NAPCRG 52nd Annual Meeting — Abstracts of Completed Research 2024.

Submission Id: 7035

Title

Development of a ChatGPT-Powered Crisis and Suicidal Ideation Management Module for an HIV Self-Management Chatbot

Priority 1 (Research Category)

Infectious Diseases (not respiratory tract)

Presenters

Sebastian Villanueva, Yuanchao MA, MSc, Bertrand Lebouché, MD, PhD, David Lessard, PhD

Abstract

Background: MARVIN is a chatbot promoting self-management of medication among people with HIV (PWH) and currently providing support for medication adherence and HIV-related general knowledge. PWH face significant stigma and mental health challenges (e.g., stress, depression, and isolation), and are at greater risk of distress than the general population.

Objectives: To enhance MARVIN's ability to respond to users' potential 'extreme messaging' (e.g. messages suggesting suicidal thoughts or including self-harming or discriminatory components). Sub-objectives are to 1) develop a classification algorithm of extreme messaging, 2) integrate this algorithm into MARVIN, and 3) assess its usability and impacts on patient-chatbot communication, including conversation clarity and user satisfaction.

Methods: We integrated three public hate speech databases from an online catalog (hatespeechdata.com) mixed with recorded MARVIN-user conversation dataset and MARVIN synthetic data (e.g. permutations of our own data), to train ChatGPT to identify three categories of messages: self-harm, insults, and normal (e.g., messages which intents are not self-harm or insult). We prompt-tuned the model and tested its performance with different prompts and the one-shot prompting technique. We quantified its performance with recall, precision, accuracy, and F1 score metrics. We then integrated ChatGPT into MARVIN's classification tree through requests to the OpenAI servers. We constructed a 2-hour test guide of 14 scenarios and asked six testers, including three PWH and three professionals (e.g. two engineers and one doctor), to do each scenario, and then to fill out a two-item on conversation clarity and user satisfaction.

Results: Using one-shot prompting and a prompt using the CO-START framework, ChatGPT attained a value of 96% across all the metrics. The hybrid ChatGPT-MARVIN model successfully generated appropriate responses to extreme messaging with emergency contacts, while relying on MARVIN's original responses for messages with 'normal' intentions. When applying scenarios, testers considered MARVIN's responses as clear and concise, and were satisfied with the responses and overall experience.

Conclusion: This 'anti-extreme module' represents the first filter for MARVIN. Large language models such as ChatGPT can help identify extreme intentions like self-harm and insults and to generate appropriate responses for healthcare chatbots. Providing users with non-judgmental support

Downloaded from the Annals of Family Medicine website at www.AnnFamMed.org.Copyright © 2024 Annals of Family Medicine, Inc. For the private, noncommercial use of one individual user of the Web site. All other rights reserved. Contact copyrights@aafp.org for copyright questions and/or permission requests.