Usability Study Design

Impact of Interface Design on Reliability of Health Information

Shreeti Shrestha
Northeastern University

CS 6983: Research in Human-Centered NLP
Instructor: Hye Sun Yun
Date Submitted: March 14, 2025

Overview

- Objective: To design an evaluation to compare the usability and perceived trustworthiness of two chatbot interfaces designed to help users find reliable health information.
- Control: A text-based chatbot interface that provides answers without citing the source.
- Experiment: A text-based chatbot interface that provides answers with cited sources (e.g., hyperlinks to reputable health websites).

Evaluation Design

Target Participants

- The participant pool would be a mix of people from different pools. It could be broken down into a 50/50 proportion of people with/without medical background:

User Group	**Proportion (%)**
General Public (non-experts)	35
Patients (non-experts)	15
Medical Students (experts)	35
Medical Professionals (experts)	15

- Majority of the participants would be general people, which would help assess reliance
- and trustworthiness of non-medical everyday users in chatbots to look up health information.
- We'd also recruit a small proportion of patients dealing with medical issues, since going through a medical condition might make them interested and more engaged in interacting with a chatbot to understand, and/or relate their condition and provide valuable feedback based on their own experiences.
- A decent proportion would be people who have a background in medicine. These participants would be essential to verify the correctness of the information displayed by the chatbots. Majority of this population would be recruited from medical students who might be easier to access in terms of time and financial resources compared to medical professionals.

Metrics/Measures

- Ease of Use: Using System Usability Scale (SUS) in post-use survey for assessing participants' perception of interface usability for each chatbot
- Perceived Trustworthiness: Using Credibility Perception Scale in post-use survey for assessing participants' trust in results generated by each chatbot
- Interaction Frequency and Duration: The interactions with each chatbot would be recorded to track the time spent on each chatbot for each question and follow up interactions.

Methodology/Protocol

- Recruitment: Participants could be recruited through advertisements, online/social media platforms, medical school mailing lists, and health forums, and recruited based on a questionnaire to categorize them into groups based on their medical/non-medical backgrounds.
- The study protocol would be as follows:
    1. The experiment can be done anywhere in a quiet environment.
    2. Participants will be given a brief introduction to the study whereby they will be informed that the purpose would be to assess usability and trustworthiness of chatbots when it comes to accessing health information. They will not be informed of the control or the experiment sources.
    3. Each participant will then be assigned one of the chatbots at random to interact with.
    4. Participants will be given a set of health-related queries to search using the chatbot.
    5. After the interaction, they will fill out a post-task survey with questions to assess trustworthiness and usability.
    6. After a short break, they will be assigned the second chatbot that they haven't previously interacted with, to follow steps 4 and 5 with this new chatbot.
    7. Once they're done with the second survey, in a short debrief interview, participants are asked which of the two chatbots they'd prefer to use for any health-related questions they might have. After the debrief, and thanking them for their time, the study is concluded

Experimental Design

- Here, we go for a split-plot design, given the independent variables:
    - The chatbot source (control, experiment)
    - User group background (non-medical people, medical people)
    - **Within-subjects:** Each participant in the study interacts with both the chatbots (control and experiment), allowing for direct comparison between the 2 interfaces

- **Between-groups:** The participant pool is stratified into 50% non-experts (general/patients) and 50% professionals (medical students/experts) to assess perceived differences based on knowledge and expertise.

Rationale for Design Choices

- A split-plot design provides the benefits of a within-group design that include a smaller sample size and effective isolation of individual differences, while also capturing the impact of expertise levels.
- Self-report surveys provide direct feedback for usability and trustworthiness. An individual survey after each interaction offers immediate isolated perceptions, compared to a single survey at the end, by which the user might have forgotten their impressions of the first interaction.
- Having a same set list of health questions to guide user interactions helps standardizing comparisons, where leaving interactions open ended could result in a variety of queries, making comparisons harder if users ask very different questions.

Analysis Methods

Main Hypothesis:

- Is there a significant difference in usability and trustworthiness between the two chatbots?
- The null hypothesis (H0) assumes no significant difference in perceived trustworthiness and usability between the chatbots, and an alpha level of 0.05 will be used to determine statistical significance (p-value < 0.05).

Statistical Analysis

- Perceived usability: Which of the two is easier to use?
  - System Usability Scale (SUS)
  - We could conduct a paired t-test (if scores are normalized) or a Wilcoxon signed-rank test to compare mean SUS scores between the 2 chatbots.
- Perceived trustworthiness: Are chatbots with cited responses more trustworthy?
  - Credibility Perception Scale (CPS)
  - We could conduct a paired t-test (if scores are normalized) or a Wilcoxon signed-rank test to compare mean CPS scores between the 2 chatbots.

The p-values from these tests will determine whether or not the 2 chatbots are statistically different in perceived usability and trustworthiness.

Additionally, if the data is normalized, we could do a mixed-design ANOVA test. The within-subjects factor would be the chatbot type and the between-subjects factor would be the participant group. The Credibility Perception Scale scores would be the dependent variable for ANOVA test, to answer the following questions:

- Do users trust one chatbot over the other overall?
- Do medical experts rate trustworthiness differently than non-medical people?

Qualitative Analysis

- The responses from the final question in which participants are asked which chatbot they would prefer to use and why, could be analyzed to validate the results from statistical analysis. These open-ended responses could be categorized into themes of engagement, credibility, ease of use, and satisfaction, that could provide additional information about perceived usability and trustworthiness generated by statistical analysis.

Expected Challenges

- Participants might trust or prefer the chatbot with citations due to pre-existing biases.
    - Solution: Not mentioning that one is cited, and the other is not. It remains implicit for the purposes of the study.
- Fatigue and disengagement of having to fill out individual surveys post each interaction.
    - Solution: Instead of having individual questionnaires for usability (SUS), trustworthiness (CPS), compile a comprehensive questionnaire to generate usability and trustworthiness scores. A break in between each interaction might be helpful.
- Set of health questions/conditions might be unfamiliar to non-medical people, making it harder for them to understand the responses, and thus, negatively affecting notions of usability.
    - Solution: Compile a mix of common health questions that are easy to comprehend by daily people and questions for specific conditions that can be verified by experts.