Review on Fake News Detection Using Natural Language Processing

Nandani Sahu[1], Shanu Gour[2]
[1] *M. Tech Scholar, Computer Science & Engineering, Bharti Vishwavidyalaya, Durg, Chhattisgarh, India*
[2] *Assistant Professor, Computer Science & Engineering, Bharti Vishwavidyalaya, Durg, Chhattisgarh, India*

Abstract – The unprecedented growth of digital media and social networking platforms has accelerated the spread of misinformation, posing significant threats to societal trust, public health, political stability, and democratic processes. Traditional manual fact-checking mechanisms are insufficient to counter the volume, speed, and sophistication of fake news circulation. This study presents a comprehensive analysis of Natural Language Processing (NLP)-based approaches for automated fake news detection, synthesizing contemporary machine-learning, deep-learning, and hybrid multimodal methods. Through a detailed examination of linguistic patterns, transformer-based architectures, stance analysis, knowledge-augmented models, and graph-based frameworks, the research identifies key features and performance parameters critical to effective misinformation detection. Findings highlight that hybrid systems integrating textual semantics, contextual metadata, social-behavioral signals, and adversarial defense strategies outperform text-only models. The study further emphasizes the importance of multilingual datasets, cross-domain adaptability, explainability, and ethical AI deployment. This research contributes to strengthening digital information integrity and provides a foundation for developing scalable, transparent, and robust NLP-driven fake news detection systems.

Index Terms - Fake News Detection, Natural Language Processing (NLP), Machine Learning, Deep Learning

I. INTRODUCTION

The exponential growth of digital media and social networking platforms has revolutionized the way information is generated, consumed, and shared. While this digital transformation has democratized access to information, it has also facilitated the rapid spread of misinformation and fake news. Fake news refers to fabricated, misleading, or manipulated content designed to deceive readers or influence public opinion. In recent years, the proliferation of fake news has emerged as a serious societal challenge, influencing political outcomes, harming public health, triggering social unrest, and creating distrust in legitimate information sources. The COVID-19 pandemic, political elections across various countries, and communal incidents have demonstrated how rapidly misinformation can spread, driving real-world consequences. Traditional methods of detecting fake news—such as manual verification by journalists or fact-checking organizations—are time-consuming, labor-intensive, and not scalable to the vast volume of online content produced every second. Consequently, automated fake news detection has become a significant research area. With advancements in computational linguistics and artificial intelligence, Natural Language Processing (NLP) has emerged as a promising solution to address misinformation challenges. NLP-based models can analyze textual patterns, linguistic cues, semantic meanings, and context to classify news content as real or fake. Machine learning and deep learning techniques, such as Support Vector Machines (SVM), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and transformers like BERT, have significantly improved accuracy in detecting deceptive and biased content. Despite substantial progress, fake news detection remains a complex problem due to linguistic ambiguity, evolving writing patterns, domain dependency, and the adversarial nature of misinformation creators. There is a growing need for robust, context-aware, multilingual, and reliable detection systems that can adapt to dynamic digital environments. This research aims to explore and

evaluate advanced NLP techniques for fake news detection, identify key textual features contributing to misinformation classification, and propose a model capable of enhancing detection accuracy across diverse datasets. The study further highlights ethical considerations and encourages responsible deployment of AI tools to maintain information integrity in digital ecosystems.

II. LITERATURE SURVEY

Shu et al. (2017) examined fake news from a computational social science perspective, emphasizing that misinformation is deliberately created to manipulate users and therefore often mimics legitimate linguistic patterns. They highlighted limitations of traditional feature-based text models and argued that fake news detection requires integration of content cues, social context, user profiling, and propagation behavior. Their analysis established that deceptive information spreads faster and more broadly than genuine news due to emotional and sensational framing. The authors developed a conceptual model based on psychological theories such as confirmation bias and social influence, identifying challenges like early detection, evolving language, and adversarial behavior. The study laid the foundation for multimodal and graph-based approaches by emphasizing hybrid learning techniques that combine NLP features, social network analytics, and contextual metadata for more robust and scalable fake-news classification systems.

Wang et al. (2017) introduced the LIAR dataset, a benchmark comprising thousands of fact-checked political claims labeled across multiple veracity categories ranging from “true” to “pants-on-fire.” Their research demonstrated that simple textual features alone were insufficient to capture complex deception patterns, particularly in short statements typical of political discourse. The authors compared logistic regression, SVM, and neural architectures, revealing moderate but inconsistent performance due to linguistic ambiguity and subtle rhetorical manipulation. Importantly, the study highlighted the value of metadata such as speaker identity, party affiliation, and context, proposing hybrid architectures that combine textual and contextual signals. The LIAR dataset remains a significant milestone because it encourages fine-grained classification, supports generalization tests, and continues to serve as a benchmark for transformer-based models and stance-aware veracity frameworks in contemporary fake-news research.

Rashkin et al. (2017) explored linguistic and stylistic differences among four categories of deceptive text: hoaxes, propaganda, satire, and real news. They built a large-scale dataset using fact-checking websites and conducted linguistic analysis to identify markers such as hyperbolic sentiment, modal verbs, hedging, and emotional intensity. Their experiments illustrated that satire poses unique challenges because it intentionally mimics news structure while exaggerating content for humor and criticism. The authors used neural language models and found that stylistic cues contribute significantly to distinguishing deceptive content but may break down when adversaries adjust writing patterns. Their work motivated researchers to incorporate pragmatic signals, semantic frames, and discourse-level interpretation into NLP systems, shifting fake-news modeling from surface-level word patterns toward more context-rich and genre-sensitive classification.

Ruchansky et al. (2017) proposed the CSI model—Capture, Score, Integrate—representing one of the earliest hybrid deep-learning architectures for fake-news detection. The model combined textual analysis, temporal propagation patterns, and user-credibility scoring, demonstrating that misinformation often spreads in coordinated clusters influenced by unreliable accounts. Their results showed significant improvements over content-only baselines, proving that network-behavioral signals play a crucial role in disinformation identification. By integrating recurrent neural networks for user-interaction sequences and embedding-based text analysis, the authors demonstrated how structural and social media dynamics provide predictive value. The study also introduced early discussions on adversarial manipulation of user networks and the limitations of relying solely on linguistic cues. CSI became a reference model inspiring future work in graph neural networks, temporal modeling, and large-scale social-context integration for fake-news detection.

Thorne et al. (2018) introduced the FEVER benchmark (Fact Extraction and VERification), which linked claims to evidence retrieved from Wikipedia. Their approach required systems to not only classify claims as supported, refuted, or unverified but also supply textual evidence supporting the decision. This marked a paradigm shift from pure classification toward explainable verification pipelines involving retrieval, sentence matching, and natural language inference. Baseline models revealed challenges in retrieving relevant evidence and interpreting factual consistency, highlighting limitations of purely end-to-end deep learning approaches. FEVER inspired advancements in retrieval-augmented transformers, rationality extraction, and explainable AI frameworks for misinformation detection. The benchmark remains central in research bridging fake-news classification with automated fact-checking, emphasizing transparency and traceability in credibility judgments.

Zellers et al. (2019) developed GROVER, a transformer-based model capable of both generating and detecting synthetic fake-news articles. They demonstrated that state-of-the-art language models can produce highly convincing deceptive narratives conditioned on metadata such as topics and outlets. Interestingly, GROVER was also most effective in detecting its own generated text, emphasizing adversarial dynamics between text generation and detection. Their study revealed weaknesses in conventional fake-news systems, as stylistic and lexical cues become unreliable when adversaries employ powerful neural generators. This work triggered new research directions involving watermarking, adversarial training, and provenance-based detection. GROVER remains foundational in understanding the arms race between generative AI misinformation and automated defense mechanisms.

Shahi and Nandini (2020) created the FakeCovid dataset, compiling multilingual misinformation narratives related to the COVID-19 pandemic. The dataset contained manually verified claims categorized across misinformation types such as medical myths, conspiracy theories, and fabricated policy announcements. Their evaluation of traditional machine-learning and transformer models revealed the difficulty of detecting evolving misinformation, especially in non-English languages. The study emphasized the need for cross-lingual transfer learning, culturally adaptive preprocessing, and domain-adaptive fine-tuning to handle shifting global narratives. By documenting misinformation themes and regional variations, the authors underscored the importance of multilingual resources and contextual awareness in health communication crises, setting a precedent for future research in pandemic-related infodemic detection.

Kaliyar et al. (2021) introduced FakeBERT, a hybrid deep-learning architecture combining BERT embeddings with convolutional neural networks to capture contextual and local n-gram-level features simultaneously. Their experiments on benchmark datasets demonstrated superior accuracy and stability compared to standard BERT, CNN, and LSTM models, especially in distinguishing nuanced deceptive content. The study highlighted that pure transformer encoders may overlook localized deception cues and benefit from convolutional layers that preserve structural patterns. Moreover, FakeBERT addressed computational efficiency concerns, offering a lightweight yet effective architecture suitable for real-time detection environments. Their results sparked renewed interest in hybrid neural architectures blending transformer-based semantics with feature-level granularity to enhance robustness in fake-news classification.

Kondamudi et al. (2023) presented a comprehensive survey of fake-news detection methodologies across social media, focusing on multimodal, graph-based, and adversarial-resilient architectures. They categorized detection methods into content-based, propagation-based, source-credibility-based, and hybrid systems, highlighting the limitations of single-signal approaches. The authors stressed challenges such as temporal drift, dataset bias, misinformation campaigns executed by bots and coordinated groups, and the need for explainable AI. Their work proposed future directions including multimodal fusion, early detection, and human-AI collaboration in fact-checking. This survey remains a key reference for understanding evolving threat landscapes and guiding implementation of scalable and ethical misinformation defense systems.

Zhou et al. (2020) presented an extensive survey on computational approaches for detecting fake news, examining psychological foundations, linguistic signals, knowledge graphs, and deep-learning frameworks. Their research emphasized that fake news spreads more rapidly due to emotional and sensational framing designed to exploit cognitive biases. They categorized fake news detection into four major approaches: style-based, knowledge-based, propagation-based, and hybrid models. The authors highlighted limitations in early content-based models and advocated for integrating contextual metadata, user behavior, and credibility cues for improved detection. They also stressed the need for explainable artificial intelligence, arguing that transparency increases trust in automated systems. Additionally, they discussed challenges including low-resource language support, adversarial attacks, and generalization issues across domains. Their work provided a strategic roadmap for future research on scalable and ethical misinformation detection solutions.

Nakamura et al. (2021) investigated cross-lingual fake news detection using transformer-based language models, focusing on multilingual misinformation dynamics. They explored the ability of multilingual BERT and XLM-RoBERTa to detect fake news across English, Spanish, and Japanese datasets, demonstrating that multilingual training improves generalization but suffers performance degradation when handling culturally specific misinformation. Their work emphasized the role of semantic similarity, shared linguistic structures, and cross-lingual embeddings in building universal deception detectors. The authors highlighted the challenges of resource scarcity in certain languages and suggested leveraging transfer learning and domain adaptation to address multilingual misinformation. They also identified emerging issues, such as cultural bias and regional slang, which complicate NLP tasks. Their study contributed significantly to global misinformation mitigation strategies and highlighted the importance of culturally adaptive detection systems.

Dong et al. (2022) proposed a graph-based neural network approach to detect fake news by modeling user interactions, social influence, and content features simultaneously. Their model integrated Graph Convolutional Networks (GCNs) and attention mechanisms to capture semantic relationships and propagation patterns within social media ecosystems. The study demonstrated that misinformation often originates from clusters of coordinated accounts and spreads through tight community networks. Their experiments showed improved accuracy and robustness compared to pure text-based baselines, especially in noisy online environments. They stressed the significance of temporal features and dynamic graph updates to reflect evolving user behavior. Moreover, the authors emphasized adversarial defense capabilities, demonstrating resilience against manipulation attempts by malicious actors. This research further solidified the value of social context and graph analytics in modern fake-news detection frameworks.

Hossain et al. (2022) explored fake news detection using transformer architectures combined with context-aware embedding techniques to capture implicit semantic cues. Their model incorporated discourse features, sentiment polarity, and linguistic complexity metrics alongside BERT embeddings. Experimental results highlighted improved accuracy in detecting politically motivated misinformation and conspiracy narratives. They also evaluated the impact of emotion-driven content and concluded that fake news often employs exaggerated emotion markers, persuasive rhetoric, and polarized sentiment. The authors recommended integrating human cognitive patterns, such as argument structure and narrative framing, to enhance model interpretability. Additionally, their study emphasized model explainability, arguing that black-box systems limit adoption in critical decision-making environments. Their work demonstrated that blending linguistic intelligence with modern NLP architectures enhances reliability and interpretability in automated misinformation detection systems.

Bhatia et al. (2023) focused on multimodal fake news detection, combining textual semantics, visual features, and social metadata. Their approach utilized transformer-based text models, convolutional neural networks for image processing, and graph representations for user-network analysis. The study revealed that misinformation often uses manipulated images and emotionally charged captions to reinforce

deceptive narratives. Their multimodal fusion strategy significantly improved classification accuracy, particularly for politically biased and health-related misinformation. They also highlighted challenges such as fake image synthesis, deepfake content, and the rising sophistication of adversarial actors. The authors emphasized model scalability and the need for robust real-time deployment frameworks capable of handling high-volume social feeds. Their research contributed to the advancement of hybrid systems capable of capturing complex deception patterns beyond simple textual analysis.

Volkova et al. (2017) analyzed linguistic and psychological attributes distinguishing factual content from manipulative text on social platforms. They explored sentiment polarity, emotional intensity, pronoun usage, modality, and toxicity to identify deception tendencies. Their experiments showed that fake news typically demonstrates polarized sentiment, exaggerated emotions, and high use of subjective language, whereas credible news maintains balanced tone and factual density. The authors incorporated psycholinguistic lexicons, LIWC features, and neural classifiers, revealing that combining emotional and lexical cues yields stronger classification accuracy than surface text alone. Their work emphasized early detection and demonstrated that linguistic deception markers generalize across domains when integrated with machine-learning algorithms. They also recommended fusing behavioral cues, user credibility scoring, and engagement metadata for holistic detection. This research became foundational for affect-aware NLP models in misinformation environments and inspired hybrid psychological-language detection paradigms.

Pérez-Rosas et al. (2018) investigated linguistic feature-based approaches for fake-news classification, creating a curated dataset of manually verified fake and legitimate news articles. They extracted lexical, syntactic, and readability-based features to evaluate differences in narrative style, structural complexity, and emotional patterns. Their analysis showed that deceptive articles tend to contain simpler sentence structures, higher sentiment volatility, persuasion vocabulary, and a more conversational tone. Using SVMs, logistic regression, and deep neural networks, the authors illustrated that handcrafted linguistic features still offer competitive performance, especially in low-resource contexts without large-scale training data. The study also emphasized dataset balance, annotation reliability, and narrative structure analysis. Their findings highlighted the complementarity between classical NLP techniques and modern deep models, demonstrating that linguistic explainability enables better transparency and credibility in automated fake-news systems.

Popat et al. (2018) presented a credibility-assessment framework that combined textual evidence, source reliability, and citation context for misinformation detection. Their approach integrated attention mechanisms to assess claim-evidence alignment, relying on supporting documents from credible repositories. Results showed improved detection accuracy when systems actively verify claims through cross-checking external sources rather than solely analyzing textual style. The authors argued that veracity signals exist beyond phrasing patterns and must reflect factual grounding. Their model achieved transparency by highlighting evidence passages influencing classification decisions. Additionally, they emphasized the importance of scalable evidence retrieval pipelines, context filtering, and knowledge-graph support for automated fact-checking. The study influenced retrieval-augmented transformer architectures and motivated the shift toward explainable fact-verification models in misinformation research.

Qian et al. (2021) developed a knowledge-enhanced fake news detection system that integrates external factual information into transformer models. The authors used structured knowledge-bases and entity-linking to compare claims with verified information, aligning content representations with real-world facts. Their results demonstrated significant performance improvements, particularly in politically sensitive and domain-specific misinformation. The model incorporated semantic similarity, attention-based entity relevance scoring, and context-aware verification modules. They highlighted the need for grounding neural predictions in factual databases to counter stylistically sophisticated yet false narratives. Their work demonstrated that transformers enriched with knowledge-graph reasoning outperform text-only baselines on interpretability and stability. They also

proposed a pipeline for real-time fact-checking workflows, emphasizing the future integration of online knowledge-retrieval engines for scalable misinformation defense.

Zhang et al. (2021) introduced a hierarchical transformer framework for rumor and fake-news classification across multi-turn discussions on social networks. The model captured contextual information from conversation threads, analyzing stance patterns, reply sequences, and discourse flow. Their findings revealed that user stance cues—support, deny, question—hold strong predictive value in misinformation scenarios. The hierarchical design learned both article-level semantics and conversational interaction structures, outperforming flat sequence-based models on benchmark rumor datasets. The authors highlighted challenges like sarcasm, misinformation cascades, and echo-chambers that distort linguistic signals. They also emphasized explainability and interpretable stance-attention mechanisms, suggesting future work integrate conversational transformers with emotion-tracking and social-graph-based cues for improved reliability.

Monti et al. (2019) proposed a graph convolutional network (GCN) approach for fake-news detection on social platforms. Their model represented posts, users, and interactions as nodes within a heterogeneous graph, allowing message-passing to capture relational patterns and spread behavior. Results indicated that misinformation often originates and propagates through dense clusters of like-minded users, making structural network analysis essential. The GCN approach significantly outperformed text-based baselines in early-detection tasks, even when textual signals were limited. They also explored temporal propagation patterns and adversarial robustness. Their study highlighted the strength of graph learning in tackling complex misinformation ecosystems, providing a blueprint for graph-transformer hybrids in current research.

Huang et al. (2022) introduced a transformer-based adversarial training framework for robust fake-news detection against synthetic and style-shift attacks. Their method simulated adversarial perturbations, paraphrasing manipulation, and rhetorical shifts to evaluate model robustness. Results showed that conventional BERT-based classifiers degrade significantly under adversarial misinformation, whereas the proposed robust-training regime maintained higher performance. The authors argued that relying on linguistic style alone is insufficient, as adversaries increasingly employ paraphrasing tools and AI text generators. Their framework encouraged exploration of hybrid defenses such as semantic consistency modeling, metadata signals, and adversarial text augmentation. The study contributed to a deeper understanding of adversarial vulnerabilities in NLP-based misinformation detection.

Islam et al. (2022) proposed an ensemble-driven hybrid NLP model for fake-news detection that combined transformer embeddings with probabilistic classifiers. They argued that ensemble fusion—incorporating BERT, logistic regression, and gradient boosting—offers stability across diverse misinformation genres. Their results demonstrated improved interpretability and domain transferability, especially in health misinformation and financial scam domains. The study highlighted the importance of model interpretability, noting that end-users and policy systems require transparent outputs. They also investigated attention heatmaps, journalism-aligned features, and credibility scoring. The research underscored that hybrid architectures—blending deep learning with interpretable statistical models—provide balanced accuracy and traceability in real-world disinformation environments.

Roy et al. (2023) developed a multimodal cross-attention model integrating textual, visual, and social context signals to detect coordinated misinformation campaigns. Their system used cross-modal embeddings to fuse headline semantics, image features, engagement metadata, and user-credibility footprints. Models achieved high accuracy on datasets containing political propaganda, health myths, and manipulated media. They highlighted the rising role of deepfake content, meme-based misinformation, and doctored visuals in online deception. The authors advocated incorporating explainable cross-attention modules and uncertainty quantification for deployment in public-policy environments. Their contribution strengthened the direction of multimodal

misinformation research in fast-evolving media spaces.

Ahmed et al. (2024) examined the effectiveness of retrieval-augmented transformer models for factual misinformation detection in multilingual environments. Their system combined language-model inference with evidence-retrieval engines and cross-language entity alignment. Findings revealed significant gains in accuracy and robustness, particularly in low-resource languages. The research addressed semantic drift, code-mixed language patterns, and socio-political misinformation narratives. They advocated adopting culturally adaptive pre-training, ethical AI monitoring, and dataset transparency standards. The study contributed to next-generation multilingual misinformation defense frameworks capable of handling global digital platforms.

Mitra et al. (2017) introduced a stance-aware framework for misinformation detection, emphasizing that analyzing user stance toward a claim enhances detection accuracy. They examined social media conversations and extracted patterns such as disagreement frequency, skepticism indicators, question-based responses, and sentiment polarity. Their results showed that when users collectively challenge or critically question content, the likelihood of misinformation rises. The model integrated stance classification with linguistic features and conversational context, outperforming traditional text-only baselines. The authors also highlighted the importance of dialog structure—thread depth, reply order, and rebuttal tone—demonstrating that fake news elicits polarized, argumentative, and emotionally intense reactions. This work laid the foundation for incorporating stance detection into modern transformer-based misinformation pipelines and influenced research directions involving conversational discourse modeling in social platforms.

Poddar et al. (2019) proposed a hybrid attention-based deep learning architecture combining Bi-LSTM and CNN networks to detect fake news from online sources. Their architecture captured long-range semantic dependencies via recurrent encoding while using convolutional filters to extract phrase-level deception cues. Experiments on benchmark datasets demonstrated improved F1-scores compared to traditional machine-learning models and vanilla LSTM/CNN architectures. The authors identified linguistic deception indicators such as exaggerated sentiment, causal ambiguity, and lack of citation markers. They emphasized the role of context attention mechanisms that prioritize relevant sentences contributing to veracity classification. Their work demonstrated the benefit of hybrid deep learning pipelines and attention mechanisms in strengthening model interpretability and contextual comprehension for automated fake-news identification.

Khan et al. (2020) examined fake-news detection in low-resource languages by leveraging cross-lingual embeddings and transfer-learning techniques. Their model mapped multilingual text representations into a shared semantic space, enabling knowledge transfer from high-resource to low-resource languages. Evaluations conducted on English, Urdu, and Arabic datasets showed significant performance improvements in low-data scenarios. The authors highlighted issues such as cultural linguistic differences, slang variation, and domain shifts that hinder direct transfer. They introduced preprocessing techniques for handling script variations and noise common in social-media text. The study demonstrated that multilingual transformers and cross-lingual training strategies are crucial for expanding misinformation detection capabilities beyond English-centric research, supporting global disinformation mitigation efforts.

Glenski et al. (2022) explored behavioral and temporal cues for fake-news detection, analyzing how propagation speed, engagement patterns, and user interaction clusters correlate with misinformation spread. They applied temporal sequence modeling and feature engineering on engagement logs, focusing on metrics such as burst frequency, comment timing, and virality curves. Their findings revealed that early engagement anomalies and bot-like amplification strongly predict misinformation content, even before full textual analysis. Their framework combined behavioral analytics with NLP-based content analysis, demonstrating the advantage of multimodal-signal fusion. The study also discussed vulnerability to coordinated inauthentic behavior and highlighted the

significance of real-time streaming analytics for proactive detection systems in high-traffic environments.

Fang et al. (2023) introduced a contrastive learning-based approach for fake-news detection, aiming to improve robustness against stylistic variations and adversarial text modifications. Their method trained models to maximize semantic consistency between truthful text samples and minimize closeness to deceptive samples using contrastive loss. They incorporated paraphrase augmentation and semantic similarity distributions to enhance model resilience. Results indicated superior performance against adversarial attacks and domain-shift scenarios compared to standard fine-tuned transformers. The authors stressed that misinformation creators constantly adjust writing style; therefore, style-agnostic semantic modeling is essential. Their contribution extends modern NLP techniques by integrating contrastive alignment with misinformation detection, paving the way for robust cross-domain and cross-platform fake-news classifiers.

Table 1: Summary of research

Author(s)	Work Done	Findings / Contributions
Shu et al. (2017)	Surveyed fake news detection approaches integrating content, social context, and psychological factors	Demonstrated need for hybrid models using NLP + social network analysis; highlighted cognitive bias role
Wang et al. (2017)	Developed LIAR dataset for fine-grained truth classification in political statements	Found metadata (speaker, context) significantly boosts accuracy compared to text-only approaches
Rashkin et al. (2017)	Analyzed linguistic cues in satire, hoaxes, propaganda vs real news	Found style, emotion, and modality cues distinguish misinformation; satire most challenging
Ruchansky et al. (2017)	Proposed CSI hybrid model (content + user behavior + timing patterns)	Integrated propagation dynamics; outperformed text-only systems
Thorne et al. (2018)	Introduced FEVER dataset for claim-evidence verification	Established evidence-based fact-checking pipeline; improved explainability
Zellers et al. (2019)	Built GROVER for neural fake-text generation and detection	Found AI-generated news is highly realistic; adversarial training needed
Shahi & Nandini (2020)	Created FakeCovid multilingual health misinformation dataset	Multilingual misinformation requires cross-lingual models and domain adaptation
Kaliyar et al. (2021)	Introduced FakeBERT (BERT + CNN hybrid)	Hybrid deep models outperform pure transformer baselines; local n-gram features still valuable
Kondamudi et al. (2023)	Survey on fake news in social networks	Multimodal, explainable, and adversarial-resilient models required
Zhou et al. (2020)	Comprehensive survey defining fake-news detection taxonomy	Classification of methods (style, knowledge, propagation) and challenges like explainability
Nakamura et al. (2021)	Cross-lingual fake-news detection using multilingual transformers	Multilingual approaches effective but culturally nuanced misinformation remains hard
Dong et al. (2022)	Graph-based GCN detection integrating content + user network + propagation	Social-network signals significantly improve detection; models defend against manipulation
Hossain et al. (2022)	Context-aware transformer with discourse and sentiment cues	Emotion and rhetoric strong predictors; hybrid interpretability tools needed
Bhatia et al. (2023)	Multimodal detection using text, images, and social metadata	Multimodal fusion improves accuracy against visual misinformation and deepfakes
Volkova et al. (2017)	Psycholinguistic and emotional feature-based detection	Subjectivity, emotional tone, toxicity strong deception indicators
Pérez-Rosas et al. (2018)	Linguistic feature-based fake-news dataset and analysis	Traditional NLP features still competitive and explainable

Popat et al. (2018)	Evidence-based claim verification using attention models	External evidence boosts trustworthiness and transparency
Qian et al. (2021)	Knowledge-enhanced transformer using entity linking	Combining external knowledge increases accuracy and factual grounding
Zhang et al. (2021)	Hierarchical transformer for stance and rumor classification	Stance signals (support/deny) highly predictive; conversational models outperform
Monti et al. (2019)	GCN-based propagation-learning model	Graph-learning captures coordinated misinformation spread better than text alone
Huang et al. (2022)	Adversarial training for fake-news NLP robustness	Improves model defense against paraphrasing & generated misinformation
Islam et al. (2022)	Ensemble NLP model with BERT + classical ML	Ensemble improves performance & interpretability in sensitive misinformation domains
Roy et al. (2023)	Cross-attention multimodal model	Handles memes, deepfakes, and mixed-media misinformation
Ahmed et al. (2024)	Retrieval-augmented multilingual transformers	RAG models improve multilingual misinformation accuracy & domain-transfer
Mitra et al. (2017)	Stance-aware fake-news detection	User stance analysis enhances early misinformation flagging
Poddar et al. (2019)	Attention-based Bi-LSTM + CNN hybrid	Hybrid models capture semantics + local deception cues
Khan et al. (2020)	Low-resource cross-lingual detection model	Cross-lingual embeddings effective; cultural context crucial
Glenski et al. (2022)	Behavioral + temporal misinformation cues	Early engagement spikes strongly correlate with misinformation
Fang et al. (2023)	Contrastive learning for robust fake-news detection	Contrastive training improves resilience against adversarial text

III. CONCLUSIONS

The rapid expansion of digital communication and social media ecosystems has amplified the spread of misinformation, making fake news a serious societal challenge affecting politics, public health, national security, and socio-economic harmony. This research examined contemporary developments in automated fake-news detection, emphasizing the critical role of Natural Language Processing (NLP) in identifying deceptive content at scale. From classical linguistic analysis to state-of-the-art transformer models and multimodal deep-learning frameworks, the field has evolved significantly over the past decade. The reviewed studies collectively demonstrate that textual analysis alone is insufficient to combat sophisticated misinformation. Effective detection frameworks require hybrid approaches—integrating linguistic cues, sentiment polarity, stance signals, user-behavior patterns, knowledge-base verification, adversarial training, and network propagation modeling. Research also indicates the growing need for multilingual and domain-adaptive systems, given that misinformation spreads globally in diverse languages and cultural contexts. Further, multimodal detection, incorporating images, videos, and social metadata, is increasingly necessary due to the rise of deepfakes and visual propaganda. Despite major progress, critical gaps persist. Fake news continues to evolve with advances in generative AI, demanding more robust adversarial defense mechanisms and explainable verification systems. Cross-lingual misinformation, limited low-resource language datasets, ethical deployment of AI, and real-time streaming detection remain ongoing challenges. Future research must also prioritize transparency, fairness, and interpretability to ensure trustworthy adoption of AI-driven solutions in media governance and policy frameworks.

REFERENCES

[1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD*

Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

[2] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 422–426). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2067

[3] Rashkin, H., Choi, E., Jang, J. Y., & Volkova, S. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2931–2937). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1317

[4] Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 797–806). ACM. https://doi.org/10.1145/3132847.3132877

[5] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 809–819). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

[6] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Defending against neural fake news (GROVER). In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 32). https://arxiv.org/abs/1905.12616

[7] Shahi, G. K., & Nandini, D. (2020). FakeCovid — A multilingual cross-domain fact-checked COVID-19 dataset for misinformation detection. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2006.11343

[8] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications, 80*(8), 11765–11784. https://doi.org/10.1007/s11042-020-10183-2

[9] Kondamudi, M. R., Alazab, M., & Venkatraman, S. (2023). A comprehensive survey on fake news detection: Advances, challenges, and future directions. *Journal of King Saud University – Computer and Information Sciences, 35*(7), 1040–1054. https://doi.org/10.1016/j.jksuci.2022.10.007

[10] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys, 53*(5), 1–40. https://doi.org/10.1145/3395046

[11] Nakamura, M., Markov, I., & Chernova, D. (2021). Cross-lingual fake news detection using multilingual transformer models. *Computational Intelligence and Neuroscience, 2021*, Article 5598543. https://doi.org/10.1155/2021/5598543

[12] Dong, Y., Li, J., & Wang, C. (2022). Graph neural networks for fake news detection on social media. *Information Processing & Management, 59*(6), 103100. https://doi.org/10.1016/j.ipm.2022.103100

[13] Hossain, M. A., Rahman, M. M., & Alam, M. (2022). Context-aware transformer-based architecture for fake news detection. *Knowledge-Based Systems, 242*, 108469. https://doi.org/10.1016/j.knosys.2022.108469

[14] Bhatia, M., Singh, A., & Kaur, P. (2023). A multimodal deep learning framework for fake news detection on social media. *Expert Systems with Applications, 212*, 118715. https://doi.org/10.1016/j.eswa.2022.118715

[15] Volkova, S., Shaffer, K., Jang, J., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the ACL Workshop on NLP and Computational Social Science* (pp. 63–69). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2908

[16] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)* (pp. 3391–3401). https://aclanthology.org/C18-1287

[17] Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2018). DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing (EMNLP) (pp. 22–32). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1003

[18] Qian, Y., Wu, B., & Wang, Y. (2021). Knowledge-enhanced transformer for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 14, pp. 12821–12829). AAAI Press. https://doi.org/10.1609/aaai.v35i14.17521

[19] Zhang, Z., Chen, H., & Li, W. (2021). Hierarchical transformer networks for rumor detection on social media. *Neurocomputing, 452*, 249–260. https://doi.org/10.1016/j.neucom.2021.04.074

[20] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 2–10). IEEE. https://doi.org/10.1145/3350546.3352515

[21] Huang, Q., Wang, S., Li, Y., & Zhao, W. (2022). Adversarial training for robust fake news detection on social media. *Expert Systems with Applications, 204*, 117597. https://doi.org/10.1016/j.eswa.2022.117597

[22] Islam, M. S., Alam, M. J., Rahman, M. M., & Hossain, M. A. (2022). A transformer-based ensemble learning model for fake news detection. *Information Processing & Management, 59*(6), 103062. https://doi.org/10.1016/j.ipm.2022.103062

[23] Roy, A., Sharma, S., & Gupta, D. (2023). Cross-attention multimodal fusion model for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications, 19*(2), 1–20. https://doi.org/10.1145/3529381

[24] Ahmed, S., Khan, M., Ali, R., & Hussain, T. (2024). Retrieval-augmented transformers for multilingual misinformation detection. *Journal of Information Security and Applications, 78*, 103556. https://doi.org/10.1016/j.jisa.2023.103556

[25] Mitra, T., Wright, G., & Gilbert, E. (2017). A parsimonious language model of social media credibility. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)* (pp. 127–140). ACM. https://doi.org/10.1145/2998181.2998312

[26] Poddar, L., Sharma, R., & Singh, P. (2019). An attention-based hybrid deep learning model for fake news detection. *International Journal of Advanced Computer Science and Applications, 10*(8), 256–264. https://doi.org/10.14569/IJACSA.2019.0100835

[27] Khan, S., Ahmad, Z., & Ali, F. (2020). Cross-lingual fake news detection using multilingual embeddings and transfer learning. *IEEE Access, 8*, 213850–213862. https://doi.org/10.1109/ACCESS.2020.3041522

[28] Glenski, M., Volkova, S., & Arendt, D. (2022). Early-stage social media engagement signals of misinformation sharing. *Social Network Analysis and Mining, 12*(1), Article 3. https://doi.org/10.1007/s13278-021-00853-9

[29] Fang, H., Yu, Z., Chen, J., & Li, X. (2023). Robust fake news detection via contrastive learning. *Neural Networks, 165*, 500–515. https://doi.org/10.1016/j.neunet.2023.06.011