AI-Driven Detection and Mitigation of Misinformation Spread in Generated Content

[1]Mr.Sidharth Sharma

[1] Manager – Forensic Investigation, American Internation Group (AIG). Inc, 80 Pine St New York, NY 10005 - US

Abstract. Misinformation has been a persistent and detrimental phenomenon in our society in many ways, including individuals' physical health and economic security. With the advent of short video platforms and associated applications, dissemination of multi-modal misinformation, including images, texts, audios, and videos, have increased these issues. The advent of generative AI models such as ChatGPT and Stable Diffusion has further enhanced the complexity, providing give rise to Artificial Intelligence Generated Content (AIGC) and posing new challenges in the detection and mitigation of misinformation. Therefore, the conventional methods of misinformation detection and intervention have proved to be insufficient in this new paradigm. This paper examines the challenge posed by AIGC in the context of misinformation. It analyzes the problem from a psychological and societal point of view, and investigates the subtle manipulation traces in AIGC at signal, perceptual, semantic, and human levels. Through examining manipulation traces like signal manipulation, semantic inconsistency, logical incoherence, and psychological strategies, our goal is to address AI-generated misinformation and present a conceptual design of systematic explainable solution. Finally, we hope this paper will provide useful insights into the fight against misinformation, especially in the age of AIGC.

Keywords. AI- Generated Content (AIGC), Multimodal, Generative AI Models Misinformation Detection.

1. INTRODUCTION

The emergence of Large Language Models (LLMs) has marked a significant shift in the field of natural language processing, enabling machines to produce text that closely matches the intricacy and cohesion of content created by humans. Models like GPT-3 and Deberta have shown exceptional skill in activities from translating languages to crafting creative content, making it increasingly challenging to distinguish between text produced by humans and machines. While these advancements have opened new frontiers in AI research and application, they have also raised profound questions regarding the authenticity and trustworthiness of the generated content.

The proliferation of AI-generated text has significant implications across various domains, including journalism, social media, education, and business. However, alongside the potential benefits come inherent risks, manipulation of public opinion, and the erosion of trust in digital communication channels. Addressing these challenges requires robust mechanisms for distinguishing between AI-generated and human generated content, a task that remains inherently complex due to the evolving nature of AI technologies. Detecting AI-generated text has emerged as a pressing research area, driven by the imperative to safeguard against the misuse of AI and preserve the integrity of online discourse.

Traditional approaches to text classification, such as TF-IDF, have long been foundational in natural language processing, providing information on the term distribution within a corpus. These approaches are likely to fail, though, when faced with the sophisticated linguistic subtleties typical of AI-generated text. To counter these limitations, our work suggests a new hybrid solution that combines conventional feature extraction methods. By taking advantage of the strengths of TF IDF complemented by sophisticated algorithms like Bayesian classifiers, Stochastic Gradient Descent (SGD), Categorical Gradient Boosting (CatBoost), and the highly Deberta-v3-large models, our approach is designed to attain unparalleled precision in differentiating between text produced by AI and that created by humans. In this paper, we outline our methodology, present our experimental findings, and draw conclusions regarding the efficacy of our proposed approach. Through extensive experimentation on a diverse dataset comprising both human and AI-generated text samples, we demonstrate the superiority of our method in accurately discerning between the two. By advancing AI-generated text detection techniques, our research seeks to mitigate the risks associated with the proliferation of AI generated content and foster trust in digital communication platforms.

2. LITERATURE SURVEY

Throughout history, misinformation1 has had a negative effect on people and societies (Fig. 1). For instance, in South Africa between 2000 and 2005, misinformation that HIV was not linked to AIDS led to misdirected policies and about 330,000 excess deaths [1]; during the 2016 presidential election campaign, a staggering 126 million Americans were exposed to politically oriented misinformation [2].The development of social media and short video platforms has energized the mass spread of multi-modal misinformation. New generative AI models, like 'deepfakes', need little or no training, lowering the cost of creating multi-modal misinformation [3]. The models facilitate the manipulation and creation of various digital media content according to particular instructions, e.g., Stable Diffusion creates images from text prompts [4]. The spread of Artificial Intelligence Generated Content (AIGC) is a serious challenge in the fight against misinformation [5].

Misinformation detection has been widely researched, encompassing text-based, visual, and audio disinformation [6,7]. The early methods employed text content, writing style, watermarks, and types of manual features. For visual features, artifacts, camera fingerprints, and biological signals have been used [8]. Audio features such as MFCC [9] and spectrogram [10] have been used. Although advancements, earlier models were not explainable and scalable for general multi-modal disinformation [11]. Machines cannot comprehend manipulative tactics or ploys used by manipulators, and models based on small datasets cannot evaluate the semantic truthfulness of AIGC. Successful counteraction of future disinformation needs comprehension of highly sophisticated deception and application of rich background knowledge. To overcome these issues, this work presents a comprehensive survey of new generative AI models and their possible applications for misinformation. We also discuss future challenges of AIGC proliferation and the psychological outlook of drivers of misinformation and consumer behavior. According to the above, we put forward a conceptual design of multi-modal misinformation detection architecture specifically suited for the AIGC era.

Our approach suggests a cascade of detection mechanisms to cater to a broad range of fabrication operations, including human-editing, AI-manipulation and human propagation. We classify the deceptive traces into four layers: signal, perceptual, semantic, and human (psychology), where "signal" and "perceptual" are associated with low-level inconsistencies, "semantic" with logical inconsistencies, and "human" with behavioral psychology. The designed architecture is intended to thoroughly detect multi-modal misinformation with an explainable model, offering a strong defense against the multi-faceted threats of misinformation using AIGC. The contribution of this work is in three folds:

• We examine the possible misinformation scenarios that can arise with the participation of AIGCs. This entails investigating the possible uses of generative AI models in misinformation creation, examining the features of misinformation in the future, and discussing both challenges and opportunities of countermeasures in the AIGC era.

• We clarify social science factors influencing misinformation, such as confirmation bias and social proofing, and provide insightful recommendations for the fight against misinformation from a multi-disciplinary perspective.

• Our proposal is a complete multi-modal misinformation detection system that examines manipulation traces – signal and perceptual genuineness, semantic trustworthiness, and behavioral psychology hints — across the whole fabrication process, with an emphasis on explainability and scalability in the AIGC era

FIGURE 1. A synopsis of false knowledge throughout human history.

Left: Since ancient Rome, misinformation has been a part of human communication. Octavian began a "fake news" campaign in 44–30 B.C. against Anthony Mark. Coins with slogans that portrayed Antony as a drunk and a womanizer helped Octavian defeat him. Middle: In 1835, the New York newspaper The Sun published a story titled "The Great Moon Hoax," which asserted that an extraterrestrial society existed on the moon. Correct: The US stock market saw a brief sell-off on May 22, 2023, in response to an AI-generated depiction of the Pentagon explosion.

3. PROPOSED SYSTEM

During the process of misinformation generation and dissemination, we identify three common elements: techniques, misleading contents, and propagation processes, each of which leaves behind discernible fabrication traces. 1) Techniques: Despite the technical development of generative models, AIGCs still exhibit inherent technical flaws. Current large generative models rely on data training on collected digitized samples, lacking features of physical signals in the recording and transmitting process. Besides, the digital samples are less complex than real world, with limited views, dimensions and length. Therefore, AI lack of overall understanding of objects from the multiple perspectives, e.g., it is not easy for AI to learn physical laws from these simplified samples. Therefore, AIGC often have perceptual artifacts in realism and naturalism. Hence, we propose analyzing the above technique flaws at two levels: the signal level and the perceptual level. 2) Misleading contents: Diverse misleading contents need understanding of both content and background knowledge. Therefore, we propose semantic level for misleading contents. 3) Propagation operations: Our analyses throw up four psychological features of users, which introduces new insights into propagation operations. Building on these findings, we propose psychological cues as human-level features.

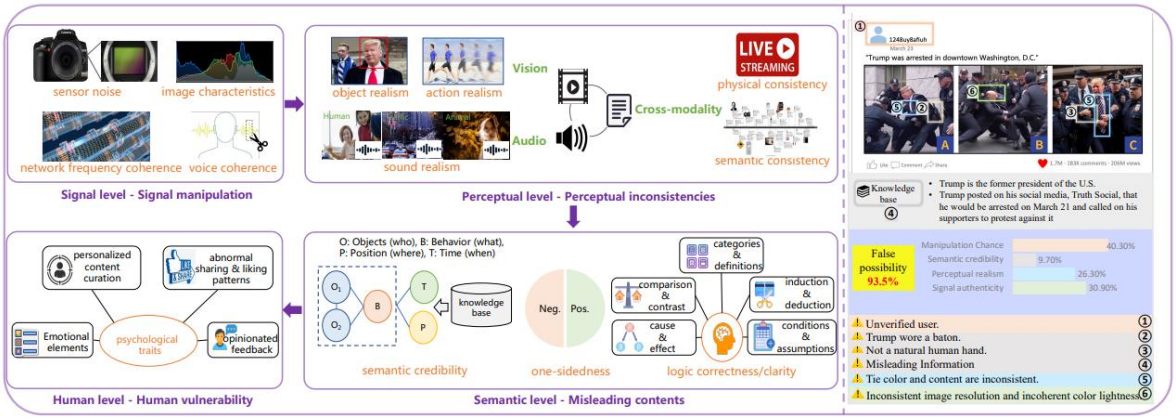

Figure 2. A visual representation of the suggested conceptual framework for the identification of multi-modal disinformation.

Four levels of fabrication traces are extracted by the framework: signal, perceptual, semantic, and human. Interface Conceptual Prototype (b). In the images, the four color boxes represent the four levels of misinformation characteristics; the total false possibility score, the specific scores of the four levels of misinformation characteristics, explanations, and reference knowledge are exhibited.

Signal-level

In the process of collecting, digitally transforming, and transmitting authentic visual and audio content, various types of traces are formed, such as camera sensor noise [13] and electrical network frequency [12]. These traces are absent in AIGCs. Additionally, natural contents exhibit certain signal characteristics, such as continuous lightness and sharpness in images, and coherent MFCC, tone, and sound quality in voices (both main sounds and background sounds).
However, in AIGCs or manipulated contents, these natural characteristics may be absent or distorted. Therefore, these changes in signal characteristics can be utilized to represent signal authenticity — one key attribute of misinformation.

Perceptual-level

Authentic content should be easily comprehensible to humans. Conversely, content that is\ unrecognizable or misleading is often the result of a potential AI generation. At the perceptual level, the assessment of content realism extends to both uni- and multi-modal contexts. On uni modality, visual realism involves the realism of object appearance and movement in both spatial and temporal dimensions, audio realism entails that each audio composition corresponds a physical sound in the aspects of the physical properties (e.g., 'timbre') and semantic properties (e.g., language). On cross modality, realism refers to the cross-modal consistency. For narrative multi-modal content, it means semantic consistency, e.g., an image showing a 'bird' is described as 'plane' in text. While for live videos it also includes physical consistency, which means the sound and vision should belong to the same subjects. For example, the mouth shapes and utterances should be synchronous, and different speaker should have voice different timbre. In summary, the perceptual level rules out the unrealistic contents in physical world.

Semantic-level

Authenticity of multimedia information at the signal and perceptual levels presents a primary challenge. Manipulation of information is still the most important distinguishing point between accurate information and misinformation. Manipulators can create events or mix up existing events in deceptive manners. Therefore, identification of created events and nonsensical relationships becomes imperative.

Moreover, filtration and selective presentation of information for the advantage of some groups can create misinformation by encouraging one-sidedness. At the semantic level, we judge the believability of each item of information against a human knowledge base and judge the logical consistency and contextual plausibility. In addition, we judge the overall one-sidedness of the information on the basis of subjective expression and scope of coverage.

4. CONCLUSION

The problems of disinformation in the AIGC era are examined in this research. In order to address the upcoming problems of massively created disinformation and the influence of human psychology, we present the conceptual framework for multi-modal misinformation detection. Our design paradigm encompasses the full process of disinformation manipulation and functions on four levels: signal, perceptual, semantic, and human. It seeks to close the current gap by offering thorough, intelligible, and real-time detection methods. Finally, there is a never-ending game of cat and mouse between those who spread false information and those who catch it. To keep ahead, distributors change their strategies while detectors improve their methods. Regulation, legislation, and education are crucial in holding bad actors accountable and equipping people with media literacy and critical thinking skills, even while technical solutions are crucial.

REFERENCES

1. Hunt, E. B. (2014). *Artificial intelligence*. Academic Press.
2. Holmes, J., Sacchi, L., & Bellazzi, R. (2004). Artificial intelligence in medicine. *Ann R Coll Surg Engl*, *86*, 334-8.
3. Winston, P. H. (1992). *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc..
4. Winston, P. H. (1984). *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc..
5. Boden, M. A. (Ed.). (1996). *Artificial intelligence*. Elsevier.
6. Thepade, D. S., Mandal, P. R., & Jadhav, S. (2015). Performance Comparison of Novel Iris Recognition Techniques Using Partial Energies of Transformed Iris Images and Enegy CompactionWith Hybrid Wavelet Transforms. In *Annual IEEE India Conference (INDICON)*.