

AI for Detecting Misinformation: A Discussion on the Case of COVID-19 in Indonesia

Santi Indra Astuti¹, Dyanning Pangestika²

¹The Faculty of Communication Science

Universitas Islam Bandung (Unisba), Indonesia, Jl. Tamansari No.1, Tamansari, Kec. Bandung Wetan, Kota Bandung, Jawa Barat 40116, Indonesia

(santi@unisba.ac.id)

²Integrated Marketing Communication Department, School of Communication

Level 20, MoF Inc. Tower, Platinum Park, 9, Persiaran KLCC, Kuala Lumpur, 50088 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia

(dyanningp@student.usm.my)

Abstract

The advent of generative Artificial Intelligence (AI) was viewed as a threat to the information ecosystem due to generative AI's ability to create 'stories' that might easily be twisted as misinformation. Generative AI was alleged to be a high-risk tool for fact checkers, journalists, public officials, and others responsible for verifying and sharing correct information, increasing the possibility of widespread misinformation and disinformation in the society. However, as technology evolves, so does humans' comprehension of the new machines. There are many benefits that could be explored from generative AI, as the platform offers a plethora of potential applications. To explore the possibility, this research investigates the potential of AI to detect health misinformation. Focusing on COVID-19 misinformation in Indonesia, the study employing Qualitative Content Analysis (QCA) to examine the result of AI machines, namely Copilot, ChatGPT, and Gemini, by using specific prompts in two languages (Indonesia and English) which applied in two different times (August and September 2024). This study concludes that the aforementioned AI platforms are capable of detecting misinformation while providing supporting claims to substantiate their reasoning. However, while generative AI has the potential to be utilized as a tool for detecting misinformation, further improvements should be made to refine the output. Furthermore, due to the nature of generative AI learning through deep learning, users who wish to utilize generative AI platforms to debunk hoaxes have to perform additional research to complement their findings.

Keywords: artificial intelligence, pandemic, verifying, health issues, information

1.0 Introduction

Along with the rapid development of AI-generated content and its applicability in various sectors of living, there are also growing concerns about the unwanted effects of AI, particularly in the field of combating misinformation. By following and modelling the structure of fake news or other forms of misinformation, AI has produced tons of misinformation, fake news, and alternate facts effectively throughout every digital platform (Dalkir, 2021). Not only opening up opportunities to create increasingly realistic AI-generated fake content, AI also facilitates the dissemination of disinformation to a targeted audience and at scale by malicious stakeholders (Bontridder & Poulet, 2021). The emergence of threat by AI in the battle of misinformation/disinformation has become an international concern. The term of AI as synthetic media was followed by 'synthetic disinformation' which multiplied the pressure of handling information disorder (Bereskin, 2023).

Despite the moral panic that surfaced along with the emergence (and glorification) of AI, there's considerable effort to use AI to detect the misinformation, and incorporated it as part of prebunking management. The discourse about 'fighting fire with fire' has coined by several prominent members of global world leader, such as Russia President who claimed that 'artificial intelligent is the future' by adopting the logic of AI-powered system as the next step to upscale the battle against politically motive disinformation (Kreps et al., 2022). Since then, effort to model AI as tools for detecting misinformation was growing across multiple platforms and multiple issues. Not only political, the use of AI for detecting misinformation also became part of infodemic management for public health issues (Purnat, 2021).

Technology will naturally continue to develop for as long as humanity exists. It's crucial not to treat AI as an enemy or a replacement for humanity, but rather as a tool to support society. Since generative AI is a relatively new technology, there is much to learn about it. One thing is certain, though: Generative AI has great potential for detecting misinformation because it helps users simplify analysis and reduce the time it takes to sort information. This research aims to help practitioners and researchers interested in integrating AI, or generative AI specifically, into fact-verification tools understand how generative AI works for this purpose and which tool is most suitable.

2.0 Literature Review

Misinformation, as false or inaccurate information, is a communication risk that is as old as the information and communication system itself. While misinformation is unintentional, the problem arises when people believe it and spread the misinformation (Wardle & Derakshan, 2017). Disinformation is another scale of misinformation with the specific intent to create, disseminate, and even promote false, inaccurate, or misleading information that causes public harm for profit (Santos, 2023; Irwansyah, 2024). The deliberate nature of such malicious intent to deceive marked the difference between misinformation and disinformation (Bereskin, 2023). Both misinformation and disinformation had detrimental effects on individuals and society (Xu et.al, 2023).

AI has been shadowing the crusade over truthful information and brought the effort of combatting disinformation to a new level (Nazar & Rustam, 2020). Its ability to generate increasingly realistic AI fake content has facilitated the dissemination of dubious content from textual misinformation to deep fake videos (Bontridder & Poulet, 2021; Kreps et al., 2022). During its initial development, AI generates more textual based

misinformation. However, as the technology expanded rapidly, and becoming more 'user-friendly', the risk of AI generated image and video for producing misinformation have obviously proliferated in such alarming bouts (Fatimah et al., 2024).

On a different playing field, AI also had the capacity to manipulate public opinion by producing synthetic text and generating bots to amplify the misinformation (Dalkir, 2021). It is suspected that during the pandemic of COVID-19, where infodemic term emerged and constitutes the abundance of misinformation on various themes of the disease, AI generated misinformation also playing a significant role in disrupting the health protocol and hampering vaccine acceptance (Dalkir, 2021; Monteith et al., 2024).

In a recent publication on the Global Risk Perception Survey 2023-2024, the World Economic Forum listed AI-generated misinformation/disinformation in the second place (53%). Furthermore, in the next two years, AI-generated misinformation/disinformation is predicted to rank first in global risk severity in both the short and long term. However, in the next 10 years, the severity of AI problems is predicted to be in fifth place, assuming that the technology is able to overcome the main obstacles of AI application at that time.

AI usage in detecting misinformation gained momentum during the pandemic of COVID-19. Over the course of time, a lot of misinformation regarding the pandemic flooded uncontrollably, stirring confusion and eroded trust toward the government and health authorities. This situation is labeled as infodemic, which is signified by excessive information, including false or misleading information, in digital and physical environments during an acute public health event (Purnat et al., 2021). Applying AI to detect COVID-19 misinformation became part of infodemic management

which rely on the fastness of response and the ability to prebunk the misinformation as proposed by inoculation strategies (Akhtar et al., 2022). Earlier effort was marked by using feature extraction applied to train the algorithm developed by machine learning (Khan et al., 2022).

Using AI for combating misinformation or disinformation, however, poses several ethical concerns. First of all, there are bias and fairness challenges resulting from algorithmic bias, content moderation bias, and data selection bias. AI may censor or disproportionately flag certain perspectives, and may inherit bias from previous data training (Omdena, 2024). Second, freedom of expression and censorship. AI might exercise over-censorship by flagging valid content incorrectly, or become under-censorship and miss harmful misinformation. A chilling effect also detected in which AI could potentially abuse free speech (Avey, 2024). Third, judgment and context. Although AI can mimic the route of human thinking, AI still struggles with human nuances and context. Hence, AI may misinterpret satirical or humorous content (Santos, 2023).

Regardless of the risk and how promising design models and tools developed for fact-checking AI look like, the technology cannot do their tasks alone. Generative AI results are characterized as anthropomorphised, or described and characterized as having human traits, by the general public, media and AI researchers. There are tendencies of bias, subjectivity, partial truth, or any symptoms mentioned as 'hallucinations' – referring to undesirable text errors based on LLM (Language Learning Machine) (Monteith, 2024). A collaboration between researchers, private and public sectors are needed exploring what AI can do to overcome the harmful intent of disinformation (Bereskin, 2023; Amnesty International, 2024; Irwansyah, 2024).

3.0 Methodology

This research aimed to explore the possibility of using AI to identify COVID-19 misinformation by utilizing Copilot, ChatGPT, and Gemini. A Qualitative Content Analysis (QCA) focused on the result of each prompt inputting to the designed 'task' is employed. QCA is a research approach for the description and interpretation of textual data using the systematic process of coding (Assarroudi et al., 2018). By nature, qualitative content analysis is more subjective than quantitative content analysis, because QCA attempted to analyze text by employing inductive reasoning—a process of developing conclusions from collected data by weaving together new information into theories or concepts (Bengtsson, 2016). The aim of QCA is identifying themes according to specific contexts which will be further analyzed to analyze the situation being problematized.

To conduct this process, this study designed an experiment by inputting three most popular hoaxes in Indonesia to three AI platforms, namely ChatGPT, Gemini, and Copilot. These platforms are chosen as they are the most popular generative AI tools in Indonesia (Khasanah, 2024). Each generative AI processed similar misinformation by using similar prompts. The result is then clustered into several domains/themes, and analyzed comparatively.

Prior to determining the prompts, this study is prefaced by research to identify potential hoax topics. Upon referring to fact checking sources provided by Mafindo, Cek Fakta, and Kominfo regarding most circulated topics of COVID-19 misinformation in Indonesia during the pandemic (2020-2021), further analysis was conducted to narrow down the topics to three of the most popular hoaxes surrounding COVID-19 at the time, namely garlic as COVID-19 cure, microchip in

COVID-19 vaccine, and oxygen deficiency caused by mask usage.

After determining the topics, the next step is crafting the suitable prompts for testing. The prompts were written in both English and Indonesian, with the latter being selected due to considerations of accessibility for Indonesians, the location upon which this research was based. Each prompt is tested four times in their respective languages to observe consistency. We tested each prompt in two different periods: (1) August 2024 and (2) September 2024. In each period we tested the prompt four times, in order to capture possible different results.

To assess the quality of information from each prompt-testing result, we employed several codes for coding purposes, namely 'topic', 'source', 'bias', and 'timestamp'. Based from the coding process, we clustered the result to do further analyzes in terms of (1) Identifying Main Idea and Supporting Details of each AI-texts; (2) Assessing Source Credibility and Bias; and (3) Critical Issue: Language and Timestamp.

4.0 Result

To answer these research questions, we formulated several prompts that are designed to ascertain whether ChatGPT, Gemini, and Copilot could provide responses that could be utilized to assist the fact-checking process. Prompt creation is essential as it serves as a foundation for communication between users and generative AI models (Knoth et al., 2024). Although generative AI has endless potential, users still have to be able to create a well-crafted prompt to generate outcomes that are still within the context of what they desire to see, as AI's capabilities are reliant on the prompts given to them (Bozkurt & Sharma, 2023). Prompts should also be designed in a way that is able to cater users' individual

contexts so that it could enhance the effectiveness of the AI system (Robertson et al., 2024).

To explain the differences between Copilot, ChatGPT, and Gemini approaches in detecting COVID-19 misinformation through the aforementioned prompts, and the quality of the information provided by those three generative AI based on the 'designed task' as expressed through the 'prompt', the findings were presented based on the categories outlined previously.

4.1 The Structure of AI-Texts

Upon responding to the prompt, each generative AI articulated a similar main idea. There's no difference between the three AIs in identifying the main idea based on the prompt offered to them. The explanations offered by Gemini, ChatGPT, and Copilot use a similar structure, claiming that there is a lack of scientific evidence to substantiate each claim established by the mis/disinformation. ChatGPT offers a longer explanation by describing the hoaxes, starting with an explanation of when the hoaxes first emerged, a description of the hoaxes' narratives themselves, and describing the objects in each hoax and what they were commonly used for (i.e., what is garlic, what is a microchip, and how to wear masks properly). Copilot and Gemini, however, immediately announced the unfounded claim of the hoax and only briefly described the hoax with no technical details.

4.2 Source Credibility and Bias

ChatGPT, Gemini, and Copilot discourage the users from taking the hoaxes face value by adding debunking statements in their prompt result. Gemini and Copilot describe the hoaxes as insubstantial by prefacing their explanations with: "the idea that [...] is simply not supported by scientific evidence" or "the idea that [...] is simply a myth" before delving into their explanation, in which they provided a brief description of the hoaxes, and

a scientific fact to explain why the hoaxes are untrue by providing an excerpt of articles or academic journals.

In ChatGPT, debunking statements are included at the end of the prompt result as a part of the concluding paragraph, in which ChatGPT stated: "The idea that [...] is a conspiracy theory with no basis in reality." Compared to the other platforms, Copilot includes a statement from the World Health Organization in introduction, thus allowing users to locate a verified source to perform fact-checking procedures. In general, the majority of platforms employ a similar set of keywords—such as "not supported by scientific evidence" and "conspiracy theory"—in an attempt to dissuade users from considering these hoaxes an accurate solution to overcome COVID-19. Furthermore, they also encourage the users to seek more information from reputable organizations such as the World Health Organization, or health agencies in the users' respective countries.

4.3 Critical Issue: Language and Timestamp

The success of AI depends on the quality of the data used to train it. This data is designed to produce the best possible results by following the instructions. However, there is a possibility that biases may emerge in the process of creating and managing content, potentially leading to situations where certain content is censored or not censored. The most important thing is the quality of the results, which depend on two different instructions (Indonesian and English) and timestamps. A crucial question to address here is whether the test performs differently when conducted at different times, in which the study discovered that there are no significant differences between Indonesian and English results, except for word length and references. In Copilot, Indonesian results directed to local news articles. Additionally, all platforms encourage users to reach out to

Kementerian Kesehatan Republik Indonesia (Indonesian Ministry of Health) to verify information provided on the platforms.

Although ChatGPT, Gemini, and Copilot have provided useful assistance in relaying information to debunk hoaxes, it should be taken into consideration that generative AI improves its output through deep learning (Imtiaz et al., 2024). It must be acknowledged that the outcomes may be subject to alteration at any given moment. This is due to the intrinsic characteristics of generative AI, which adapts based on the data inputted into its platforms and therefore requires continuous monitoring. However, in this study, the research team conducted two rounds of testing in August and September. Despite the time gap between the two tests, the results demonstrated no significant differences in terms of wording and structure.

Our findings indicate that ChatGPT offers the most detailed information among the three platforms, providing not only a brief account of the hoax's origin but also a list of scientific facts that challenge the claim. Nevertheless, due to its lengthy elaboration, a number of terms utilized by ChatGPT are still regarded as overly technical and potentially challenging for non-medical professionals to comprehend. In comparison to ChatGPT, both Gemini and Copilot provide explanations of a shorter length. It is noteworthy, however, that Copilot is the only platform that offers citations to supplement the generated results.

5.0 Discussion

To verify misinformation/disinformation, people today rely on mass media and fact-checking websites set up for

this purpose. The result and quality of verification depends on several variables: (1) response time; (2) human capital; (3) technology; and (4) non-technological support. Response time refers to attempts to respond to the misinformation as quickly as possible to prevent the escalating effects that are likely to increase the harmful effects of the misinformation. Human capital refers to the quality and competence of fact checkers in performing their task. Technology plays a central role in the fight against misinformation as a supporting tool. Meanwhile, non-technology variables refer to the support of society, law, state and platforms to cooperate and join forces, play their respective roles to maintain the atmosphere of a high integrity information ecosystem as the backbone of democracy or any state system to govern society (Irwansyah, 2024).

The aforementioned variables require significant effort and time to implement. AI offers an opportunity to overcome the threat of slow response by detecting misinformation in a very short time before the outbreak of malicious content. Training full-fledged human fact-checkers requires a lot of capital: time, money, opportunities to upgrade and scale the operation, number of employees, etc. Here, AI could replace human capital, at least to do the hard work, such as scraping tons of suspicious information, which requires human resources (Gifu, 2023). As a technology, the results of this research have shown what AI could potentially be used by laypeople. Nevertheless, the use of AI for early detection of misinformation could be a game changer, as long as the existing support from various sectors is bound to capitalize.

AI operated through the logic of machine learning, meaning that the capability of AI for generating contents depended on the availability of the desirable content, and the 'training' of AI to perform such functions

(Bozkurt & Sharma, 2023). Therefore, similar content prompts were frequently tested to determine whether AI learned in a given period. The results clearly showed that the structure, narrative, and facts of the texts did not differ between the two periods. Two possibilities could explain this: either no additional content on the topic circulating on digital platforms, or the AI platform's limited ability to "learn".

6.0 Conclusion

In general, there is no significant distinction between the performance of Copilot, ChatGPT, and Gemini in detecting false information related to COVID-19. Using the same prompt in each platform will yield a similar result. In terms of wording, there are no significant differences, except with regard to the length of explanations and the selected terms used to describe the information. In comparison to the other two platforms, ChatGPT provides a more structured explanation. Nevertheless, the absence of citations, coupled with the platform's inclination towards technical terminology required users to undertake further research in order to substantiate claims with accurate information. Conversely, although Copilot offers briefer explanations, it also offers links to supporting evidence, enabling users to verify hoaxes and identify the sources of information with greater ease.

To conclude, this study believes that generative AI has the potential to facilitate fact checking efforts. The results of prior experiments indicate that generative AI platforms are capable of providing information by inputting simple commands, which could help reduce the time required to perform the fact-checking process. However, given that generative AI is largely constructed based on users input from prompts and words, it is also

essential to recognize the necessity for adjustments to enhance its performance.

7.0 Limitations and Solutions

As generative AI improves its output by learning through datasets, this study acknowledged that the results may change over time. Furthermore, as language continues to evolve, the output generated during this research period may differ in the future. This study also acknowledged that the results from generative AI are not reliable at all times. It is recommended that individuals or organizations intending to utilize generative AI platforms as a means of debunking misinformation treat the platform as a secondary source of information and conduct their own research by analyzing research or articles from credible news sources, experts, or scientific journals.

Additionally, this study also recommends individuals who wish to utilize generative AI to detect misinformation by continuously training generative AI platforms to provide credible data by entering accurate information to assist with the hoax debunking process. To achieve its maximum performance, AI should be facilitated by producing and circulating 'safe' content as much as possible. As a consequence of this, AI cannot work alone. There should be a supporting system to enable AI to learn and to draw the content.

8.0 Disclosure Statement

The authors report that there are no competing interests to declare.

References

- Akhtar, P., Ghouri, A. M., Khan, H. U. R., Amin ul Haq, M., Awan, U., Zahoor, N., ... & Ashraf, A. (2023). Detecting fake news and disinformation using

- artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of operations research*, 327(2), 633-657.
- Amnesty International. (2024). Unravelling a Murky Network of Spyware Exports To Indonesia a Web of Surveillance.
- Assarroudi, A., Heshmati Nabavi, F., Armat, M. R., Ebadi, A., & Vaismoradi, M. (2018). Directed qualitative content analysis: the description and elaboration of its underpinning methods and data analysis process. *Journal of research in nursing*, 23(1), 42-55.
- Avey, C. (2024). AI Misinformation: Concerns and Prevention Methods. *GlobalSign Blog*. <https://www.globalsign.com/en/blog/ai-misinformation-concerns-and-prevention>
- Bengtsson, M. (2016, January). How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2, 8–14.
- Bereskin, C. (2023). Commonwealth Parliamentary Association. (2023, December). *Parliamentary Handbook On Disinformation, AI and Synthetic Media*. Commonwealth Parliamentary Association (CPA)
- Bontridder, N., & Pouillet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3, e32.
- Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education*, 18(2), i-vii.
- Dalkir, K. (2021). Fake News and AI: Fighting Fire with Fire. In *CEUR Workshop Proc* (Vol. 2942, pp. 112-115).
- Fatimah, R., Mumtaz, A., Fahrezi, F. M., & Zakaria, D. (2024). AI-Generated Misinformation: A Literature Review. *Indonesian Journal of Artificial Intelligence and Data Mining*, 7(2), 241-254.
- Gifu, D. (2023). An intelligent system for detecting fake news. *Procedia Computer Science*, 221, 1058-1065.
- Imtiaz, A., Pathirana, N., Saheel, S., Karunanayaka, K., & Trenado, C. (2024). A Review on the Influence of Deep Learning and Generative AI in the Fashion Industry. *Journal of Future Artificial Intelligence and Technologies*, 1(3), 201-216.
- Irwansyah. (2024). ASEAN Guideline On Management Of Government Information In Combating Fake News and Disinformation In The Media. Ministry of Communications and Informatics Republic of Indonesia.
- Khan, S., Hakak, S., Deepa, N., Prabadevi, B., Dev, K., & Trelova, S. (2022). Detecting covid-19-related fake news using feature extraction. *Frontiers in Public Health*, 9, 1–9. <https://doi.org/10.3389/fpubh.2021.788074>
- Khasanah, U. (2024). *7 AI Terpopuler 2024, Mana yang Paling Banyak Digunakan?*. IDN Times. <https://www.idntimes.com/tech/gadget/ai-terpopuler-2024-00-vs37w-8fchm6>
- Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225.
- Kreps, S., McCain, R. M., & Brundage, M. (2022). All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, 9(1), 104–117. doi:10.1017/XPS.2020.37
- Monteith, S., Glenn, T., Geddes, J. R., Whybrow, P. C., Achtyes, E., & Bauer, M. (2024). Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2), 33-35.
- Nazar, S., & Bustam, M. R. (2020). Artificial intelligence and new level of fake news. In *IOP conference series: materials science and engineering* (Vol. 879, No. 1, p. 012006). IOP Publishing.

- Purnat, T. D., Vacca, P., Czerniak, C., Ball, S., Burzo, S., Zecchin, T., Wright, A., Bezbaruah, S., Tanggol, F., Dubé, È., Labbé, F., Dionne, M., Lamichhane, J., Mahajan, A., Briand, S., & Nguyen, T. (2021). Infodemic signal detection during the COVID-19 pandemic: Development of a methodology for identifying potential information voids in online conversations. *JMIR Infodemiology*, 1(1). <https://doi.org/10.2196/30971>
- Robertson, J., Ferreira, C., Botha, E., & Oosthuizen, K. (2024). Game changers: A generative AI prompt protocol to enhance human-AI knowledge co-construction. *Business Horizons*, 67(5), 499-510.
- Santos, F. C. C. (2023). Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis. *Journalism and Media* 4, 2 (2023), 679–687.
- The Ethical Role of AI in Media: Combating Misinformation. (2024). Omdena. <https://www.omdena.com/blog/the-ethical-role-of-ai-in-media-combating-misinformation>
- Xu, D., Fan, S., & Kankanhalli, M. (2023). Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9291-9298).