Enhancing Health Communication:
Developing LLM-based Models to Generate Effective Counter Speech Against Misinformation

Kevin Gautier[1], Anirban Saha Anik[2], Sriram Kandadai[1], Xiaoying Song[2], Lingzi Hong (Mentor)[2]
[1]Department of Computer Science and Engineering, University of North Texas, [2]Department of Information Science, University of North Texas

Abstract

The current internet ecosystem allows for a breeding ground of health misinformation. This health misinformation could potentially cause susceptible readers to make choices that could be detrimental for their health. Counter speech to health misinformation presents verified information that provides context, clarity, or correction to the original statements made. The goal of this project is to train an LLM model to detect health misinformation and generate counter speech automatically. The trained model can then be used to help alleviate the work of moderators and community members. We have completed the first step of this project, identifying effective counter speech to health misinformation. This will be the foundation for the next steps of the project, which will include modeling the effectiveness of the counter speech and developing the LLM-based generation of counter speech.

Methods

- Data collection
  - Used the Praw api in to scrape posts and comments from Reddit
  - Found health-related posts using keyword filtering
- Human Labeling
  - Reddit posts were labeled by 3 individuals independently as “Health Misinformation” or “Not Health Misinformation”
  - Reddit comments were labeled as “Counter speech” or “Not Counter speech
- LLM Labeling
  - Used the Llama3 model labeling posts as health misinformation and comments as counter speech

References

1. Yu, Xinchen, Eduardo Blanco, and Lingzi Hong. "Hate speech and counter speech detection: Conversational context does matter." arXiv preprint arXiv:2206.06423 (2022).
2. Hong, Lingzi, et al. "Outcome-Constrained Large Language Models for Countering Hate Speech." *arXiv preprint arXiv:2403.17146* (2024).
3. Suarez-Lledo, Victor, and Javier Alvarez-Galvez. “Prevalence of Health Misinformation on Social Media: Systematic Review.” *Journal of medical Internet research* vol. 23,1 e17187. 20 Jan. 2021, doi:10.2196/17187
4. Islam, M.R., Liu, S., Wang, X. et al. Deep learning for misinformation detection on online social networks: a survey and new perspectives. Soc. Netw. Anal. Min. 10, 82 (2020). https://doi.org/10.1007/s13278-020-00696-x
5. Krishna, A., & Thompson, T. L. (2021). Misinformation About Health: A Review of Health Communication and Misinformation Scholarship. American Behavioral Scientist, 65(2), 316-332. https://doi.org/10.1177/0002764219878223

Project Outline

Step 1

Identifying effective counter speech to health misinformation

We first located health misinformation posts on Reddit, then collected all the comments from these posts. For the comments, we first manually labeled them as counter speech or not, then will build LLM-based classifiers for the automatic identification of counter speech.

Step 2

Modeling the effectiveness of counter speech

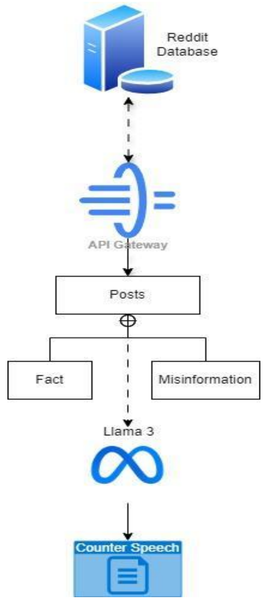
We will then collect the user interactions and follow-up conversations of counter speech to health misinformation. Using these interactions, we automatically assess the effectiveness of various counter speech.

Step 3

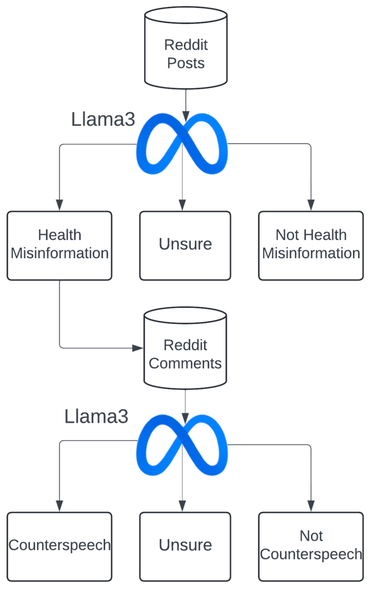
LLM-based generation of effective counter speech

We will then develop LLM models using fine tuning and reinforcement learning to incorporate the effectiveness into the counter speech generation process.

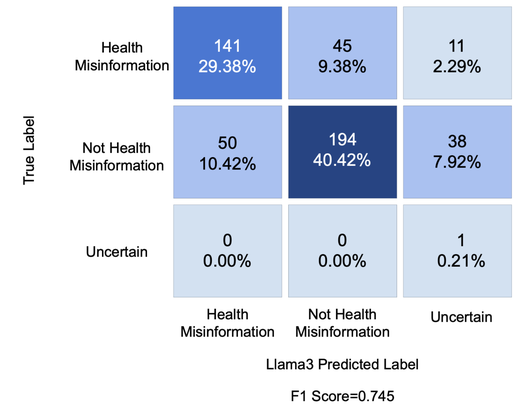
System Design

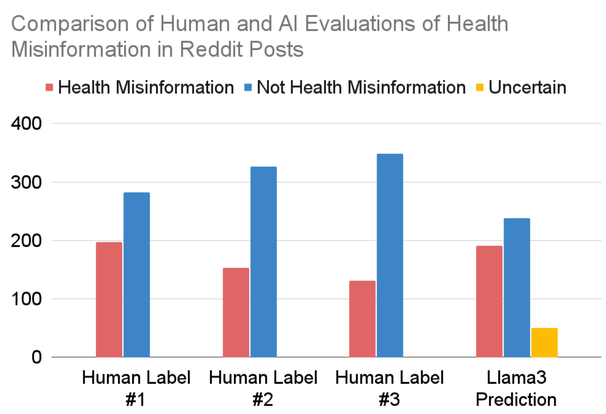


LLM Pipeline



Results





Conclusion

Using the collected Reddit data, we were able to build a starting dataset that will be used for LLM training. This is only the first step in the project, where we focused on collecting and identifying counters speech to health misinformation. The future of this project will focus on ensuring the labeled data is reliable, and determining what is effective counter speech to health misinformation

Acknowledgements

We extend our sincere gratitude to the Department of Information Science at UNT for making this project possible. Special thanks to our advisor, Dr. Lingzi Hong, for her invaluable guidance, mentorship, and support in facilitating this research. We also acknowledge the financial support from the Institute of Museum and Library Services under Grant Numbers LG-256661-OLS-24 and LG-256666-OLS-24.