Expressive Interviewing: A Conversational System for Coping with COVID-19

Charles Welch$^{\diamond}$**, Allison Lahnala**$^{\diamond}$**, Verónica Pérez-Rosas**$^{\diamond}$**, Siqi Shen**$^{\diamond}$**, Sarah Seraj**$^{\Phi}$**,**
Larry An$^{\iota}$**, Kenneth Resnicow**†**, James Pennebaker**$^{\Phi}$**, Rada Mihalcea**$^{\diamond}$
$^{\diamond}$Computer Science & Engineering, University of Michigan
$^{\Phi}$Department of Psychology, University of Texas
$^{\iota}$ Medical School, University of Michigan
† School of Public Health, University of Michigan
{cfwelch,alcllahn,vrncapr,shensq,lcan,kresnic,mihalcea}@umich.edu
{sarahseraj,pennebaker}@utexas.edu

Abstract

The ongoing COVID-19 pandemic has raised concerns for many regarding personal and public health implications, financial security and economic stability. Alongside many other unprecedented challenges, there are increasing concerns over social isolation and mental health. We introduce *Expressive Interviewing*– an interview-style conversational system that draws on ideas from motivational interviewing and expressive writing. Expressive Interviewing seeks to encourage users to express their thoughts and feelings through writing by asking them questions about how COVID-19 has impacted their lives. We present relevant aspects of the system's design and implementation as well as quantitative and qualitative analyses of user interactions with the system. In addition, we conduct a comparative evaluation with a general purpose dialogue system for mental health that shows our system's potential in helping users to cope with COVID-19 issues.

1 Introduction

The COVID-19 pandemic has changed our world in unimaginable ways, dramatically challenging our health system and drastically changing our daily lives. As we learned from recent large-scale analyses that were performed on social media datasets and extensive surveys, many people are currently experiencing increased anxiety, loneliness, depression, concerns for the health of family and themselves, unexpected unemployment, increased child care or homeschooling, and general concern with what the future might look like.[1]

Research in expressive writing (Pennebaker, 1997b) and motivational interviewing (Miller and Rollnick, 2012) has shown that even simple interactions where people talk about one particular experience can have significant psychological value. Numerous studies have demonstrated their effectiveness in improving people's mental and physical health (Vine et al., 2020; Pennebaker and Chung, 2011; Resnicow et al., 2017). Both expressive writing and motivational interviewing rely on the fundamental idea that by putting emotional upheavals into words, one can start to understand them better and therefore gain a sense of agency and coherence of the thoughts and emotions surrounding their experience.

In this paper, we introduce a new interview-style dialogue paradigm called *Expressive Interviewing* that unites strategies from expressive writing and motivational interviewing through a system that guides an individual to reflect on, express, and better understand their own thoughts and feelings during the pandemic.

By encouraging introspection and self-expression, the dialogue aims to reduce stress and anxiety. Our system is currently online at https://expressiveinterviewing.org and available for anyone to try anonymously.

2 Related Work

2.1 Expressive Writing

Expressive writing is a writing paradigm where people are asked to disclose their emotions and thoughts about significant life upheavals. Originally studied in the scope of traumatic experiences (Pennebaker and Beall, 1986), study participants are usually asked to write about an assigned topic for about 15 minutes for one to five consecutive days. Later studies expanded to specific experiences such as losing a job (Spera et al., 1994). Expressive writing has been shown to be effective on both physical and mental health measures by multiple meta-analyses (Frattaroli, 2006; Reinhold et al., 2018), finding its association with drops in physician visits, positive behavioral changes, and

[1] http://trackingsocial.life

long-term mood improvements. No single theory at present explains the cause of its benefits, but it is believed that the process of expressing emotions and constructing a story may play a role for participants in forming a new perspective on their lives (Pennebaker and Chung, 2011).

2.2 Motivational Interviewing

Motivational interviewing (MI) is a counseling technique designed to help people change a desired behavior by leveraging their own values and interests. The approach accepts that many people looking for a change are ambivalent about doing so as they have reasons to both change and sustain the behavior. Therefore, the goal of an MI counselor is to elicit their client's own motivation for changing by asking open questions and reflecting back on the client's statements. MI has been shown to correlate with positive behavior changes in a large variety of client goals, such as weight management (Small et al., 2009), chronic care intervention (Brodie et al., 2008), and substance abuse prevention (D'Amico et al., 2008).

2.3 Dialogue Systems

With the development of deep learning techniques, dialogue systems have been applied to a large variety of tasks to meet increasing demands. In recent work, Afzal et al. (2019) built a dialogue-based tutoring system to guide learners through varying levels of content granularity to facilitate a better understanding of content. Henderson et al. (2019) applied a response retrieval approach in restaurant search and booking to provide and enable the users to ask various questions about a restaurant. Ortega et al. (2019) built an open-source dialogue system framework that navigates students through course selection.

There are also dialogue system building tools such as Google's Dialogflow[2] and IBM's Watson assistant,[3] which enable numerous dialogue systems for customer service or conversational user interfaces.

2.4 Chatbots for Automated Counseling

Two dialogue systems for automated counseling services available on mobile platforms are Wysa[4] and Woebot.[5] These chatbots provide cognitive behavioral therapy with the goal of easing anxiety and depression by allowing users to express their thoughts. A study of Wysa users over three months showed that more active users had significantly improved symptoms of depression (Inkster et al., 2018). Another study shows that young students using Woebot significantly reduced anxiety levels after two weeks of using the conversational agent (Fitzpatrick et al., 2017). These findings suggest a promising benefit of automated counseling for the nonclinical population.

Our system is distinct from Wysa and Woebot in that it is designed specifically for coping with COVID-19 and allows users to write more topic related free-form responses. It asks open-ended questions and encourages users to introspect, and then provides visualized feedback afterward, whereas the others have a conversational logic mainly based on pre-coded multiple choice options.

3 Expressive Interviewing

Our system conducts an interview-style interaction with the users about how the COVID-19 pandemic has been affecting them. The system's goal is to encourage users to write as much as possible about themselves, building upon previous findings regarding the psychological value of writing about personal upheavals and the use of reflective listening for behavioral change (Pennebaker, 1997b; Miller and Rollnick, 2012).

The interview consists of a set of writing prompts in the form of questions about specific issues related to the pandemic. The system guides the interaction based on users responses, and provides reflective feedback and asks additional questions whenever appropriate. In order to provide reflective feedback, the system automatically detects the topics being discussed (e.g., work, family) or emotions being expressed (e.g., anger, anxiety), and responds with reflective statements that ask the user to elaborate or to answer a related question to explore that concept more deeply. For instance, if the system detects *work* as a topic of interest, it responds with "How has work changed under COVID? What might you be able to do to keep your career moving during these difficult times?"

After the interaction ends, i.e., all prompts have been answered by the user, the system provides detailed visual and textual feedback.

[2] https://dialogflow.com/

[3] https://www.ibm.com/cloud/watson-assistant

[4] https://wysa.io/

[5] https://woebot.io/

3.1 Guiding Questions

During the formulation of the guiding questions used by our system, we worked closely with our psychology and public health collaborators to identify a set of questions on COVID-19 topics that would motivate individuals to talk about their personal experience with the pandemic. We formulated the following question as the system's conversation starting point:

[Major issues] What are the major issues in your life right now, especially in the light of the COVID outbreak?

We also formulated three follow-up questions, which were generated after several refining iterations.[6] The order of these questions is randomized across users of the system.

[Looking Forward] What do you most look forward to doing once the pandemic is over?

[Advice to Others] What advice would you give other people about how to cope with any of the issues you are facing?

[Grateful] The outbreak has been affecting everyone's life, but people have the amazing ability to find good things even in the most challenging situations. What is something that you have done or experienced recently that you are grateful for?

3.2 Language Understanding and Reflection Strategies

Our system's capability for language understanding relies on identifying words belonging to various lexicons. This simple strategy allowed us to quickly develop a platform upon which we intend to implement a more sophisticated language understanding ability in future work.

When a user responds to one of the main prompts, the system looks for words belonging to specific topics and word categories. The system examines the user responses to identify dominant word categories or topics and triggers a reflection from a set of appropriate reflections. [7] If none of these types are matched, it responds with a generic reflection. The system flow is fully described in Algorithm 1.

Algorithm 1: Algorithm for an Expressive Interview showing the order of precedence for reflection types. The $IsNew$ and functions of pp prevent the system from looping and asking the same questions.

```
Intro;
Ask(Main,$Q_1$);
while MainQs Asked <4 do
    ps = GetPreviousUserStatement();
    rt = GetResponseTime();
    pp = GetPreviousSystemPrompt();
    $d_t$ = GetDominantTopic(ps);
    $d_e$ = GetDominantEmotion(ps);
    $d_p$ = GetDominantPronoun(ps);
    if (rt <15s or len(ps) <100) and not
     IsReflection(pp) and not
     IsShortResponse(pp) then
        Ask(ShortResponse, Random);
    else if IsNew($d_t$) and not
     IsReflection(pp) then
        Ask(TopicReflection, $d_t$);
    else if IsNew($d_e$) and not
     IsReflection(pp) then
        Ask(EmotionReflection, $d_e$);
    else if IsNew($d_p$) and not
     IsReflectioin(pp) then
        Ask(PronounReflection, $d_p$);
    else
        Ask(Main, Random);
    GetResponse;
end
```

The word categories are derived from the LIWC, WordNet-Affect, and MPQA lexicons (Pennebaker et al., 2001; Strapparava et al., 2004; Wiebe et al., 2005) and include pronouns (I, we, others), negative emotion (anger, anxiety, and sadness), positive emotion (joy) and positive and negative words. The COVID-19 related topics include finances, health, home, work, family, friends, and politics. Most of the topics are covered by the LIWC lexicon, with the exception of politics. For this category, we use the *politics* category from the Roget's Thesaurus (Roget, 1911) and add a small number of proper nouns covered in recent news (e.g., Trump, Biden, Fauci, Sanders).

We formulate a set of specific reflections for each word category and topic, which were refined

[6] We removed an additional question about how people's lives have changed since the outbreak, as well as a question about what people missed the most about their previous lives.

[7] A dominant word category is defined as a word type, where the frequency of occurrence is at least 50% higher than the second highest frequency category for that group.

by our psychology and public health collaborators. For instance, if the dominant emotion category is anxiety, the system responds “You mention feelings such as fear and anxiety. What do you think is the best way for people to cope with these feelings?” Initially, we also considered reflections for different types of pronouns, but found that they did not steer the dialogue in a meaningful direction. Instead, we flag responses with dominant use of impersonal pronouns and lack of references to the self and reflect that fact back to the user and further ask them how they are specifically being affected. We also crafted generic reflections to be applicable to a large number of situations though the system does not understand the content of what the user has said (e.g. “I see. Tell me about a time when things were different”, and “I hear you. What have you tried in the past that has worked well”).

3.3 User Feedback

After the interview, the system provides visual and textual feedback based on the user’s responses and provides links to resources (i.e., mental health resources) appropriate given their main concerns.

The visual feedback consists of four pie charts showing the relative usage of different word categories, including: discussed topics (work, finance, home, health, family, friends and politics), affect (positive, negative), emotions (anger, sadness, fear, anxiety, joy), and pronouns (I, we, other).

The textual feedback includes a comparison with others (to normalize the user’s reactions) and interpretations of where the user falls within normalized scales. The system also presents a summary of the most and least discussed topics and how they compare to the average user, along with normalized values for meaningfulness, self-reflection, and emotional tone (using a 0-10 scale) along with textual descriptors for the shown scale values. [8] These metrics are inspired by previous work on expressive writing and represent the self-reported meaningfulness, usage of self-referring pronouns, and the difference in positive and negative word usage (Pennebaker, 1997a). Finally, the system provides relevant resources for further exploration (e.g. for the work topic it lists external links to COVID related job resources and safety practices).

3.4 Online Interface

The system is implemented as a web interface so it is accessible and easy to use. The interface is built with the Django platform and jQuery and uses Python on the backend (Django Software Foundation, 2019).

Before the interaction users are asked to report on a 1-7 scale: (1) **[Life satisfaction]** how satisfied they are with their life in general, and (2) **[Stress$_{before}$]** what is their level of stress. The user then proceeds to the conversational interaction with our system. **After** the interaction, the user is asked again about (3) **[Stress$_{after}$]** what is their level of stress; (4) **[Personal]** how personal their interaction was; and (5) **[Meaningful]** how meaningful their interaction was. Once this is submitted, the user can proceed to the feedback page and view details about what they wrote and how their interaction compares to a sample of recent users. The user is finally presented with a list of resources triggered by the topics discussed.

We made an effort to make our system appear human-like to make users more comfortable while interacting with it, although this can vary for different individuals. In future work, we hope to explore individual personas and more sophisticated rapport building techniques. We named our dialogue agent ‘C.P.’, which stands for *Computer Program*. This name acknowledges that the user is interacting with a computer, while at the same time it makes the system more human by assigning it a name. When responding to the user, C.P. pauses for a few seconds as if it is thinking and then proceeds to type a response one letter at a time with a low probability of making typos – similarly to how human users would type.

4 Analysis of User Interactions

After the system was launched (and up to when we conducted this analysis), we had 174 users interact with the system. We analyze these interactions to evaluate system usefulness, user engagement, and reflection effectiveness.

4.1 System Usefulness

We examine the system’s ability to help users cope with COVID-19 related issues by analyzing the different ratings provided by users before and after their interaction with C.P. First, we look at the change in stress for the users after using our system as stress decrease is one target by which we can see

[8] Textual descriptions are predefined for different ranges of each scale

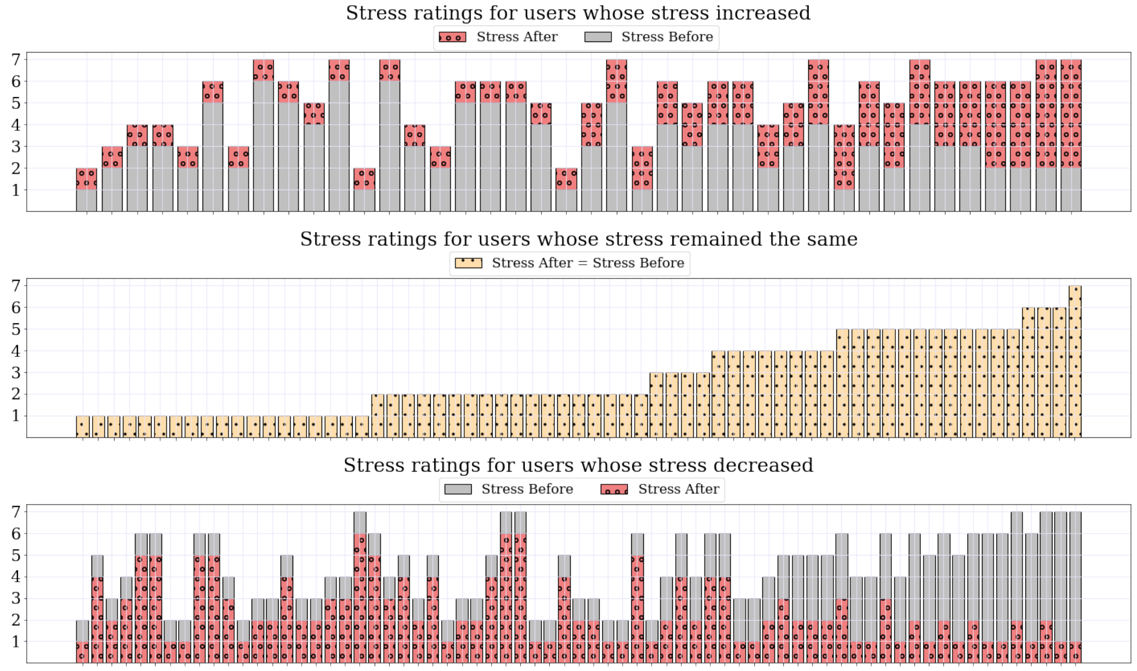

Figure 1: Top: before and after stress ratings by users whose stress increased after interaction with C.P. Middle: before and after stress ratings by users whose stress remained the same after interaction with C.P. Bottom: before and after stress ratings by users whose stress decreased after interaction with C.P. The bars are ordered by the magnitude of change (top and bottom), or by the static stress rating (middle).

if the system is useful for users. We observe that stress decreased for most users, which is shown in Figure 1 with bar charts to reflect the users' stress levels before and after using the system.

Throughout this discussion, we use ΔStress to indicate how the users stress rating differs before and after the interaction: ΔStress = Stress$_{after}$ - Stress$_{before}$. Negative values for ΔStress are therefore an indicator of stress reduction, whereas positive values for ΔStress reflect an increase in stress. We start by measuring the Spearman correlation between the different ratings for the 174 interactions with C.P. Results are shown in Table 1.

The strongest correlation we observe is between the *personal* and *meaningful* ratings, suggesting that interactions that are more *meaningful* appear to feel more *personal*, or vice versa.

We also observe a strong negative correlation between Δ Stress and the meaningfulness of the interaction, suggesting that the interactions that the users found to be meaningful are associated with a reduction in stress.

4.2 User Engagement

We examine user engagement by analyzing the time users spend in the interaction and the number of words they write throughout the session. Figure 2 shows histograms of the session lengths in the number of words used by the user and of the session duration in seconds. The rightmost column of Table 2 shows Spearman correlation coefficients between user ratings and the length and duration of the sessions. We find a significant negative correlation between *Stress*$_{before}$ and *Stress*$_{after}$ with session duration and number of words, suggesting

Rating 1	Rating 2	rho
Life satisfaction	Stress$_{before}$	**-0.261**
Life satisfaction	Stress$_{after}$	**-0.166**
Life satisfaction	Δ Stress	0.083
Life satisfaction	Personal	**0.285**
Life satisfaction	Meaningful	**0.243**
Meaningful	Stress$_{before}$	0.033
Meaningful	Stress$_{after}$	**-0.226**
Meaningful	Δ Stress	**-0.202**
Meaningful	Personal	**0.675**
Personal	Stress$_{before}$	0.065
Personal	Stress$_{after}$	-0.067
Personal	Δ Stress	-0.073

Table 1: Spearman correlation coefficients between pairs of ratings for the 174 interactions. Bold indicates significance with $p < 0.05$.

an association between user engagement and lower stress. There is also a weak negative correlation between duration of session and reduction in stress (ΔStress).

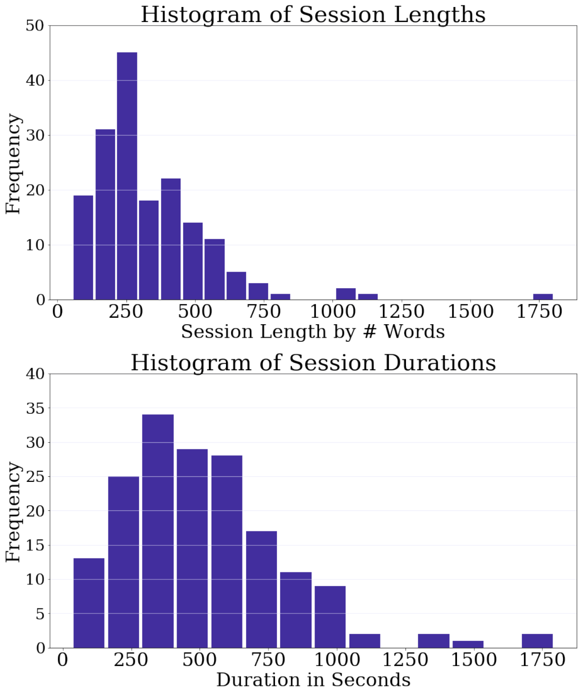

Figure 2: Histograms of overall user engagement measured by session length and duration.

We also investigate if there is a relationship between the pre- and post-session ratings and how engaged a user was with each prompt in terms of length of and duration in writing their response. Table 2 shows Spearman correlation coefficients for these relationships. It appears that *Life Satisfaction* has no correlation with the length of any prompt response except a potentially weak negative correlation with length on the *Major Issues* prompt ($p = 0.052$). A lower rating may relate with having more personal challenges to write about.

$Stress_{before}$ has a weak negative correlation between the number of words used and the duration spent in the response to *Looking Forward*. Higher stress may relate to present concerns, which may make one less inclined to spend time thinking and writing about positive aspects of their future than someone with less stress. We presume this could be the case for the *Grateful* prompt, which likewise correlates weakly and negatively with $Stress_{before}$.

$Stress_{after}$ has a negative correlation between duration spent on every prompt response except for the time spent on *Major Issues*. This could be a reflection of the fact that those who have a lot to write about major issues in their life also incur high levels of stress.

The *Personal* rating shows no correlations with the duration spent on any of responses, except potentially *Advice to Others* ($p = 0.074$). We do observe weak negative correlations between *Personal* ratings and response *lengths* on *Major Issues* and *Looking Forward*, and potentially on *Grateful* ($p = 0.054$) and *Advice to Others* ($p = 0.08$). Perhaps if a user writes more, there is a greater expectation for more personal reflections. We discuss engagement related to reflections more deeply in the next section.

The *Meaningful* rating shows weak negative correlations with length on *Major Issues*, *Advice to Others*, and possibly on *Grateful* ($p = 0.052$) and *Looking Forward* ($p = 0.062$). We do not observe a significant correlation with *duration* on *Major Issues* or *Grateful*, but we do observe positive correlations between *duration* and *Looking Forward* and *Advice to Others*. Users who spend more time thinking about advice they would give others facing their issues may find the interaction more meaningful, and may experience benefits having reflected on their agency in managing their challenges.

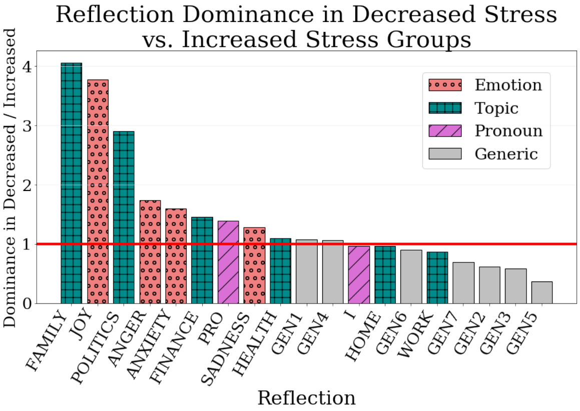

Figure 3: The dominance of each reflection triggered for users whose stress decreased divided by each reflection's dominance for users whose stress increased. Scores above 1 (red line) correspond to a decrease in stress; score below 1 correspond to an increase in stress. See Table 3 for sample reflections, including the GENERIC reflections.

4.3 Reflection Effectiveness

To investigate the effectiveness of Expressive Interviewing reflections, we compare the reflections that were triggered for users whose stressed decreased to the reflections that triggered for the users whose stress increased. For each of these user groups,

	Major issues	Grateful	Looking Forward	Advice to Others	Overall
	Length in Words				
Life Satisfaction	-0.148	-0.121	-0.079	-0.096	-0.070
Personal	**-0.156**	**-0.147**	**-0.185**	-0.134	**-0.159**
Meaningful	**-0.181**	**-0.148**	-0.142	**-0.151**	**-0.151**
Stress$_{before}$	-0.001	-0.076	**-0.161**	-0.083	**-0.151**
Stress$_{after}$	-0.020	-0.135	-0.130	-0.129	**-0.177**
Δ Stress	-0.067	-0.106	-0.039	-0.112	-0.092
	Duration in Seconds				
Life Satisfaction	-0.057	0.016	0.048	0.091	0.066
Personal	-0.017	-0.041	0.053	0.136	0.066
Meaningful	-0.036	0.099	**0.173**	**0.205**	0.143
Stress$_{before}$	-0.067	**-0.252**	**-0.178**	-0.099	**-0.198**
Stress$_{after}$	-0.120	**-0.241**	**-0.207**	**-0.192**	**-0.233**
Δ Stress	-0.069	-0.023	-0.052	-0.092	-0.068

Table 2: Spearman correlation coefficients between each rating provided by a user and (top) the length in number of words of the user's response to each particular prompt, and (bottom) duration in seconds of the user's response to each particular prompt, from 174 full interactions. Bold denotes significance with $p < 0.05$.

HEALTH	I'd like to know more about your feelings surrounding your own health and the health of people close to you. What actions can you take to help keep you healthy during these challenging times?
FAMILY	What can you do to keep your family resilient during these tough times?
POLITICS	What is it about the political world that may be hooking you? What are your reactions saying about you?
GENERIC	Interesting to hear that. How does what you say relate to your values?
GENERIC	I see. Tell me about a time when things were different.

Table 3: Sample topic specific and generic reflections.

we compute the dominance of each reflection as its proportion of times it was triggered out of all reflections triggered. In Figure 3, we compare the dominance of each reflection across these user groups by dividing the reflection dominance in the decreased-stress group by that of the increased-stress group.

Importantly, we observe that all emotion reflections and more topic reflections were triggered at a higher rate for users whose stress decreased, whereas more generic reflections were triggered at a higher rate for users whose stress increased. While we do not presume that increased stress was due to generic reflections, the correspondence between emotion and topic reflections with stress reduction aligns with expectations of effective reflections from motivational interviewing–generic reflections and specific reflections resemble *simple reflections* and *complex reflections* respectively, as referred to in Motivation Interviewing. While both types of reflections serve a purpose, complex reflections both communicate an understanding of what the client has said and also contribute an additional layer of understanding or a new interpretation for the user, whereas simple reflections focus on the former (Rollnick and Allison, 2004).

In qualitatively analyzing the instances where generic reflections were triggered, we observe that contextual appropriateness seems to be the best indicator of their success (in terms of ability to elicit a deeper thought, feeling, or interpretation) given that the user was invested in the experience. As these generic reflections are selected at random, their contextual appropriateness was inconsistent, illuminating the scenarios in which they are more or less appropriate. For instance, out of the seven times the reflection "*Interesting to hear that. How does what you say relate to your values?*" was triggered for the increased-stress users, one user expanded on their previous message, one expressed confusion about the question, and another copied and pasted the definition of *core values*[9] as their response. Two other instances of this reflection were triggered when a user had expressed negative feelings such as worry and feeling lazy which appeared misplaced, and the last case was triggered by a message that was not readable. Out of the thirteen times the same reflection was triggered for the decreased-stress group, one user expressed not having much to say, another gave one word responses before and after, and all others expanded on their previous message in relation to their values or gave a simple response to indicate a degree that it relates. This reflection appeared more "successful" (based on if the user expanded on their previous message or values) when it was triggered by a message with more neutral to positive sentiment, such as when the user was expressing what they were looking forward to, or when they had several pieces of advice to offer for a friend in their situation, as opposed to one with more negative sentiment like the messages expressing worry or laziness.

In instances of other generic reflections, we observed that another issue for appropriateness was whether the reflection matched the user's frame of thought in terms of past, present, or future. For instance, the reflection *"I see. Tell me about a time when things were different,"* best matched scenarios when users described thoughts about changes to their daily lives, but not when users described future topics such as what they were looking forward to, nor when they were already describing the past.

Based on our observations of the reflections in action, we have three main takeaways. First, topic and emotion specific reflections are more associated with the group of users whose stress decreased. These reflections are only triggered if the system determines a dominant topic or emotion, which depends on the effectiveness of its heuristics, as well as the amount of detail and context that a user expresses. This leads to the next takeaway, that the system appears to be more effective when users approach the experience with an intention for expression, or conversely it seems less effective when the intent to not engage and express is explicit. Third, the generic reflections were developed with the intent to function in generic contexts, but we learned in practice that some clashed with emotional and situational content or were confusing given the context. As we did observe many, if not more, successful instances of generic reflections, we are able to contrast these contexts to the unsuccessful contexts, and can develop a heuristic for selecting the generic reflections rather than selecting at random, as well as adapt the language of our current generic reflections to be more appropriate for the Expressive Interviewing setting.

5 Comparative Evaluations

To assess the extent to which our Expressive Interviewing system delivers an engaging user experience, we conduct a comparative study between our system and the conversational mental health app Woebot (Fitzpatrick et al., 2017). We chose Woebot over Wysa as it was more widely available for COVID-19 mental support at the time we were conducting our experiments and also for its use of predefined prompts, which was similar to our own strategies.

We recruited 12 participants and asked them to interact independently with each system to discuss their COVID-19 related concerns. Participants include graduate and undergraduate students in our lab, 4 female and 8 male and with ages ranging between 22 to 35 years. More specifically, we asked them to use each system during 10-15 minutes and provide evaluative feedback pre- and post-interaction. The interaction with our system ended whenever the participant responded the four main prompts and reached the feedback page. While interacting with Woebot, participants stopped after they interacted with the system for at least 10 minutes. To avoid cognitive bias, we randomized the order in which each participant evaluated the systems. In addition, we randomized the order in which the questions are shown in the evaluation form. To protect participant's privacy we did not link their interaction logs with each system with their evaluation responses.

Before interacting with either system, participants rated their life satisfaction and their stress level. After the interaction, participants reported again their stress level and rated several aspects of their interaction with the system, including ease of use, usefulness (in terms of discussing COVID-19 related issues and motivation to write about it), overall experience, and satisfaction using mainly binary scales. For example, the questions "Did <system> motivate you to write at length about

[9] https://www.indeed.com/career-advice/career-development/core-values

	Woebot	Expressive Interviewing
Stress_{before}	91%	91%
Stress_{after}	73%	64%

Table 4: Percentage of users reporting high levels of stress ($>$ 3 on a 7-point Likert scale) before and after using Woebot and Expressive Interviewing.

	Woebot	Expr. Interv.
Ease of Use	82%	91%
Useful	18%	73%
Motivation to Write	27%	91%
User Satisfaction	36%	36%
Meaningful Interaction	64%	73%
Overall Experience	36%	46%

Table 5: Comparative evaluation of Woebot and Expressive Interviewing. Percentage of users reporting positive/high ratings scores (>3 on a 7-point Likert scale) on usability aspects after interacting with Woebot and Expressive Interviewing.

your thoughts and feelings? yes/no" and "How useful was C.P. to discuss your concerns about COVID? useful/not useful" assess whether the system encouraged the user to write about their thoughts and feelings about COVID and whether the system provided guidance for it. Tables 4 and 5 show the percentage of users that provided positive or high scores ($>$ 3 on a 7-point scale) for each of these aspects after interacting with both systems.

As observed, there are fewer participants reporting high levels of stress after using either system. However, we see a smaller fraction of participants reporting high levels of stress after interacting with Expressive Interviewing, thus suggesting that our system was more effective in helping participants to reduce their stress levels.

Overall, participants reported that Expressive Interviewing was easier to use, more useful to discuss their COVID concerns and motivated them to write more than Woebot. Similarly, users reported a more meaningful interaction and a better overall experience. While these results are encouraging, it is important to note that this study has been conducted with a small pool of participants and that Woebot was not specifically designed for discussing COVID-19 concerns and it is of more general purpose than our system. Nonetheless, our analysis suggests that a dialogue system such as Expressive Interviewing can be effective in helping users cope with COVID-19 issues as compared to a general purpose dialogue system for mental health.

6 Ethical and Privacy Considerations

During development, we followed suggestions of previous research on automated mental health counseling and adopted the goals of being respectful of user privacy, following evidence based methods, ensuring user safety, and being transparent in system capabilities (Kretzschmar et al., 2019). However, there are a number of open questions surrounding the use of conversational agents in healthcare and we need to continue to evaluate our system to ensure that it benefits users (McGreevey et al.).

The practices of motivational interviewing and expressive writing have numerous studies supporting their efficacy (Miller and Rollnick, 2012; Pennebaker and Chung, 2007). The combination of these methods in an interviewing format has not previously been studied and we intend to continue publishing our findings as the user population expands and becomes more diverse. We will also continue to improve our system and assessment.

In addition, we have taken efforts to secure users' data. For instance, we do not ask for identifiers and data is stored anonymously by session ID. The website is secured with SSL. Furthermore, the collected data is only accessible to researchers directly involved with our study. Our study has been approved by the University of Michigan IRB (HUM00182586).

7 Conclusion

In this paper, we introduced an interview-style dialogue system called Expressive Interviewing to help people cope with the effects of the COVID-19 pandemic. We provided a detailed description on how the system is designed and implemented.

We analyzed a sample of 174 user interactions with our system and conducted analyses on aspects such as system usefulness, user engagement and reflection effectiveness. We also conducted a comparative evaluation study between our system and Woebot, a general purpose dialogue system for mental health. Our main findings suggest that users benefited from the reflective strategies used by our system and experienced meaningful interactions leading to reduced stress levels. Furthermore, our system was judged to be easier to use and more useful than Woebot when discussing COVID-19 related concerns.

In future work we intend to explore the applicability of the developed system to other health-related domains.

Acknowledgments

This material is based in part on work supported by the Precision Health initiative at the University of Michigan, the National Science Foundation (grant #1815291), and the John Templeton Foundation (grant #61156). Any opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect the views of the Precision Health initiative, the National Science Foundation, or the John Templeton Foundation.

References

Shazia Afzal, Tejas Dhamecha, Nirmal Mukhi, Renuka Sindhgatta, Smit Marvaniya, Matthew Ventura, and Jessica Yarbro. 2019. Development and deployment of a large-scale dialog-based intelligent tutoring system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 114–121, Minneapolis, Minnesota. Association for Computational Linguistics.

David A Brodie, Allison Inoue, and David G Shaw. 2008. Motivational interviewing to change quality of life for people with chronic heart failure: a randomised controlled trial. *International journal of nursing studies*, 45(4):489–500.

Elizabeth J D'Amico, Jeremy NV Miles, Stefanie A Stern, and Lisa S Meredith. 2008. Brief motivational interviewing for teens at risk of substance use consequences: A randomized pilot study in a primary care clinic. *Journal of substance abuse treatment*, 35(1):53–61.

Django Software Foundation. 2019. Django.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Joanne Frattaroli. 2006. Experimental disclosure and its moderators: a meta-analysis. *Psychological bulletin*, 132(6):823.

Matthew Henderson, Ivan Vulić, Iñigo Casanueva, Paweł Budzianowski, Daniela Gerz, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. PolyResponse: A rank-based approach to task-oriented dialogue with application in restaurant search and booking. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 181–186, Hong Kong, China. Association for Computational Linguistics.

Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Kira Kretzschmar, Holly Tyroll, Gabriela Pavarini, Arianna Manzini, Ilina Singh, and NeurOx Young People's Advisory Group. 2019. Can your phone be your therapist? young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical informatics insights*, 11:1178222619829083.

John D McGreevey, C William Hanson, and Ross Koppel. Clinical, legal, and ethical aspects of artificial intelligence–assisted conversational agents in health care. *JAMA*.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Daniel Ortega, Dirk Väth, Gianna Weber, Lindsey Vanderlyn, Maximilian Schmidt, Moritz Völkel, Zorica Karacevic, and Ngoc Thang Vu. 2019. ADVISER: A dialog system framework for education & research. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–98, Florence, Italy. Association for Computational Linguistics.

James W Pennebaker. 1997a. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162–166.

James W Pennebaker and Sandra K Beall. 1986. Confronting a traumatic event: toward an understanding of inhibition and disease. *Journal of abnormal psychology*, 95(3):274.

James W Pennebaker and Cindy K Chung. 2007. Expressive writing, emotional upheavals, and health. *Foundations of health psychology*, pages 263–284.

James W Pennebaker and Cindy K Chung. 2011. Expressive writing: Connections to physical and mental health. Oxford University Press.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

J.W. Pennebaker. 1997b. Writing about emotional experiences as a therapeutic process. *Psychological Science*, (8):162–166.

Maren Reinhold, Paul-Christian Bürkner, and Heinz Holling. 2018. Effects of expressive writing on depressive symptoms—a meta-analysis. *Clinical Psychology: Science and Practice*, 25(1):e12224.

K. Resnicow, PJ Teixeira, and GC Williams. 2017. Efficient allocation of public health and behavior change resources: The ”difficulty by motivation” matrix. *American Journal of Public Health*, 107(1):55–57.

Peter Mark Roget. 1911. *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company.

Stephen Rollnick and Jeff Allison. 2004. Motivational interviewing. *The essential handbook of treatment and prevention of alcohol problems*, pages 105–116.

Leigh Small, Deborah Anderson, Kimberly Sidora-Arcoleo, and Bonnie Gance-Cleveland. 2009. Pediatric nurse practitioners' assessment and management of childhood overweight/obesity: Results from 1999 and 2005 cohort surveys. *Journal of Pediatric Health Care*, 23(4):231–241.

Stefanie P Spera, Eric D Buhrfeind, and James W Pennebaker. 1994. Expressive writing and coping with job loss. *Academy of management journal*, 37(3):722–733.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: An affective extension of wordnet. In *Lrec*, volume 4, page 40. Citeseer.

V. Vine, R.L. Boyd, and J.W. Pennebaker. 2020. Feelings in many words: Natural emotion vocabularies as windows on distress and well-being. *Nature Communications*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.