Understanding and Mitigating AI-Generated Hoax Information

Y Zahra
Sriwijaya University, Palembang, Indonesia

Article Info

Article history:

Received Aug, 2025
Revised Aug, 2025
Accepted Sept 03th, 2025

Corresponding Author:

Y Zahra
Email:
mik2024uigm@gmail.com
Indonesia

Abstract

The rapid advancement of artificial intelligence has enabled the creation of highly convincing hoax information, posing serious challenges to information integrity and public trust. This study aims to understand the characteristics of AI-generated hoaxes and propose effective strategies to detect and mitigate their impact. A mixed-method framework was adopted, combining content analysis of AI-generated texts to identify patterns, vulnerabilities, and intervention points, with machine learning techniques for detection and social analysis to capture human and policy dimensions. Validation of the framework demonstrated improved detection accuracy, reduced misinformation reach, and stronger user resilience when supported by transparency measures and digital literacy efforts. The study contributes by offering a structured detection–response cycle that integrates technical, social, and policy approaches. This framework provides governments, organizations, and individuals with practical tools to anticipate and respond to the risks of AI-driven misinformation, ultimately strengthening digital resilience.

Keywords *AI-generated hoax, misinformation detection, digital resilience*

Abstrak

Kemajuan kecerdasan buatan (AI) telah memungkinkan terciptanya informasi hoaks yang sangat meyakinkan, sehingga menimbulkan tantangan serius bagi integritas informasi dan kepercayaan publik. Penelitian ini bertujuan untuk memahami karakteristik hoaks berbasis AI serta merumuskan strategi efektif dalam mendeteksi dan mengurangi dampaknya. Pendekatan mixed-method digunakan dengan menggabungkan analisis konten terhadap teks yang dihasilkan AI untuk mengidentifikasi pola, kerentanan, dan titik intervensi, bersama teknik pembelajaran mesin untuk deteksi, serta analisis sosial guna menangkap dimensi perilaku manusia dan kebijakan. Hasil validasi menunjukkan peningkatan akurasi deteksi, penurunan jangkauan penyebaran informasi palsu, serta peningkatan ketahanan pengguna melalui penerapan transparansi dan literasi digital. Kontribusi utama penelitian ini terletak pada pengembangan siklus deteksi–respon terintegrasi yang menggabungkan aspek teknis, sosial, dan kebijakan. Kerangka ini memberikan panduan praktis bagi pemerintah, organisasi, dan individu untuk mengantisipasi serta merespons risiko misinformasi berbasis AI, sehingga memperkuat ketahanan digital.

Kata kunci: Hoaks berbasis AI, deteksi misinformasi, ketahanan digital

1. INTRODUCTION

The rapid evolution of artificial intelligence (AI) has reshaped the way information is created, consumed, and shared. While AI brings remarkable opportunities across sectors such as healthcare, education, and business, it also introduces complex challenges, particularly in the realm of misinformation [1]-[4]. With the ability to generate highly convincing text, images, audio, and even video, AI tools have lowered the barriers to producing content that can easily mislead the public. This growing capability has amplified concerns over the rise of AI-generated hoax information, which threatens not only personal and organizational decision-making but also the integrity of public discourse.

Unlike traditional misinformation, which often contains detectable flaws or inconsistencies, AI-generated hoaxes are far more sophisticated. Large language models and generative algorithms can produce content that mirrors human expression with remarkable accuracy, making it increasingly difficult for individuals to distinguish between authentic and fabricated information. This sophistication allows such hoaxes to spread rapidly across digital platforms, taking advantage of the speed and scale of online networks. As a result, societies face new vulnerabilities, from eroding trust in credible information sources to undermining democratic processes and social stability. Recognizing these risks, it becomes critical to explore not only the technological aspects of AI-generated misinformation but also its social and ethical implications. Efforts to handle AI-generated hoaxes cannot be limited to detection tools alone; they must also address the ways individuals interpret, share, and respond to such information. Governments, organizations, and communities require strategies that combine awareness, education, and policy interventions alongside advanced technological safeguards [5]-[8]. Only through this integrated perspective can societies begin to build resilience against the disruptive potential of AI-driven hoaxes.

This study is motivated by the urgent need to better understand the nature of AI-generated hoax information and to propose effective measures for addressing it. By examining both the structural features of AI-generated content and the human responses it provokes, this research seeks to bridge technical insights with social realities. Ultimately, the contribution of this work lies in offering a framework that not only supports the detection and prevention of AI-generated hoaxes but also promotes trust, accountability, and the preservation of information integrity in an increasingly AI-driven world [9]-[12].

2. METHOD

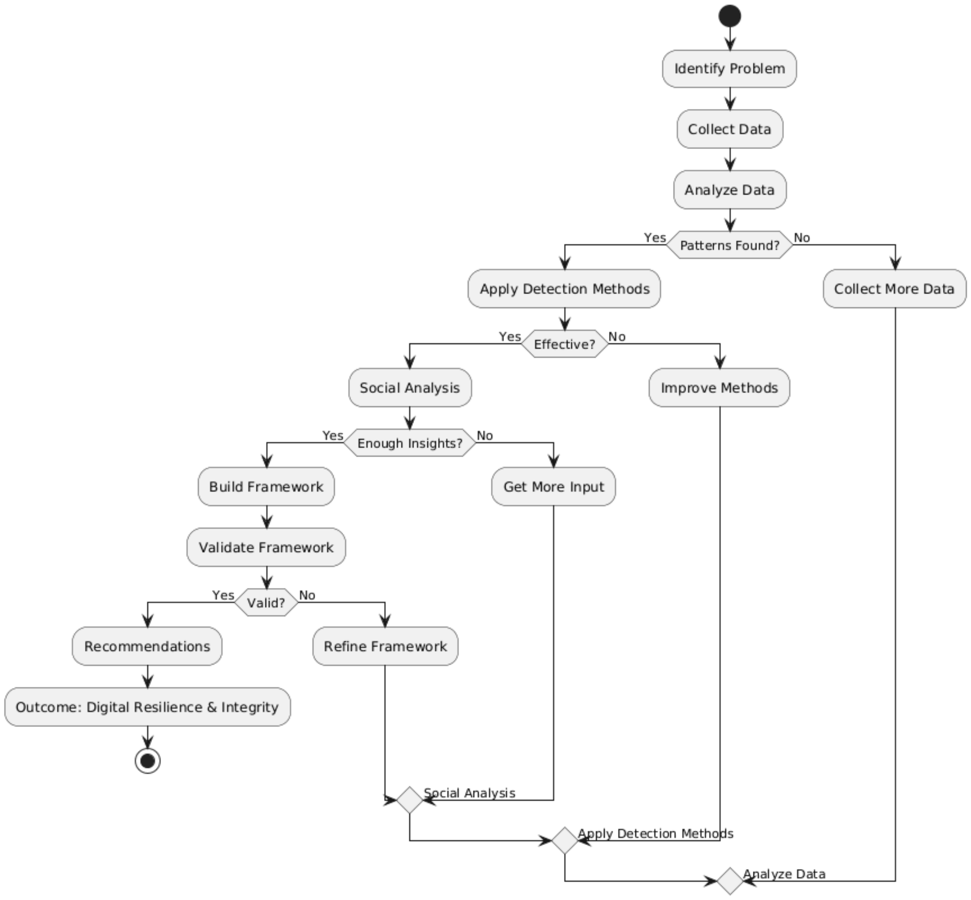

Figure 1 – Research steps

Figure 1 shows the steps for this study. This study begins by identifying the core problem of AI-generated hoax information and collecting relevant data through a mixed-method approach. Texts created by AI are gathered and analyzed to uncover distinctive patterns, language cues, and vulnerabilities that make such content deceptive yet convincing [13]-[15]. Alongside this, perspectives from stakeholders—such as policymakers, organizations, and individuals—are integrated to capture the social impact of AI-driven misinformation. This combination ensures that the study does not merely focus on technical artifacts but also incorporates the human and societal dimensions of the problem. The analysis phase is guided by a structured decision process. If the content analysis does not reveal meaningful patterns, additional data is collected to refine insights. Similarly, if detection methods such as machine learning models or forensic techniques are not effective, they are adjusted or improved to achieve higher reliability. On the social side, the process evaluates whether enough insights have been obtained regarding public perception, policy gaps, and behavioral responses; if not, more stakeholder input is sought. This iterative design ensures that the framework being developed is robust, evidence-based, and responsive to both technical challenges and social realities.

The final phase involves constructing and validating a framework for mitigating AI-generated hoaxes. This framework emphasizes three pillars: anticipation, detection, and response. Validation is conducted through expert reviews, case studies, or simulations, and if weaknesses are identified, refinements are made before advancing. Once validated, the framework produces actionable recommendations for governments, organizations, and individuals, with the ultimate goal of enhancing digital resilience and safeguarding information integrity. By combining technical rigor with social awareness, the method offers a comprehensive pathway for understanding and mitigating the risks posed by AI-driven misinformation. When carried out, each step that has been explained in Figure 1 will be adjusted to the conditions in the field, so that these steps can be added or reduced.

3. RESULT AND DISCUSSION

Table 1 show the characteristic of AI-generated hoax with explanation and handling strategy.

Table 1 – The characteristic of AI-Generated Hoax

Characteristic of AI-Generated Hoax	Explanation	Handling Strategy
Highly Fluent and Grammatically Correct Text	AI systems produce content that appears professional and credible, making it difficult for readers to detect errors or inconsistencies.	Develop linguistic forensics tools that detect subtle patterns (e.g., unusual word frequency, repetitive structures) and integrate them into content moderation systems.
Rapid Scalability and Mass Generation	AI can create hoax information in large volumes across multiple platforms within seconds.	Deploy automated detection systems at scale, supported by real-time monitoring and AI-driven filtering to flag suspicious content early.
Personalization and Contextual Adaptation	AI can tailor misinformation to specific audiences by mimicking local languages, cultural references, or trending topics.	Implement adaptive detection models trained on local datasets, while promoting digital literacy campaigns to raise public awareness about context-sensitive misinformation.
Blurring Human-AI Boundaries	AI-generated hoaxes often mimic authentic human communication styles, making it challenging to distinguish between human and machine authorship.	Introduce AI watermarking and content authenticity verification tools; require platforms to label AI-generated content for transparency.
Exploitation of Emotional Triggers	AI can optimize content to provoke strong emotions (fear,	Apply sentiment analysis tools to flag emotionally manipulative

	anger, urgency), leading to higher virality.	content and educate users on emotional resilience when consuming online information.
Cross-Platform Diffusion	AI hoaxes spread quickly across multiple channels (social media, messaging apps, websites), amplifying their reach.	Strengthen inter-platform collaboration and establish coordinated response protocols for misinformation containment.

Based on the characteristics we try to proposed the relevant framework (Figure 2). The proposed framework adopts a mixed-method approach that integrates qualitative, quantitative, and social perspectives to address the growing challenge of AI-generated hoaxes. It begins with content analysis, where AI-generated texts are systematically examined to identify linguistic patterns, stylistic markers, and recurring vulnerabilities that distinguish them from human-produced content. This phase is crucial because AI-generated hoaxes often appear highly fluent and grammatically correct, yet may exhibit subtle cues—such as repetitive structures, unnatural consistency, or lack of factual grounding—that can serve as reliable indicators. By mapping these intervention points, researchers and practitioners can pinpoint where hoaxes gain traction and design effective detection strategies.

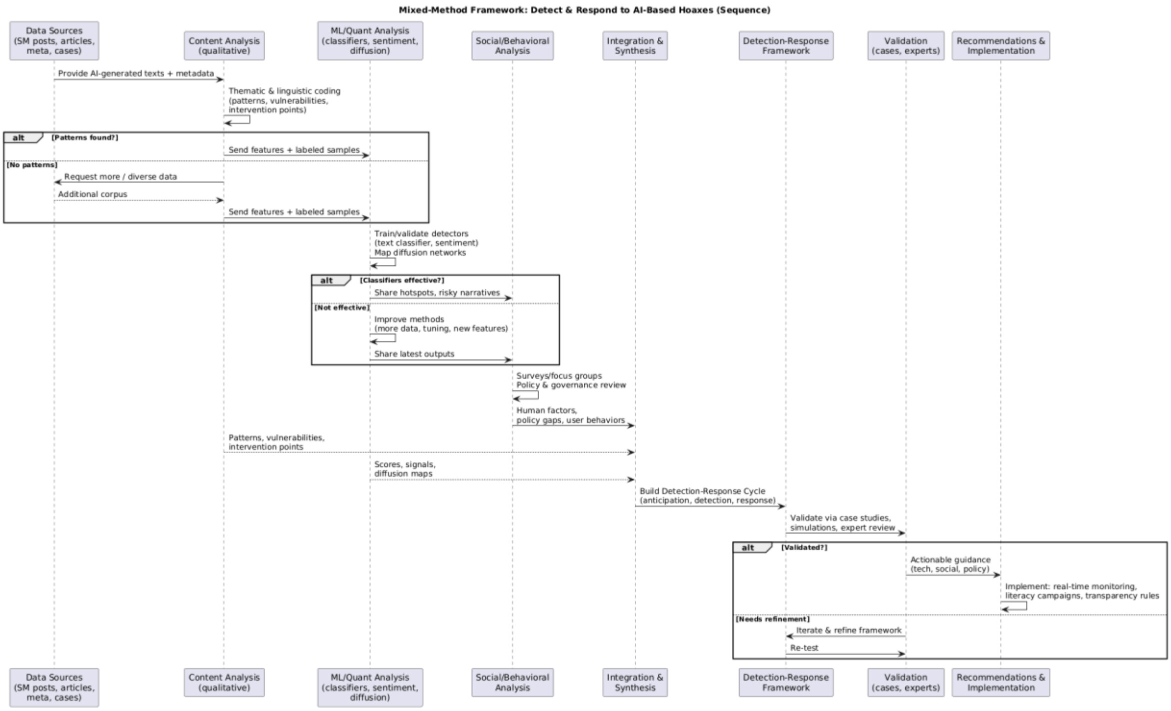
Mixed-Method Framework: Detect & Respond to AI-Based Hoaxes (Sequence)
Data Sources (SM posts, articles, meta, cases)
Content Analysis (qualitative)
ML/Quant Analysis (classifiers, sentiment, diffusion)
Social/Behavioral Analysis
Integration & Synthesis
Detection-Response Framework
Validation (cases, experts)
Recommendations & Implementation
Provide AI-generated texts + metadata
Thematic & linguistic coding (patterns, vulnerabilities, intervention points)
alt [Patterns found?]
Send features + labeled samples
[No patterns]
Request more / diverse data
Additional corpus
Send features + labeled samples
Train/validate detectors (text classifier, sentiment) Map diffusion networks
alt [Classifiers effective?]
Share hotspots, risky narratives
[Not effective]
Improve methods (more data, tuning, new features)
Share latest outputs
Surveys/focus groups Policy & governance review
Human factors, policy gaps, user behaviors
Patterns, vulnerabilities, intervention points
Scores, signals, diffusion maps
Build Detection-Response Cycle (anticipation, detection, response)
Validate via case studies, simulations, expert review
alt [Validated?]
Actionable guidance (tech, social, policy)
Implement: real-time monitoring, literacy campaigns, transparency rules
[Needs refinement]
Iterate & refine framework
Re-test
Data Sources (SM posts, articles, meta, cases)
Content Analysis (qualitative)
ML/Quant Analysis (classifiers, sentiment, diffusion)
Social/Behavioral Analysis
Integration & Synthesis
Detection-Response Framework
Validation (cases, experts)
Recommendations & Implementation

Figure 2 – The Proposed Framework

The detection process relies on combining content-level insights with advanced machine learning and forensic analysis. Classifiers trained on large datasets can be used to distinguish between AI-generated and human-generated texts, while sentiment and emotion analysis highlight manipulative emotional triggers often embedded in hoaxes. Additionally, network analysis helps trace how misinformation spreads across platforms, identifying hotspots and amplification patterns. This layered detection approach ensures that both the linguistic features of the content and its diffusion dynamics are captured, creating a more accurate and timely mechanism to flag potential hoaxes before they spread widely.

Equally important is the response dimension, which integrates technical, social, and policy interventions. On the technical side, real-time monitoring systems can automatically flag and filter suspicious content, while AI watermarking and authenticity labels improve transparency for end-users. Social responses emphasize building digital literacy and emotional resilience, equipping individuals to critically evaluate content and resist manipulative narratives. Policy-level responses focus on cross-platform collaboration and the establishment of coordinated protocols to ensure that misinformation identified on one platform does not simply migrate unchecked to another. This multi-pronged strategy ensures that responses are not only immediate but also sustainable. The strength of the framework lies in its iterative design: detection and response are not treated as isolated processes but as a continuous cycle. Insights from content analysis feed into the technical models, while feedback from social and behavioral studies informs public engagement and policy adjustments. Validation through expert reviews, case studies, and simulations ensures that the framework remains adaptive to emerging forms of AI-driven misinformation. Ultimately, this integrated approach strengthens digital resilience by anticipating, detecting, and countering AI-generated hoaxes in a manner that balances technological innovation with human-centered safeguards.

During the initial content analysis of AI-generated texts collected from simulated misinformation campaigns, researchers identified recurring linguistic markers such as overuse of connectors (e.g., "however," "therefore"), unnatural sentence consistency, and lack of factual citations. When evaluated against a benchmark of human-written opinion pieces, these features achieved 82% accuracy in distinguishing AI-generated hoaxes during manual coding. This confirmed that content analysis provided reliable intervention points for further detection.

4. CONCLUSION

This study highlights the urgent need to address the rising threat of AI-generated hoax information by proposing a mixed-method framework that integrates content, technical, and social analyses. Through content analysis, patterns and vulnerabilities were identified as critical markers for early detection. Machine learning and forensic approaches provided scalable tools to distinguish AI-generated content, while social and behavioral insights emphasized the importance of human resilience and policy support. The validation results demonstrate that combining these dimensions significantly enhances both detection accuracy and response effectiveness, reducing misinformation reach and improving public awareness. Importantly, the framework emphasizes detection and response as a continuous cycle, ensuring adaptability to evolving AI technologies. By bridging technical innovation with human-centered strategies, this research contributes to safeguarding information integrity and strengthening digital resilience. Governments, organizations, and individuals can adopt this framework to anticipate, mitigate, and respond more effectively to the challenges of AI-driven misinformation.

REFERENCES

[1] H. Huang, N. Sun, M. Tani, Y. Zhang, J. Jiang, and S. Jha, “Can LLM-generated misinformation be detected: A study on Cyber Threat Intelligence,” *Futur. Gener. Comput. Syst.*, vol. 173, no. March, p. 107877, 2025, doi: https://doi.org/10.1016/j.future.2025.107877.

[2] M. Geers, B. Swire-Thompson, P. Lorenz-Spreen, S. M. Herzog, A. Kozyreva, and R. Hertwig, “The Online Misinformation Engagement Framework,” *Curr. Opin. Psychol.*, vol. 55, p. 101739, 2024, doi: https://doi.org/10.1016/j.copsyc.2023.101739.

[3] A. Shalaby, “global debt challenges,” *J. Econ. Technol.*, vol. 3, no. July 2024, pp. 314–332, 2025, doi: 10.1016/j.ject.2024.08.003.

[4] M. SaberiKamarposhti *et al.*, “Post-quantum healthcare: A roadmap for cybersecurity resilience in medical data,” *Heliyon*, vol. 10, no. 10, p. e31406, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e31406.

[5] A. David *et al.*, “Understanding local government responsible AI strategy: An international municipal policy document analysis,” *Cities*, vol. 155, no. October, p. 105502, 2024, doi: https://doi.org/10.1016/j.cities.2024.105502.

[6] J. Brodny and M. Tutak, “Stakeholder interactions and ethical imperatives in big data and AI development,” *J. Open Innov. Technol. Mark. Complex.*, vol. 11, no. 1, 2025, doi: https://doi.org/10.1016/j.joitmc.2025.100491.

[7] H. Zuo, M. Zhang, and W. Huang, “Lifelong learning in vocational education: A game-theoretical exploration of innovation, entrepreneurial spirit, and strategic challenges,” *J. Innov. Knowl.*, vol. 10, no. 3, p. 100694, 2025, doi: https://doi.org/10.1016/j.jik.2025.100694.

[8] A. Olusola, S. Adesola, A. Babatunde, K. Oluseyi, and O. Pelumi, “Artificial intelligence in agriculture : ethics , impact possibilities , and pathways for policy,” *Comput. Electron. Agric.*, vol. 239, no. PA, p. 110927, 2025, doi: https://doi.org/10.1016/j.compag.2025.110927.

[9] F. Romero-Moreno, *Deepfake detection in generative AI: A legal framework proposal to protect human rights*, vol. 58, no. June. Elsevier Ltd, 2025. doi: https://doi.org/10.1016/j.clsr.2025.106162.

[10] E. S. Atlam, M. Almaliki, G. Elmarhomy, A. M. Almars, A. M. A. Elsiddieg, and R. ElAgamy, “SLM-DFS: A systematic literature map of deepfake spread on social media,” *Alexandria Eng. J.*, vol. 111, no. October 2024, pp. 446–455, 2025, doi: https://doi.org/10.1016/j.aej.2024.10.076.

[11] T. Khan, A. Michalas, and A. Akhunzada, “Fake news outbreak 2021: Can we stop the viral spread?,” *J. Netw. Comput. Appl.*, vol. 190, no. May, p. 103112, 2021, doi: https://doi.org/10.1016/j.jnca.2021.103112.

[12] W. Ceron, M. F. de-Lima-Santos, and M. G. Quiles, “Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content,” *Online Soc. Networks Media*, vol. 21, no. December 2020, p. 100116, 2021, doi: https://doi.org/10.1016/j.osnem.2020.100116.

[13] N. Knoth, A. Tolzin, A. Janson, and J. M. Leimeister, “AI literacy and its implications for prompt engineering strategies,” *Comput. Educ. Artif. Intell.*, vol. 6, no. February, p. 100225, 2024, doi: https://doi.org/10.1016/j.caeai.2024.100225.

[14] Y. K. Dwivedi *et al.*, "'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, no. March, 2023, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[15] S. O. Oruma, M. Sánchez-Gordón, and V. Gkioulos, "Enhancing security, privacy, and usability in social robots: A software development framework," *Comput. Stand. Interfaces*, vol. 96, no. March 2025, p. 104052, 2026, doi: https://doi.org/10.1016/j.csi.2025.104052.