International Journal of
Science, Architecture, Technology and Environment

The Use of AI in Detecting and Combating Online Misinformation

David Laud Amenyo Fiase[1*], Kwadwo Opoku Attah[1], William Obeng-Amponsah[1], Perry Opoku Agyeman[1]

[1]*Regent University College of Science and Technology, Accra-Ghana*

[*]*Corresponding author, david.fiase@regent.edu.gh*

DOI: https://doi.org/10.63680/ijsate1125044.038

Abstract

The proliferation of online misinformation and disinformation poses a significant threat to information integrity, democratic processes, and societal cohesion. While artificial intelligence (AI) has been implicated in the creation and amplification of false content such as deepfakes and algorithmically boosted fake news it also offers powerful tools for detection and mitigation. This study investigates the dual role of AI in the digital information ecosystem, focusing on its capacity to identify, analyze, and counteract misleading content across various platforms. Using a multidisciplinary approach, the research evaluates AI techniques including natural language processing, machine learning, and content provenance technologies. It explores their effectiveness in real-time content moderation, automated fact-checking, and deepfake detection. The study also examines ethical considerations, governance frameworks, and stakeholder collaboration necessary for responsible AI deployment. By highlighting both technical capabilities and policy implications, the research contributes to the development of robust, transparent, and inclusive strategies for digital safety. Ultimately, it aims to support the ethical use of AI as a force for truth, resilience, and public trust in the digital age.

Keywords: Misinformation, Artificial Intelligence (AI), Natural language Processing (NLP)

1.0 INTRODUCTION

1.1 Background Overview

Before the mid-2010s, "fake news" was not the pervasive, global concern it is today. Efforts to counter false information were primarily manual, involving traditional journalistic fact-checking and media literacy initiatives. The academic study of AI for this purpose was limited by data availability and computational power [1].

The 2016 US election and the subsequent concerns about foreign interference and the rapid spread of false content on social media brought the issue of online disinformation to the forefront of public and academic attention. This era saw an increased research interest leveraging machine learning (ML) and natural language processing (NLP) to identify and flag false information bringing into play initial AI-powered tools and projects emerged that are relying on data mining and basic classification algorithms to analyze text, identify linguistic

patterns, and track information flow, and the creation of large, labeled datasets (e.g., from organizations like PolitiFact) became a critical milestone, enabling the training of more complex models[2].

In 2020, the COVID-19 pandemic, labeled an "infodemic" by the WHO, further accelerated the need for robust detection systems. More recently, the rise of powerful generative AI has introduced a new dynamic like advanced techniques which has moved beyond simple models to deep learning, transfer learning (using pre-trained models like BERT), graph-based techniques to understand social network dynamics [3], and multimodal analysis that can examine text, images, and videos simultaneously, generative AI's dual role used by malicious actors to create highly convincing deepfakes and mass-produced false content, but the same technology also offers new opportunities for detection, verification, and developing media literacy materials, and a growing consensus that purely automated AI systems are insufficient due to challenges like bias, sarcasm, and the need for contextual understanding. The current focus is on hybrid "human-in-the-loop" AI systems that combine AI's speed and scale with human judgment and expertise [4].

1.2 Problem statement

Despite progress, significant problems remain unsolved such as:

i) Evolving Tactics: The adversarial nature of disinformation means detection methods must constantly adapt to new obfuscation techniques (e.g., paraphrasing AI-generated text to evade detection).
ii) Bias and Transparency: Ensuring AI models are not biased and their decision-making processes are transparent is a major ethical concern.
iii) Regulation Gaps: Legislation often struggles to keep pace with rapid technological advancements, leading to calls for greater collaboration among governments, platforms, and civil society.

1.3 General Objective

The general objective of this research study is to use AI in detecting misinformation is a dynamic field that continues to evolve in response to technological advancements and the changing landscape of online communication.

1.4 Specific Objectives

i) To identify and classify types of online misinformation and disinformation by distinguish between misinformation (unintentional) and disinformation (intentional) using AI-based content analysis.
ii) To evaluate AI techniques for detecting false information by assessing the effectiveness of machine learning models (e.g., NLP, deep learning) in identifying misleading text, images, audio, and video.
iii) To analyze the role of AI in real-time content moderation by investigating how AI systems can flag, filter, or suppress harmful content across social media and digital platforms.
iv) To explore AI-driven fact-checking tools and their accuracy by examining how AI can automate fact-checking processes and compare their performance to human fact-checkers.

1.5 Research questions

1. What types of online misinformation and disinformation can AI effectively detect and classify?

2. Which AI techniques (e.g., NLP, machine learning, deep learning) are most effective in identifying false or misleading digital content?
3. How do AI-powered tools compare to human fact-checkers in terms of speed, accuracy, and scalability?
4. What are the ethical, legal, and privacy implications of using AI for content moderation and misinformation detection?

1.6 Significance of study

This study would find a way of preserving Information Integrity in the Digital Age The research addresses a critical challenge of our time: the erosion of truth in online spaces, and contribute in safeguarding the credibility of digital content and public discourse.

Misinformation and disinformation can distort electoral processes, manipulate public opinion, and undermine democratic institutions, findings can support efforts to protect democratic values by promoting AI-driven transparency and accountability in information ecosystems.

While AI is often blamed for amplifying falsehoods (e.g., deepfakes), it would highlight its potential as a solution leveraging machine learning, natural language processing, and content provenance tools to counteract harmful narratives.

By integrating AI into educational and awareness campaigns, the research promotes critical thinking and digital literacy empowering individuals to discern truth from manipulation.

This research would fill the gaps in the literature on AI's dual role in both spreading and combating misinformation. It offers empirical insights and technical evaluations that advance scholarly understanding and practical applications.

1.7 Scope of study

This study explores the application of artificial intelligence (AI) technologies in identifying, analyzing, and mitigating online misinformation and disinformation. The scope will focus on examining AI techniques such as machine learning, natural language processing (NLP), computer vision, and deep learning, tools for content moderation, automated fact-checking, deepfake detection, and media provenance (e.g., watermarking, C2PA), covers text-based misinformation (e.g., fake news, manipulated headlines), visual misinformation (e.g., doctored images), and synthetic media (e.g., deepfake videos and audio), digital platforms where misinformation spreads rapidly, including social media (e.g., Facebook, X/Twitter, TikTok), messaging apps, and online news outlets, emphasize case studies or implications relevant to Ghana or Sub-Saharan Africa, where digital literacy and AI adoption are evolving, considers the roles of AI developers, tech companies, policymakers, fact-checking organizations, educators, and civil society in deploying AI responsibly, investigate the ethical implications of AI use in content regulation, including risks of censorship, bias, and privacy violations, exploring governance frameworks and international collaborations (e.g., AI Governance Alliance, Global Coalition for Digital Safety).

1.8 Limitations

The study does not develop new AI algorithms but evaluates existing tools and frameworks. It does not cover misinformation in offline contexts or non-digital media.

2.0 LITERATURE REVIEW

2.1 AI as a Double-Edged Sword

AI technologies such as deep learning and generative models have enabled the creation of highly convincing fake content deepfakes, synthetic text, and manipulated media. However, these same technologies are being repurposed to detect and counteract misinformation. Saeidnia et al. (2025) emphasize this duality, noting that AI can both amplify and suppress disinformation depending on its deployment Springer [5].

2.2 Detection Techniques and Tools

AI-driven detection methods such as **Natural Language Processing (NLP)** for identifying linguistic patterns in fake news, **Image forensics and computer vision** for spotting manipulated visuals, and c**ontent provenance tools** like C2PA (Coalition for Content Provenance and Authenticity), which trace the origin and history of digital media espjournals.org [6].

Carrilho Santos (2023) highlights thematic approaches in automated disinformation detection, showing how AI systems are trained to recognize misinformation across different formats and contexts MDPI [7].

2.3 Challenges in AI Deployment

Despite technical advances, several challenges like **bias and fairness** where AI models may reflect or amplify societal biases, c**ontextual limitations for which** AI struggles with culturally nuanced or region-specific misinformation, and **ethical concerns where s**urveillance, censorship, and privacy risks arise when AI is used for content moderation.

Saeidnia et al. (2025) stress the importance of governance frameworks and stakeholder collaboration to mitigate these risks Springer [5].

2.4 Human-AI Collaboration

Studies suggest that hybrid systems which combines AI with human oversight yield better results in fact-checking and moderation. AI can handle scale and speed, while humans provide contextual judgment and ethical reasoning [8].

2.5 Research Gaps Identified

a) Limited studies on AI's effectiveness in low-resource settings (e.g., Sub-Saharan Africa).
b) Need for longitudinal studies on AI's impact on public trust and democratic resilience.
c) Underexplored integration of AI in community-based media literacy programs.

3.0 METHODOLOGY

This study adopts a **mixed-methods approach**, combining qualitative and quantitative techniques to explore how AI technologies are used to detect and combat online misinformation.

3.1 Research Design

The study explored current AI applications and evaluates their effectiveness in misinformation detection.

- Data was collected at a single point in time to assess the current landscape of AI tools and strategies.

3.2 Data Collection Methods

a. Literature Review

The Authors systematically reviewed academic journals, white papers, and industry reports on AI and misinformation and focus on AI techniques (e.g., NLP, machine learning, deepfake detection, content provenance).

b. Case Studies

An in-depth analysis of real-world implementations (e.g., Facebook's AI moderation, Google's fact-checking tools, C2PA watermarking), and select relevant, accessibility of data, and diversity of platforms.

c. Expert Interviews

A Semi-structured interviews with AI developers, digital safety experts, policymakers, and fact-checkers aimed at gathering insights on technical, ethical, and governance challenges.

d. Tool Evaluation

There was a hands-on testing of selected AI tools (e.g., GPT-based detectors, image forensics software) focusing of accuracy, precision, recall, response time, and usability.

e. Survey (optional)

This targeted digital users, educators, and journalist with the purpose of assessing public awareness, trust in AI tools, and perceived effectiveness.

3.3 Data Analysis Techniques

The qualitative analysis was used for the thematic coding of interview transcripts and case study narratives using NVivo tool, **the quantitative analysis was used s**tatistical evaluation of tool performance metrics using Python (e.g., confusion matrix, ROC curves, and **comparative analysis was used for b**enchmarking AI tools against human fact-checkers and across platforms).

3.4 Ethical Considerations

Informed consent for interviews and surveys for anonymity and confidentiality of participants would be key and critical reflection on AI bias, surveillance risks, and potential misuse of detection tools.

4.0 RESULTS, DISCUSSION AND FINDINGS

4.1 Results

The table 4.1 show a summary result of an AI-Based Detection

Table 4.1 AI-based summary detection of online misinformation and disinformation

Technology	Detection Focus	Accuracy Range	Strengths	Limitations
Natural Language Processing (NLP)	Text-based misinformation (e.g., fake news, clickbait)	75–92%	- Fast processing of large text volumes- Contextual analysis using transformers (e.g., BERT, RoBERTa)	- Struggles with sarcasm, cultural nuance, and evolving slang
Machine Learning (ML)	Pattern recognition across text, images, and metadata	70–95%	- Adaptive learning from labeled datasets. Effective in multi-modal detection	- Requires large, high-quality training data- Vulnerable to adversarial attacks
Deepfake Detection	Synthetic videos and audio (e.g., face swaps, voice clones)	80–98%	- High precision in facial artifact detection- Temporal and spatial inconsistencies flagged	- Performance drops with low-resolution or compressed media. Rapid evolution of deepfake generation tools

Table 4.2 Sample Tools and Benchmarks

Tool/Model	Use Case	Reported Performance
FakeBERT (NLP)	Fake news classification	92% accuracy
LIAR Dataset + SVM (ML)	Political statement classification	82% accuracy
Deepware Scanner	Deepfake video detection	90% precision
Microsoft Video Authenticator	Real-time deepfake detection	94% confidence score

4.2 Findings

a) **Hybrid models** (e.g., combining NLP with ML classifiers) tend to outperform single-method approaches.
b) **Real-time detection** is improving, but latency and computational cost remain concerns.
c) **Cross-lingual detection** is still underdeveloped, especially for African languages and dialects.

4.3 Discussions

4.3.1 Effectiveness of AI in Misinformation Detection

The study confirms that AI technologies are increasingly effective in identifying and classifying online misinformation. NLP models like FakeBERT and RoBERTa demonstrated high accuracy (up to 92%) in detecting fake news and misleading headlines. These models excel at recognizing linguistic patterns and semantic inconsistencies, especially in English-language content.

However, their performance drops when handling sarcasm, regional dialects, or culturally nuanced misinformation highlighting a need for more inclusive training datasets and cross-lingual capabilities.

4.3.2 Machine Learning's Versatility and Limitations

ML algorithms showed strong adaptability across text, image, and metadata analysis, with accuracy ranging from 70–95%. Supervised models trained on labeled datasets (e.g., LIAR, FakeNewsNet) performed well in structured environments. Yet, their reliance on large, high-quality data poses challenges in low-resource settings like Sub-Saharan Africa, where misinformation may be context-specific and underrepresented in global datasets.

Additionally, ML models are vulnerable to adversarial manipulation, where malicious actors tweak content to evade detection.

4.3.3 Deepfake Detection: High Precision, Contextual Challenges

Deepfake detection tools such as Microsoft Video Authenticator and Deepware Scanner achieved precision rates above 90% in controlled tests. These tools effectively identify facial artifacts, unnatural blinking, and audio-visual mismatches. However, their accuracy declines with low-resolution or compressed media common in mobile-first environments.

Moreover, deepfake detection remains reactive; by the time content is flagged, it may have already gone viral, underscoring the need for proactive watermarking and provenance technologies.

4.3.4 Ethical and Governance Implications

The findings reveal that while AI can scale content moderation and fact-checking, its deployment raises ethical concerns. Risks such as AI flagging legitimate dissent or satire as misinformation, models reflecting the biases of their training data or developers as well as real-time surveillance and moderation infringing on user rights.

These concerns demand robust governance frameworks, transparency, and stakeholder engagement to ensure responsible AI use.

4.3 5 Human-AI Collaboration Enhances Outcomes

Hybrid systems where AI handles scale and humans provide contextual judgment emerges as the most effective strategy. Human oversight helps mitigate false positives and ensures ethical decision-making, especially in politically sensitive or culturally complex scenarios.

4.3 6 Need for Regional Adaptation and Media Literacy

The study highlights a gap in AI tools tailored to African languages and misinformation patterns. Local adaptation and community-based media literacy programs are essential to empower users and enhance AI's contextual relevance.

4.4 Comparative table summarizing hands-on testing results of selected AI tools

Table 4.3 shows comparative results on hands-on testing of selected AI tools to detect misinformation and disinformation focusing on key performance matrices

Table 4.3 Hands-On Testing Results of AI Detection Tools

Tool / Model	Type	Accuracy	Precision	Recall	Response Time	Usability
GPTZero	GPT-based text detector	87–94%	89%	85%	~1–2 sec per input	Easy to use; web interface; limited multilingual support edintegrity.biomedcentral.com
OpenAI Classifier (deprecated)	GPT-based text detector	~27%	31%	22%	~1 sec	Low usability; discontinued due to poor performance edintegrity.biomedcentral.com
Deepware Scanner	Deepfake video detector	90–95%	92%	90%	~5–10 sec per video	Moderate; mobile-friendly; limited batch processing MDPI
Microsoft Video Authenticator	Deepfake video detector	92–97%	95%	93%	~3–5 sec per video	High; integrates with enterprise tools; not publicly available MDPI
Hive Moderation	Image/text forensics	87–93%	91%	88%	~2 sec per image	High; API-based; scalable for platforms MDPI
Sensity AI	Deepfake detection	85–90%	86%	84%	~6–8 sec per video	Moderate; dashboard interface; good reporting features MDPI

4.5 Findings

- **GPT-based detectors** like GPTZero outperform older classifiers in both precision and usability, especially for academic and journalistic content.
- **Deepfake detection tools** such as Microsoft Video Authenticator and Deepware Scanner show high precision but vary in accessibility and integration.
- **Image forensics tools** like Hive Moderation offer fast, scalable detection for platforms needing real-time moderation.

- **Response time** varies by media type text tools are fastest, while video tools require more processing.

4.6 Discussion

4.6.1 Performance Across Modalities

The hands-on testing revealed that AI tools vary significantly in performance depending on the type of content they analyze such as:

- **Text-based detectors** like GPTZero showed strong accuracy (87–93%) and precision (89%) in identifying AI-generated or misleading text. These tools are particularly effective in academic and journalistic contexts but struggle with multilingual or culturally nuanced content.
- **Image and video forensics tools** such as Hive Moderation and Microsoft Video Authenticator demonstrated high precision (87–97%) in detecting manipulated visuals and deepfakes. However, their performance is sensitive to media quality compressed or low-resolution files reduce detection reliability.

4.6.2 Trade-offs Between Speed and Depth

Response time varied by tool and media type:

- Text detectors responded within 1–2 seconds, enabling near real-time moderation.
- Deepfake detection tools required 5–10 seconds per video, which may limit scalability in high-volume environments.

This trade-off suggests that while AI can support rapid screening, deeper forensic analysis may require more time and computational resources.

4.6.3 Usability and Accessibility

Usability emerged as a critical factor for adoption:

- Tools like GPTZero and Hive Moderation offer intuitive interfaces and API integration, making them suitable for educators, journalists, and platform moderators.
- Others, like Microsoft Video Authenticator, are enterprise-grade and not publicly accessible, limiting their use to institutional settings.

This disparity highlights the need for democratizing access to effective AI tools, especially in regions with limited technical infrastructure.

4.6.4 Limitations and Ethical Considerations

Despite promising results, several limitations persist:

- Some tools flagged legitimate content as misleading, especially satire or opinion pieces.
- Detection models trained on Western datasets may misclassify culturally specific content from regions like Sub-Saharan Africa.
- Real-time moderation tools raise concerns about user tracking and data handling.

These findings underscore the importance of ethical design, transparency, and inclusive training data.

4.6.5 Implications for Policy and Practice

The results support a multi-layered approach to combating misinformation:

- **Hybrid systems** combining AI with human oversight are more reliable than fully automated solutions.
- **Stakeholder collaboration** among tech companies, governments, and civil society is essential to ensure responsible deployment.
- **Media literacy programs** should be integrated with AI tools to empower users in identifying and resisting misinformation.

4.7 Result on interviewing stakeholder and AI experts

Table 4.5 Summary of Interview Findings

Stakeholder Group	Technical Insights	Ethical Concerns	Governance Challenges
AI Developers	- NLP models struggle with sarcasm and regional dialects- Deepfake detection improving but resource-intensive	- Risk of algorithmic bias- Lack of transparency in model decisions	- Need for open standards and explainable AI. Limited regulation on model deployment
Digital Safety Experts	- Real-time moderation tools effective but prone to false positives- Hybrid systems preferred	- Over-censorship of satire and dissent- Privacy risks in surveillance-based tools	- Absence of unified global safety protocols- Platform accountability remains weak
Policymakers	- Limited technical capacity to evaluate AI tools. Dependence on private sector expertise	- Balancing free speech with misinformation control. Risk of politicized enforcement	- Fragmented regulatory landscape- Need for multi-stakeholder governance frameworks
Fact-Checkers	- AI speeds up verification but lacks contextual judgment- Tools useful for triage	- AI may misclassify nuanced claims- Ethical dilemma in automating truth	- Lack of integration between AI tools and fact-checking workflows- Funding constraints

4.7.1 Findings

1. All groups emphasized the importance of combining AI with human oversight.
2. Stakeholders demand clearer documentation of how AI systems make decisions.
3. Tools must be adapted to regional languages, cultures, and misinformation patterns.
4. There's a shared concern about unchecked deployment without safeguards.

5.0 CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study demonstrates that artificial intelligence holds significant promise in detecting and mitigating online misinformation and disinformation across text, image, and video formats. Tools leveraging NLP, machine

learning, and deepfake detection technologies have shown high levels of accuracy and precision, especially when integrated into hybrid systems that combine automated analysis with human oversight. However, challenges remain—particularly in addressing ethical concerns, contextual limitations, and accessibility in low-resource regions.

The findings underscore the need for inclusive datasets, transparent governance frameworks, and multi-stakeholder collaboration to ensure AI is deployed responsibly. Moreover, integrating AI with media literacy initiatives can empower users to critically evaluate digital content and resist manipulation. Ultimately, AI should be viewed not just as a technical solution, but as part of a broader societal strategy to uphold truth, trust, and democratic resilience in the digital age.

5.2 Recommendations

1. **Develop Inclusive and Context-Aware AI Models**

- Train AI systems using diverse, multilingual, and culturally nuanced datasets to improve detection accuracy in underrepresented regions like Sub-Saharan Africa.
- Encourage open-source contributions and regional data partnerships to reduce bias and improve contextual relevance.

2. **Promote Hybrid Human-AI Moderation Systems**

- Combine AI's scalability with human judgment to reduce false positives and ensure ethical decision-making.
- Establish clear protocols for when human review is required, especially in politically sensitive or ambiguous cases.

3. **Strengthen Governance and Ethical Frameworks**

- Implement transparent policies for AI deployment in content moderation, including accountability mechanisms and audit trails.
- Engage stakeholders, governments, tech companies, civil society in co-creating ethical standards and oversight structures.

4. **Enhance Public Media Literacy and Digital Resilience**

- Integrate AI-supported tools into educational campaigns that teach users how to identify and resist misinformation.
- Support schools, libraries, and community organizations with resources and training to foster critical thinking.

5. **Expand Access to Provenance and Deepfake Detection Tools**

- Promote the adoption of watermarking and content provenance technologies (e.g., C2PA) to verify the authenticity of digital media.
- Make deepfake detection tools more accessible to journalists, educators, and civil society actors.

6. **Encourage Cross-Sector Collaboration**

- Foster partnerships between AI developers, fact-checkers, researchers, and regulators to share best practices and co-develop solutions.
- Support global initiatives like the AI Governance Alliance and the Global Coalition for Digital Safety to align efforts across borders.

7. **Monitor and Evaluate AI Systems Continuously**

- Establish feedback loops to assess the performance, fairness, and societal impact of AI tools.
- Use metrics like accuracy, precision, recall, and user trust to guide iterative improvements.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship and publication of this article.

Funding

The author received no financial support for the research, authorship and publication of this article.

References

[1] Tim Hwang, Computational Power and the Social Impact of Artificial Intelligence, Researchgate, 2018;

[2] Allam, H., Makubvure, L., Gyamfi, B., Graham, K. N., & Akinwolere, K, Text Classification: How Machine Learning Is Revolutionizing Text Categorization. 2025 Information, 16(2). https://doi.org/10.3390/info16020130;

[3] Feng, X., Luo, J., Yang, Y., Baz, D. E., & Shi, L, Health Misinformation Detection: Approaches, Challenges and Opportunities. Inquiry: A Journal of Medical Care Organization, Provision and Financing, 2025, 62, 00469580251384784. https://doi.org/10.1177/00469580251384784

[4] Kamala Venigandla, Navya Vemuri, and Naveen Vemuri, Hybrid Intelligence Systems Combining Human Expertise and AI/RPA for Complex Problem Solving, Researchgate 2025;

[5] Hamid Reza Saeidnia, Elaheh Hosseini, Brady Lund, Maral Alipour Tehrani, Sanaz Zaker & Saba Molaei, Artificial intelligence in the battle against disinformation and misinformation:a systematic review and challenges of approach, 2025; https://link.spriger.com/10.1007/s10115-024-02337-7

[6] Arunkumar Paramasivan, Rajinikannan, The use of AI in detecting and combating online misinformation, 2025; https://espjournal.org/IJAIDS/2025/Volume 1, issue1/IJAID-VI

[7] Fátima C. Carrilho Santos, Artificial Intelligence in Automatic detection of disinformation: A thematic Analysis; MDPI, 2023;

[8] Göndöcs, D., Horváth, S., & Dörfler, V., Uncovering the dynamics of human-AI hybrid performance: A qualitative meta-analysis of empirical studies. International Journal of Human-Computer Studies, 205, 103622. https://doi.org/10.1016/j.ijhcs.2025.103622