), Accessed: 14.11.2024, https: //deepmind.google/discover/blog/mapping-the-misuse-of-generative-ai/.

[7] J. Li, X. Chang, Combating Misinformation by Sharing the Truth: A Study on the Spread of Fact-Checks on Social Media, Inf Syst Front, 25(2023), 1479–1493, doi:10.1007/s10796-022-10296- z.

[8] S. Omar, J.P. Van Belle, Disaster Misinformation Management: Strategies for Mitigating the Effects of Fake News on Emergency Response, Lecture Notes in Networks and Systems, 932(2024), doi:10.1007/978-3-031- 54235-0 29.

[9] Poynter, China and Russia spreading hurricane misinformation: Report, Fact-checking report (2024), Accessed: 14.11.2024, https://www.poynter.org/fact-checking/2024/china-russi a-spreading-hurricane-misinformation/.

[10] Y.H. Liu, C.Y. Kuo, SiMAIM: identifying sockpuppets and puppetmasters on a single forum- oriented social media site, J Supercomput, 79(2023), 18667–18698, doi:10.1007/s11227-023-05376- z.

[11] M. Tajrian, A. Rahman, M.A. Kabir, M.R. Islam, A Review of Methodologies for Fake News Analysis, IEEE Access, 11(2023), 73879–73893, doi:10.1109/ACCESS.2023.3294989.

[12] H. Thakar, B. Bhatt, Fake news detection: recent trends and challenges, Soc. Netw. Anal. Min., 14(2024), 176, doi:10.1007/s13278-024-01344-4.

[13] J. Ludwig, J. Sommer, Mindsets and politically motivated reasoning about fake news, Motiv Emot, 48(2024), 249–263, doi:10.1007/s11031-024-10067-0.

[14] S. Mun˜oz, C.A´. Iglesias, Exploiting Content Characteristics for Explainable Detection of Fake News, Big Data Cogn. Comput, 8(2024), 129, doi:10.3390/bdcc8100129.

[15] A. Shrestha, F. Spezzano, Textual Characteristics of News Title and Body to Detect Fake News: A Reproducibility Study, Lecture Notes in Computer Science, 12657(2021), 572–581, doi: 10.1007/978-3-030- 72240-1 9.

[16] Y. Sun, J. He, L. Cui, S. Lei, C.T. Lu, Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges, arXiv preprint, arXiv:2403.18249(2024), doi:10.48550/arXiv.2403.18249.

[17] E. Papageorgiou, C. Chronis, I. Varlamis, Y. Himeur, A Survey on the Use of Large Language Models (LLMs) in Fake News, Future Internet, 16(2024), 298, doi:10.3390/fi16080298.

[18] S. Raponi, Z. Khalifa, G. Oligeri, R. Di Pietro, Fake news propagation: A review of epidemic models, datasets, and insights, ACM Transactions on the Web (TWEB), 16(2022), 1–34, doi:10.1145/3522756.
[19] S. Yin, P. Zhu, L. Wu, C. Gao, Z. Wang, GAMC: An Unsupervised Method for Fake News Detection Using Graph Autoencoder with Masking, Proceedings of the AAAI Conference on Artificial Intelligence, 38(2024), 347–355, doi:10.1609/aaai.v38i1.27788.
[20] M.F. Abrar, M.S. Khan, I. Khan, M. ElAffendi, S. Ahmad, Towards Fake News Detection: A Multivocal Literature Review of Credibility Factors in Online News Stories and Analysis Using Analytical Hierarchical Process, Electronics, 12(2023), 3280, doi:10.3390/electronics12153280.
[21] Millet, B., Tang, J., Seelig, M., Petit, J., Sun, R., In Twitter we trust(ed): How perceptions of Twitter's helpfulness influence news post credibility perceptions and news engagement, Computers in Human Behavior, 155(2024), 108185, https://doi.org/10.1016/j.chb.2024.108185.
[22] Pragadeeswaran, S., Janarthanan, K., Anand, K., Yugapriya, S., Gowshika, S., A Review of Methodologies for Fake News Analysis, International Journal of Innovative Research in Computer and Communication Engineering, 12(4)(2024), DOI: 10.15680/IJIRCCE.2024.1204119, http:// ijircce.com/admin/main/storage/app/pdf/VIdUZM89ZaSmSrWaJBJGbFESzWBnCpPvcWO35Hqq. pdf.
[23] Alsmadi, I., Alazzam, I., Al-Ramahi, M., Zarour, M., Stance Detection in the Context of Fake News—A New Approach, Future Internet, 16(10):364(2024), https://doi.org/10.3390/fi16 100364.
[24] Mishra, A., Sadia, H., A Comprehensive Analysis of Fake News Detection Models: A Systematic Literature Review and Current Challenges, Engineering Proceedings, 59(1):28(2023), https: //doi.org/10.3390/engproc2023059028.
[25] Gong, S., Sinnott, R.O., Qi, J., Paris, C., Fake News Detection Through Temporally Evolving User Interactions, Springer, Lecture Notes in Computer Science, 13938(2023), https://doi.or g/10.1007/978-3-031-33383-5_11.
[26] Bachelot, M., Lyubareva, I., A. Epalle, T., et al., French fake news propagation: multi-level assessment and classification, Soc. Netw. Anal. Min., 14(2024), 156, https://doi.org/10.100 7/s13278-024-01319-5.
[27] Parmar, S., Rahul, Fake news detection via graph-based Markov chains, Int. j. inf. tecnol., 16(2024), 1333–1345, https://doi.org/10.1007/s41870-023-01558-3.
[28] Bobkowski, P., Younger, K., News Credibility: Adapting and Testing a Source Evaluation Assessment in Journalism, College & Research Libraries, 81(5)(2020), 822, https://doi.org/10.5 860/crl.81.5.822.
[29] Bazmi, P., Asadpour, M., Shakery, A., Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility, Information Processing & Management, 60(1)(2023), 103146, https://doi.org/10.1016/j.ipm.2022.103146.
[30] Duke Reporters' Lab, Fact-checking, (2024), https://reporterslab.org/fact-checking/, Accessed: 16.11.2024.
[31] Stencel, M., Luther, J., Fact-checkers extend their global reach with 391 outlets, but growth has slowed, Reporters' Lab (2021), https://reporterslab.org/fact-checkers-extend-their-global-reach- with-391-outlets-but-growth-has-slowed/. Accessed: 16 November 2024.
[32] European Digital Media Observatory, Fact-checking organisations in the EU (2024), https://ed mo.eu/resources/repositories/fact-checking-organisations-in-the-eu/, Accessed: 16.11.2024.
[33] Fiskkit, Fiskkit: About us (2024), https://www.fiskkit.com/about, Accessed: 16.11.2024.
[34] Community Notes, Community Notes: How it works (2024), https://communitynotes.x.com/g uide, Accessed: 16.11.2024.
[35] Debunk.org, Debunk.org: About us (2024), https://www.debunk.org/, Accessed: 16.11.2024.
[36] StopFake, StopFake: Fact-checking project (2024), https://www.stopfake.org/, Accessed: 16.11.2024.
[37] Alt News, Alt News: About us (2024), https://www.altnews.in/, Accessed: 16.11.2024.
[38] Snijders, C., Conijn, R., de Fouw, E., Van Berlo, K., Humans and Algorithms Detecting Fake News: Effects of Individual and Contextual Confidence on Trust in Algorithmic Ad- vice, International Journal of Human–Computer Interaction, 39(7)(2022), 1483–1494, https: //doi.org/10.1080/10447318.2022.2097601.
[39] Sudhakar, M., Kaliyamurthie, K. P., Detection of fake news from social media using support vector machine learning algorithms, Measurement: Sensors, 32(2024), 101028, https://doi.or g/10.1016/j.measen.2024.101028.
[40] Lee, D. Y., Liu, Y. Y., Application of Supervised Machine Learning Algorithms for Detection of Fake News

using Support Vector Machine Classifier, CTD International Journal for Media Studies, 2(2)(2024), 1–7, https://ctdjms.com/pdf/volume-2,issue-2,2024/1-Application-of-Supervised-Machine-Learning-Algorithms- for-Detection-of-Fake-News-using-Sup port-Vector-Machine-Classifier.pdf.

[41] Dedeepya, P., Yarrarapu, M., Kumar, P. P., Kaushik, S. K., Raghavendra, P. N., Chandu, P., Fake News Detection on Social Media Through a Hybrid SVM-KNN Approach Leveraging Social Capital Variables, 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, pp. 1168–1175 (2024), doi:10.1109/ICAAIC60222.2024.10575681.

[42] Matemilola, A. S., Aliyu, S., Development of an enhanced naive bayes algorithm for fake news classification, Science World Journal, 19(2)(2024), 512–517, https://doi.org/10.4314/swj.v1 9i2.28.

[43] Qubra, R., Saputra, R. A., Classification of Hoax News Using the Na¨ıve Bayes Method, In- ternational Journal Software Engineering and Computer Science (IJSECS), 4(1)(2024), 40–48, https://doi.org/10.35870/ijsecs.v4i1.2068.

[44] Bhavani, R., Balamanigandan, R., Priscilla, A. A., Analyzing the Performance of Novel Logistic Regression over Linear Regression Algorithms for Predicting Fake Job with Improved Accuracy, 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, pp. 1728–1732 (2024), doi:10.1109/ICESC60852.2024.10690033.

[45]Alguttar, A. A., Shaaban, O. A., Yildirim, R., Optimized Fake News Classification: Leveraging Ensembles Learning and Parameter Tuning in Machine and Deep Learning Methods, Applied Artificial Intelligence, 38(1)(2024), https://doi.org/10.1080/08839514.2024.2385856.

[46] Krishna, N. L. S. R., Adimoolam, M., Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm, IEEE, International Conference on Business Analytics for Technology and Security (ICBATS), pp. 1–6 (2022), doi:10.1109/ICBATS54253.2022.9758999.

[47] Oluwabunmi, A., Oluwaferanmi, R., Abdullai, A., Fake News Detection System Using Logistic Regression, Decision Tree and Random Forest, British Journal of Computer, Networking and Information Technology, 7(2024), 115–121, doi:10.52589/BJCNIT-IOYRPY7G.

[48] Abd El-Mageed, A. A., Abohany, A. A., Ali, A. H., Hosny, K. M., An adaptive hybrid African Vultures–Aquila optimizer with Xgb-Tree algorithm for fake news detection, Journal of Big Data, 11:41 (2024), https://doi.org/10.1186/s40537-024-00895-9.

[49] Singh, I., Dhanda, N., Sahai, A., Gupta, K. K., Comparative Study of Random Forest Algorithm and Logistic Regression in the Analysis of Fake News, 8th International Conference on Communication and Electronics Systems (ICCES), IEEE, pp. 1477–1482 (2023), doi:10.1109/ICCES57224.2023.10192821.

[50] Birunda, S. S., Devi, R. K., Muthukannan, M., An efficient model for detecting COVID fake news using optimal lightweight convolutional random forest, SIViP, 18(2024), 2659–2669, https://doi.org/10.1007/s11760-023- 02938-9.

[51] Cahyono, H. D., Mahadewa, A., Wijayanto, A., Wardani, D. W., Setiadi, H., Fast Na¨ıve Bayes classifiers for COVID-19 news in social networks, Indonesian Journal of Electrical Engineering and Computer Science, 34(2)(2024), 1033–1041, http://dx.doi.org/10.11591/ijeecs.v34.i 2.pp1033-1041.

[52] Singhania, S., Fernandez, N., Rao, S., 3han: A deep neural network for fake news detection, In Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part II 24, Springer International Publishing (2017), pp. 572–581, https://doi.org/10.48550/arXiv.2306.12014.

[53] Bondielli, A., Marcelloni, F., A survey on fake news and rumour detection techniques, Information Sciences, 497(2019), 38–55, https://doi.org/10.1016/j.ins.2019.05.035.

[54] Abualigah, L., Al-Ajlouni, Y. Y., Daoud, M. S., Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe, Soc. Netw. Anal. Min., 14(2024), 40, https://doi.org/10.1007/s13278-024-01198-w.

[55] Wundari, F., Amien, M. N. A. S., Irtsa, D. H., Identifying Fake News Using Long-Short Term Memory Model, Journal of Dinda: Data Science, Information Technology, and Data Analytics, 4(1)(2024), 28–34, https://doi.org/10.20895/dinda.v4i1.1424.

[56] Shalini, A., Saxena, S., Kumar, B., Automatic detection of fake news using recurrent neural network—Long

short-term memory, Journal of Autonomous Intelligence, 7(3)(2023), http://dx .doi.org/10.32629/jai.v7i3.798.

[57] Jian, W., Li, J. P., Akbar, M. A., Haq, A. U., Khan, S., Alotaibi, R. M., Alajlan, S. A., SA- Bi-LSTM: Self Attention With Bi-Directional LSTM-Based Intelligent Model for Accurate Fake News Detection to Ensured Information Integrity on Social Media Platforms, IEEE Access, Vol. 12, pp. 48436–48452 (2024), https://doi.org/10.1109/ACCESS.2024.3382832.

[58] Akhter, M., Hossain, S.M.M., Nigar, R.S., COVID-19 Fake News Detection using Deep Learning Model, Ann. Data. Sci., 11(2024), 2167–2198, https://doi.org/10.1007/s40745-023-00507-y.

[59] Xia, H., Wang, Y., Zhang, J. Z., Zheng, L. J., Kamal, M. M., Arya, V., COVID-19 fake news detection: A hybrid CNN-BiLSTM-AM model, Technological Forecasting and Social Change, 195(2023), 122746, https://doi.org/10.1016/j.techfore.2023.122746.

[60] Lakzaei, B., Haghir Chehreghani, M., Bagheri, A., Disinformation detection using graph neural networks: a survey, Artif Intell Rev, 57(2024), 52, https://doi.org/10.1007/s10462-024-107 02-9.

[61] Visweswaran, M., Mohan, J., Kumar, S. S., Soman, K. P., Synergistic Detection of Multimodal Fake News Leveraging TextGCN and Vision Transformer, Procedia Computer Science, 235(2024), 142–151, https://doi.org/10.1016/j.procs.2024.04.017.

[62] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L- .,Polosukhin, I., Attention is all you need, Advances in Neural Information Processing Systems, 30(2017), 5998–6008, https://doi.org/10.48550/arXiv.1706.03762.

[63] Kaliyar, R. K., Goswami, A., Narang, P., FakeBERT: Fake news detection in social media with a BERT-based deep learning approach, Multimed Tools Appl, 80(2021), 11765–11788, https: //doi.org/10.1007/s11042-020- 10183-2.

[64] Repede, S, .E., Brad, R., A comparison of artificial intelligence models used for fake news detection, Bulletin Of "Carol I" National Defence University, 12(1)(2023), 114–131, https://doi.org/10.53477/2284-9378-23-10.

[65] Shushkevich, E., Alexandrov, M., Cardiff, J., Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data, Inventions, 8(5)(2023), 112, https: //doi.org/10.3390/inventions8050112.

[66] Pillai, A. S., Fake News Multi Class Detection Using Transformers and Gradient Boosting Ensemble, International Research Journal of Modernization in Engineering Technology and Science, 6(2)(2024), 2582– 5208, doi:10.56726/IRJMETS49747.

[67] Zhang, Z., Lv, Q., Jia, X., Yun, W., Miao, G., Mao, Z., Wu, G., GBCA: Graph Convolution Network and BERT combined with Co-Attention for fake news detection, Pattern Recognition Letters, 180(2024), 26–32, https://doi.org/10.1016/j.patrec.2024.02.014.

[68] Al-alshaqi, M., Rawat, D. B., Liu, C., Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning, Sensors, 24(18)(2024), 6062, https://doi.org/10.3390/s24186062.

[69] Hamed, S. K., Ab Aziz, M. J., Yaakub, M. R., Enhanced Feature Representation for Multimodal Fake News Detection Using Localized Fine-Tuning of Improved BERT and VGG-19 Models, Arab J Sci Eng (2024), https://doi.org/10.1007/s13369-024-09354-2.

[70] Luqman, M., Faheem, M., Ramay, W. Y., Saeed, M. K., Ahmad, M. B., Utilizing Ensemble Learning for Detecting Multi-Modal Fake News, IEEE Access, 12(2024), 15037–15049, doi:10.1109/ACCESS.2024.3357661.

[71] Koru, G. K., Uluyol, C¸ ., Detection of Turkish Fake News from Tweets with BERT Models, IEEE Access, 12(2024), 14918–14931, doi:10.1109/ACCESS.2024.3354165.

[72] Deshpande, A., Talukdar, P., Narasimhan, K., When is BERT multilingual? isolating crucial ingredients for cross lingual transfer, arXiv preprint arXiv:2110.14782 (2021), https://doi.or g/10.48550/arXiv.2110.14782.

[73] Rybak, P., Transferring BERT Capabilities from High-Resource to Low-Resource Languages Using Vocabulary Matching, arXiv preprint arXiv:2402.14408 (2024), https://doi.org/10.48550/a rXiv.2402.14408.

[74] Repede, S, .E., Researching disinformation using artificial intelligence techniques: challenges, Bulletin Of "Carol I" National Defence University, 12(2)(2023), 69–85, https://doi.org/10.53477/2284-9378-23-21.

[75] Wang, L. Z., Ma, Y., Gao, R., Guo, B., Zhu, H., Fan, W., Ng, K. C., MegaFake: A Theory-Driven Dataset of Fake News Generated by Large Language Models, arXiv preprint arXiv:2408.11871 (2024), https://doi.org/10.48550/arXiv.2408.11871.

[76] Su, J., Cardie, C., Nakov, P., Adapting fake news detection to the era of large language models, arXiv preprint arXiv: 2311.04917 (2023), https://doi.org/10.48550/arXiv.2311.04917.

[77] Sun, Y., He, J., Lei, S., Cui, L., Lu, C. T., Med-mmhl: A multi-modal dataset for detecting human and llm-generated misinformation in the medical domain, arXiv preprint arXiv:2306.08871 (2023), https://doi.org/10.48550/arXiv.2306.08871.

[78] Jiang, B., Tan, Z., Nirmal, A., Liu, H., Disinformation detection: An evolving challenge in the age of llms, In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), Society for Industrial and Applied Mathematics, pp. 427–435 (2024), https://doi.org/10.113 7/1.9781611978032.50.

[79] Park, S., Han, S., Cha, M., Adversarial Style Augmentation via Large Language Model for Robust Fake News Detection, arXiv preprint arXiv:2406.11260 (2024), https://doi.org/10.48550/arX iv.2406.11260.

[80] Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M. Y., Wang, W. Y., On the risk of misinformation pollution with large language models, arXiv preprint arXiv:2305.13661 (2023), https://doi.or g/10.48550/arXiv.2305.13661.

[81] Chalehchaleh, R., Farahbakhsh, R., Crespi, N., Enhancing Multilingual Fake News Detection through LLM-Based Data Augmentation, In The 13th International Conference on Complex Networks and their Applications (2024), https://hal.science/hal-04733161v1/document.

[82] Dahou, A. H., Cheragui, M. A., Abdedaiem, A., Mathiak, B., Enhancing Model Performance through Translation-based Data Augmentation in the context of Fake News Detection, Procedia Computer Science, 244(2024), 342–352, https://doi.org/10.1016/j.procs.2024.10.208.

[83] Wu, J., Guo, J., Hooi, B., Fake News in Sheep's Clothing: Robust Fake News Detection against LLM-Empowered Style Attacks, In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3367–3378 (2024), https://doi.org/10.1145/3637528. 3671977.

[84] Boissonneault, D., Hensen, E., Fake News Detection with Large Language Models on the LIAR Dataset, Researchsquare (2024), https://doi.org/10.21203/rs.3.rs-4465815/v1.

[85] Liu, X., Li, P., Huang, H., Li, Z., Cui, X., Liang, J., He, Z., FakeNewsGPT4: Advancing Multimodal Fake News Detection through Knowledge-Augmented LVLMs, arXiv preprint arXiv:2403.01988 (2024), https://doi.org/10.48550/arXiv.2403.01988.

[86] Repede, S, .E., Brad, R., LLaMA 3 vs. State-of-the-Art Large Language Models: Performance in Detecting Nuanced Fake News, Computers, 13(11)(2024), 292, https://doi.org/10.3390/comp uters13110292.

[87] Ibrahim, M., Fine-Grained Language-Based Reliability Detection in Spanish News with Fine- Tuned Llama Model, In Proceedings of the Iberian Languages Evaluation Forum (IberLEF), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN), CEUR-WS.org (2024), https://ceur-ws.org/Vol-3756/FLARES2024_paper4.pdf.

[88] Zhang, X., Gao, W., Reinforcement Retrieval Leveraging Fine-grained Feedback for Fact Checking News Claims with Black-Box LLM, arXiv preprint arXiv:2404.17283 (2024), https://doi.org/ 10.48550/arXiv.2404.17283.

[89] Zhang, X., Gao, W., Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by Step Prompting Method, arXiv preprint arXiv:2310.00305 (2023), https://doi.org/10.48550/arXiv.2310.00305.

[90] Kareem, W., Abbas, N., Fighting Lies with Intelligence: Using Large Language Models and Chain of Thoughts Technique to Combat Fake News, Artificial Intelligence XL, SGAI, Lecture Notes in Computer Science, 14381, Springer, Cham (2023), https://doi.org/10.1007/978-3-031-479 94-6_24.

[91] Leite, J. A., Razuvayevskaya, O., Bontcheva, K., Scarton, C., Detecting misinformation with LLM-predicted credibility signals and weak supervision, arXiv preprint arXiv:2309.07601 (2023), https://doi.org/10.48550/arXiv.2309.07601.

[92] Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., Qi, P., Bad actor, good advisor: Exploring the role of large language models in fake news detection, Proceedings of the AAAI Conference on Artificial Intelligence, 38(20)(2024), 22105–22113, https://doi.org/10.1609/aaai.v38i20.30 214.

[93] Wan, H., Feng, S., Tan, Z., Wang, H., Tsvetkov, Y., Luo, M., Dell: Generating reactions and explanations for

LLM-based misinformation detection, arXiv preprint arXiv:2402.10426 (2024), https://doi.org/10.48550/arXiv.2402.10426.

[94] Chen, K., Wang, Z., Liu, K., Zhang, X., Luo, L., MedGraph: Malicious Edge Detection in Temporal Reciprocal Graph via Multi-Head Attention-Based GNN, Neural Computing and Applications, 35(12)(2023), 8919–8935, https://doi.org/10.1007/s00521-022-08065-9.

[95] Liu, Q., Tao, X., Wu, J., Wu, S., Wang, L., Can Large Language Models Detect Rumors on Social Media?, arXiv preprint arXiv:2402.03916 (2024), https://doi.org/10.48550/arXiv.240 2.03916.

[96] Ayoobi, N., Shahriar, S., Mukherjee, A., The Looming Threat of Fake and LLM-Generated LinkedIn Profiles: Challenges and Opportunities for Detection and Prevention, In Proceedings of the 34th ACM Conference on Hypertext and Social Media, pp. 1–10 (2023), https: //doi.org/10.1145/3603163.3609064.