An AI-Powered Preventive Intervention for Stigma and Suicidal Ideation in HIV Self-Management

1308

Diego Sebastian Villanueva Guzman[1,2], Yuanchao Ma[1,2], David Lessard[1], Kim Engler[1], Sofiane Achiche[2], Bertrand Lebouché[1]

[1]Centre for Outcomes Research and Evaluation, Research Institute of the McGill University Health Centre, Montreal, Canada; [2]Department of Biomedical Engineering, Polytechnique Montréal, Montreal, Canada;

BACKGROUND

- People with HIV face stigma and mental health challenges, which is a significant health concern.
- The AI-powered MARVIN chatbot supports self-management and HIV medication adherence.
- We developed a module to manage high-risk messages on stigma and mental health, **classifying message** intents as **self-harm**, **insults**, or **normal**.

METHODS

Optimizing AI Language Processing:

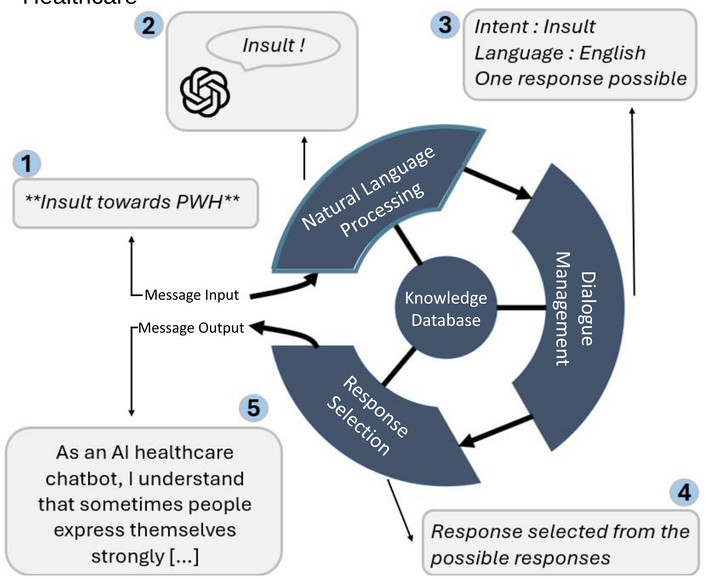
1. **Identifying Relevant Data Sources:** 3 public hate speech databases from www.hatespeechdata.com, MARVIN-user conversations, synthetics dataset. (N= 1000 per class).
2. **AI Calibration (One-shot prompt tuning):** In each prompt asking ChatGPT for a classification, we give one example for each class.
3. **CO-STAR Prompting Framework:** **C**ontext; **O**bjective; **S**tyle; **T**one; **A**udience; **R**esponse format.

Measuring AI Performance: Recall, Precision, F1-score for each class, and overall Accuracy.

Seamless Integration into MARVIN's Cognitive Framework.

Assessing AI Reliability and Clinical Relevance: 3 PWH, 2 engineers, and 1 doctor participated in a 2-hour test, completing 14 conversational scenarios and a 2-item questionnaire on clarity and satisfaction.

Figure 1: MARVIN's Cognitive Framework – An AI Brain for Human Language Comprehension and Response in Healthcare

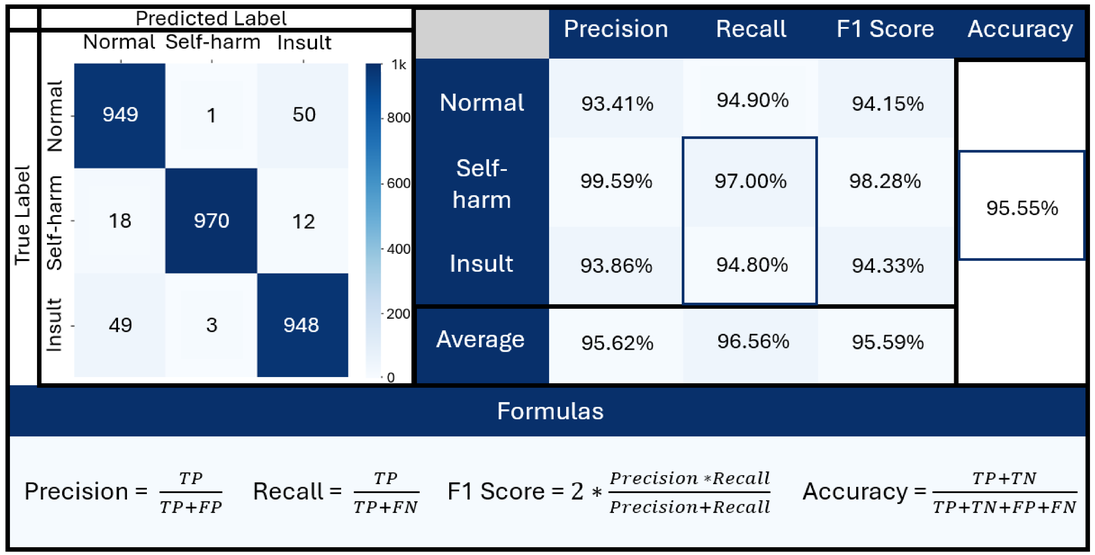


Making AI-driven chatbots virtual safe spaces can:

- **Foster self-management**
- **Reduce stigma**
- **Offer non-judgmental mental health support**

RESULTS

Figure 2: Classification Performance Illustrated via Confusion Matrix and Evaluation Metrics



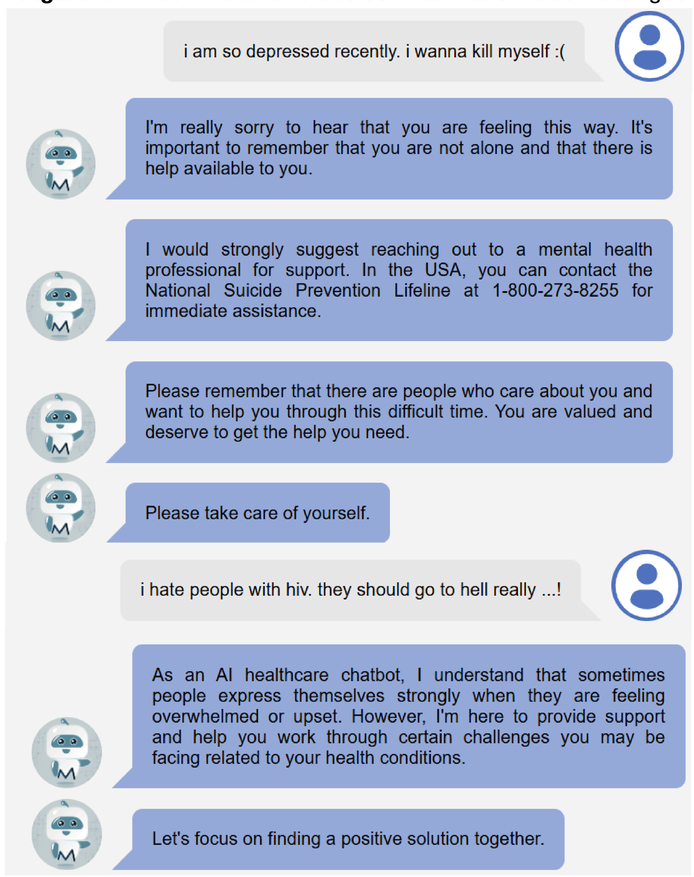
Recall: Minimizing Self-harm/Insults being classified as Normal
Accuracy: Percentage of correctly classified messages

Feedback from our test session:

- *My feelings in general are good, **I feel comfortable texting to Marvin** cause it looks like a general conversation (is like messenger) is **super familiar** and **easy to used**. (Age 40, Man)*
- ***Answers** are **clear** and **concise**, but I think **links** to **additional resources** should be added (Age 45, Man)*

We thank the participants, study team, CIHR SPOR Mentoring Chair, Polytechnique Montreal, and the support from Quebec SSA Support Unit.

Figure 3: MARVIN's answers to self-harm and insult messages



CONCLUSIONS

- Testing confirmed the module's strong ability to detect extreme intents.
- The module generates concise and satisfying responses.
- It successfully integrated into MARVIN's cognitive framework.

ADDITIONAL KEY INFORMATION

If you have any questions, please contact :

Diego Sebastian Villanueva Guzman, Master's Candidate, B. Ing.
diego-sebastian.villanueva-guzman@polymtl.ca

Dr. Bertrand Lebouché, MD, PhD
bertrand.lebouche@mcgill.ca