An Enhanced Multimodal Multilingual Dataset for Medical Misinformation Detection

Zhaoyi Sun
Biomedical Informatics and Medical Education
University of Washington
Seattle, WA
zhaoyis@uw.edu

Yujuan Fu
Biomedical Informatics and Medical Education
University of Washington
Seattle, WA
velvinfu@uw.edu

Wen-wai Yim
Health AI
Microsoft
Redmond, WA
yimwenwai@microsoft.com

Meliha Yetisgen
Biomedical Informatics and Medical Education
University of Washington
Seattle, WA
melihay@uw.edu

Fei Xia
Department of Linguistics
University of Washington
Seattle, WA
fxia@uw.edu

***Abstract*—Medical misinformation significantly challenges public health and information integrity, yet there are many types and cases of misinformation not currently studied. This study introduces a comprehensive dataset designed to advance the detection and analysis of medical misinformation across different sources, modalities, and languages. Our work extends the investigation into non-traditional platforms like podcasts, blogs, and advertisements, and addresses subtler forms of misinformation, such as AI-generated documents or ambiguous content that may mislead without being directly false. Our work include multimodal (text and visual) and multilingual (Chinese and English) contents. We will also evaluate the performance of state-of-the-art (SOTA) models against this dataset.**

***Index Terms*—Medical Misinformation, Multimodal Multilingual Dataset, Large Language Models, Misinformation Detection**

I. Introduction

Medical misinformation refers to false or misleading health-related information that is spread without regard to its accuracy, often leading to harmful consequences [1]. Despite its impact, medical misinformation can encompass a large variety of use cases beyond strict factuality, leading to challenges in consistently identifying misinformation and developing automate detection systems to help health consumers. Moreover, advanced artificial intelligence systems, such as large language models (LLMs), can generate sophisticated misinformation that is often more difficult for humans to detect than misinformation written by people with similar semantics [2]. Current efforts to track medical misinformation primarily focus on text-based content in singular languages, overlooking the significant role visual content and multilingual diversity play in spreading misinformation. [3]. This situation underscores the need for a more inclusive approach that considers both the multimodal (text and visual) and multilingual dimensions of misinformation to better understand and tackle its spread.

Several multimodal medical misinformation datasets exist [4]. However, most datasets focus primarily on news and tweets (particularly related to COVID), overlooking other sources such as podcasts, blogs and advertisements. Linguistically, these datasets frequently feature overly simplistic or obvious examples of misinformation. The misinformation generated by LLMs typically involves direct substitution at the sentence level, rendering the task unnaturally easy. Furthermore, previous studies have demonstrated limited efforts in incorporating multiple languages.

We build upon previous work by making several key contributions: (1) We broaden the scope of medical misinformation sources to include not only news articles and social media but also podcasts, blogs, and advertisements. (2) Our dataset focuses on subtler forms of misinformation, including LLM-generated documents and ambiguous content that might mislead without being outright incorrect, such as promotional or political material. (3) We incorporate content in multiple languages, specifically English and Chinese, to support more comprehensive global research on medical misinformation.

II. Specific Aims

Aim 1 - Characterize medical misinformation from multiple sources and develop an annotation schema. To bridge the gap between existing research limitations and the necessity for a comprehensive understanding of misinformation's scope, we will systematically identify various forms of misinformation, isolate particularly significant cases, and develop an annotation schema for these instances.

Aim 2 - Create a multimodal and multilingual medical misinformation dataset. Based on the annotation schema in Aim 1, we will collect data on a wide range of topics from English and Chinese news articles, social media posts, podcasts, blogs, advertisements, and LLM-generated texts. The modalities will involve text, images, and videos.

Aim 3 - Evaluate the effectiveness of SOTA models in detecting medical misinformation. We will enhance our evaluation of SOTA models on the dataset generated in Aim 2 by not only considering text but also incorporating the analysis of images and videos. The analysis of text misinformation will be conducted at various levels of granularity, from sentences to entire documents.

References

[1] W.-Y. S. Chou, A. Oh, and W. M. P. Klein, "Addressing Health-Related Misinformation on Social Media," *JAMA*, vol. 320, no. 23, pp. 2417–2418, Dec. 2018.

[2] C. Chen and K. Shu, "Can LLM-Generated Misinformation Be Detected?," *arXiv* [cs.CL], Sep. 25, 2023.

[3] K. Heley, A. Gaysynsky, and A. J. King, "Missing the Bigger Picture: The Need for More Research on Visual Health Misinformation," *Sci. Commun.*, vol. 44, no. 4, pp. 514–527, Aug. 2022.

[4] Y. Sun, J. He, S. Lei, L. Cui, and C.-T. Lu, "Med-MMHL: A Multi-Modal Dataset for Detecting Human- and LLM-Generated Misinformation in the Medical Domain," *arXiv* [cs.SI], Jun. 15, 2023.