CS484/684 Group Project 4

Reddit Health Advice ChecKer Bot (HACKBot)

Natalie Wang nwang42, Kritika Sharma ksharm19,
Gwenyth Portillo Wightman gportil2, Niharika Desaraju ndesara2

1 Introduction

Problem Definition Reddit, the ninth most popular website in the United States [8], boasts a large, engaged user base on its medical advice forums ("subreddits"), where users can crowd-source free medical opinions. Users post questions asking about health issues (e.g., "My hair is falling out"), and other Reddit users respond with medical advice (e.g., "Take X supplement"). However, much of the advice provided is neither validated nor moderated, which could lead to inappropriate care and the spread of harmful medical advice. We aim to identify misinformation in medical and health forums in Reddit comments in order to alert users of dangerous medical advice. Our goal is to increase users' trust in the medical advice they find online.

Context understanding We asked 8 current Reddit users (4 men and 4 women, ages 19-27) to find a solution to a recent health concern on Reddit and observed their interactions. After completing the task, we asked the subjects what solutions they found, how confident they were in them, whether they would verify the solution using resources outside of Reddit, if they felt like the medical advice from Reddit needed additional clarification, and why they might use Reddit to find this type of information. We summarized our work to find the main themes and insights:

- People search for medical advice online especially if they are unsure about the severity of an issue or they are embarrassed to see a doctor. Other sites like WebMD, NIH, and NCBI are seen as"trustworthy" and are often used to verify information found on Reddit.
- Online health advice is a popular solution for people that don't have access to physicians or don't believe their health issue is an emergency.
- People are willing to try medical advice suggested in online resources even if they are skeptical about the Reddit responses, especially if they have financial constraints or no insurance.

2 Background

Previous research We performed a literature review of studies looking at user's interactions with health information on Reddit. Researchers have found users who looked for health information on the platform were likely to enact advice they found regardless of if they perceived the information to be credible or not [7]. This is especially concerning because some health information could be potentially dangerous.

Kotonya and Toni [6] explored fact-checking in the medical domain and developed a framework to verify whether a health claim is right and why. We also reviewed other tools such as the Fact-Checking Tool for Public Health Claims [3] and the Google Fact Check Tools API [2]. The best performing news misinformation model we found was a SVM classifier with independent variables such as cognitive, behavioral, effective, and visual cues [5]. We aim to build a similar transparent model that is specific to health misinformation and provides a detailed explanation.

We will utilize tokenization and KNN for the binary classification part of our problem: predicting if a health statement is misinformation or not based on its similarity to known statements. We will increase

transparency by showing the user which statements are most similar to the input. This is local interpretability; we emphasize transparency in this task because it is crucial for users to trust the medical advice the bot provides. If the most similar false statement is not meaningful to the user, the user will prefer the other explanations.

3 Methods

After being summoned, our Reddit bot will read the above comment and compare it to confirmed misinformation claims in our dataset. If the similarity of the response to known-false statements is above a certain threshold, the bot will respond with up to the top three similar phrases and a quantitative measure of similarity. If it is unable to determine that the comment is misinformation, it will respond encouraging users to complete their own research. A screenshot of our prototype is in the appendix.

The main feedback we received during the first iteration of the prototype was that the similar false claims that the bot displayed blended into the rest of the message and could easily be misconstrued as facts if read quickly. We addressed this by adding red X emojis, alert emojis, and the word "False!" in bold in the bot's misinformation message.

Our research question for user evaluation was: *What is the best way to explain the Reddit bot's misinformation classification to maximize people's understanding, trust, and satisfaction in the bot?* Our hypotheses:

> **H1:** Users will show more trust, satisfaction, and understanding in conditions with an explanation than the baseline condition (No Explanation).
>
> **H2:** Users will be more satisfied if the suggestion contains an explanation with similar statements.
>
> **H3:** When provided with an explanation, users will be more likely to not follow the dangerous advice in the original comment.

We used a between-subjects design where each participant was exposed to one type of explanation: 1. No Explanation (baseline), 2. Explanation with Percentage, 3. Explanation with Similar Statements. Participants were randomly assigned to a treatment. Our dependent variables were subjective constructs such as trust, explanation satisfaction, understanding, and the likelihood of users following the advice in the original comment. We measured trust using questions from the Subjective Understanding Scale [1], the CVR version of the Explanation Satisfaction Scale [4], open-ended questions regarding understanding and satisfaction, and "Would you follow the advice in the original comment?" as a proxy for behavioral outcomes.

The study design involves training participants with an example Reddit post, comment, and bot call. Users then viewed six examples of health posts that potentially contained misinformation. Half the examples contained misinformation, and the remaining did not. Users read the Reddit post and comment, called the bot, and filled out a survey to assess their understanding. Participants also completed a post-survey with questions to determine their overall impressions of the bot (survey questions in appendix).

We chose a K-Nearest Neighbors (KNN) model with k = 3 for our problem. Our model was trained on a subset of the PUBHEALTH dataset, using only claims tagged as being health-related. The training dataset included 3,638 labeled claims from the PUBHEALTH dataset and 22 manually-selected Reddit-like statements. The additional statements were included to supplement the PUBHEALTH dataset because some of the PUBHEALTH claims were not similar to Reddit comments. The test set (447 claims) was used for evaluation. All statements were cleaned and lemmatized before being processed by the model.

4 Results

4.1 User study evaluation

We recruited 15 participants for this study (8M, 7F). Thirteen had Bachelor's degrees, one had a high school education, and the final was unknown. Nine participants reported strongly agree, agree, or somewhat agree for the statement, "I am confident using

Reddit." Participants reported a low level of familiarity with Reddit bots, and 11 participants reported that they often look up health advice online.

When the bot reports that misinformation was detected, the Similar Statements explanation consistently received higher ratings than Percentage across the survey questionnaire. No Explanation consistently performed the lowest. Similar Statements was rated as having the greatest level of irrelevant detail. The Percentage explanation was much more preferred than the No Explanation condition.

4.2 Experiments

We conducted experiments with our KNN model to determine the best hyperparameters (k neighbors) and evaluation metrics. We found k = 3 to be the best value of k after trying k = 1, 3, 5, 15. We evaluated our different models by checking the statement value (true or false) of the majority of the similar sentences against the ground truth of the claim we were testing. We also manually evaluated the models by looking at the top k similar statements and seeing how many were relevant to the claim.

5 Discussion

Our results identify users' perceptions of the bot and provide insights for the final bot design.

5.1 Preferred Explanation

Across all cases (misinformation detected/not), users preferred Similar Statements (SS), than Percentage (P), and lastly No Explanation (NE) treatments. This suggests that the percentage adds value to the bot's response but not as much as similar statements. These results support H1 - they show that users have more trust, satisfaction, and understanding with an explanation than the baseline condition. They also support H2 since users have the greatest preference for the SS explanation.

One observation was that when no misinformation was detected, participants gave varying degrees of preference, despite the same generic message across all treatments. Users in the SS condition had a more positive response to the bot, while users in the P and NE conditions reported lower levels of satisfaction. This might be because when the model provides more helpful responses (SS) when finding misinformation, the model is perceived to be more helpful in general.

5.2 Overall Impressions of the Bot

Overall, people in the SS and P groups trusted the bot more than those in the NE group, indicating they would not try the advice in the original comment. 80% of the users in the NE group said they would still follow the advice of the original comment. Our bot should have an explanation as it increases trust in the bot and validates our third hypothesis.

We also found people in the P group believed the explanation told them how accurate the bot was and felt they could predict the bot's behavior. Fewer people in the SS group believed they would be able to predict the bot's output than the NE group. Based on this and our results above, we would consider a dual P and SS treatment in the future, as there seems to be more trust in the P group but more satisfaction after each response in the SS group.

5.3 Takeaways

The user study showed that participants preferred the SS explanation, supporting our hypotheses. Based on the results of the post-experiment survey, we will add a combination Percentage and Similar Statements explanation in the final bot and based on users' suggestions to allow users to view the models themselves, we will also include a link to the project directory and website in the bot's response.

The final KNN model (k=3) had accuracy of 0.74 when comparing the ground truth labels with the majority labels of similar statements. Manual inspection of the relevance of the similar statements revealed that in the model was successful in returning relevant statements, particularly the additional Reddit-like sentences, however a larger dataset is needed in order for the model to generalize to all medical topics.

References

[1] Hao-Fei Cheng et al. "Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders". In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–12.

[2] *Google Fact Check Tool APIs*. URL: https://toolbox.google.com/factcheck/apis.

[3] Alex Gui, Vivek Lam, and Sathya Chitturi. *Fact-checking tool for public health claims*. Mar. 2021. URL: https://stanford-cs329s.github.io/reports/Fact-Checking-Tool-for-Public-Health-Claims/.

[4] Robert R Hoffman et al. "Metrics for explainable AI: Challenges and prospects". In: *arXiv preprint arXiv:1812.04608* (2018).

[5] Christian Janze and Marten Risius. "Automatic detection of fake news on social media platforms". In: (2017).

[6] Neema Kotonya and Francesca Toni. "Explainable automated fact-checking for public health claims". In: *arXiv preprint arXiv:2010.09926* (2020).

[7] Rachael A Record et al. "I Sought It, I Reddit: Examining Health Information Engagement Behaviors among Reddit Users". In: *Journal of Health Communication* 23.5 (2018). PMID: 29718799, pp. 470–476. DOI: 10.1080/10810730.2018.1465493. eprint: https://doi.org/10.1080/10810730.2018.1465493. URL: https://doi.org/10.1080/10810730.2018.1465493.

[8] *United States's top websites ranking in August 2022*. URL: https://www.similarweb.com/top-websites/united-states/.

A Pre-Experiment Questions

Enter your participant number:
Which condition will the participant go through?
Please indicate how much you agree with the following statements: {Likert scale 1-7}
1. I am confident using Reddit.
2. I am familiar with Reddit bots.
3. I often look up health advice online.
What gender do you identify as?
What is your age?
What is your highest level of education you have completed?

B Individual Case Survey

What was the main health concern you were looking up in this task?
Did the bot communicate that the original comment was dangerous? (Yes/No)
The following questions were asked regarding the explanation provided by the bot {Likert scale 1-5}
1. I understand the bot's explanation.
2. The bot's explanation was satisfying.
3. The bot's explanation has sufficient detail.
4. The bot's explanation contains irrelevant details.
5. The bot's explanation seems complete.
6. This explanation of what the bot suggests is useful to my goals.
7. I understand why the bot gave this response.
8. To prove you are paying attention, please choose the option 'strongly disagree'.

C Post-Experiment Questions

In a few sentences, describe what you think the bot is trying to tell you.
In a few sentences, describe how you think the bot came to its conclusion.
The following questions were asked regarding the explanation provided by the bot {Likert scale 1-7}
1. This explanation says how accurate the bot is.
2. This explanation lets me judge when I should trust and not trust the bot.
3. I can predict how the bot will behave.
4. I have faith that the bot would be able to cope with all kinds of situations.
5. I trust the decisions made by the bot.
6. I can count on the bot to provide a reliable

decision.
Would you follow the advice in the original comment? (Yes/No)
Would you use this bot again if it was functional in Reddit? (Yes/No)
Do you have any suggestions about the functionality of the bot?

D Interface

We provide a screenshot of the user study interface below:

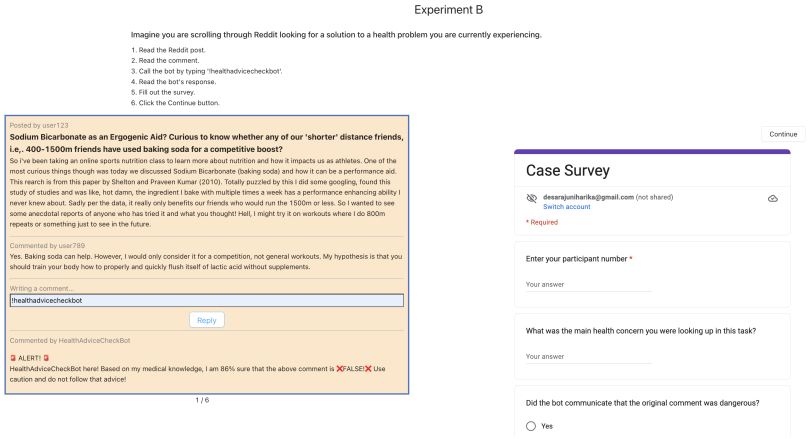

Figure 1: This is an example of what a user completing our user study will see. On the top you have instructions for how to navigate the page and the activity. The left contains the Reddit question, a comment to the question, and provides a place for our user to call the bot and see the response. On the left you see the case survey that we ask our users to fill out after each example. This screenshot shows Experiment B which shows users bot responses with no explanation.