Assessing Users' Mental Status from their Journaling Behavior through Chatbots

Masamune Kawasaki
Tokyo Institute of Technology,
NTT Communication Science Laboratories
kawasaki.m.ag@m.titech.ac.jp

Naomi Yamashita
NTT Communication Science Laboratories
naomiy@acm.org

Yi-Chieh Lee
University of Illinois Urbana-Champaign
ylee267@illinois.edu

Kayoko Nohara
Tokyo Institute of Technology
nohara.k.aa@m.titech.ac.jp

ABSTRACT

Chatbots (conversational agents) are increasingly receiving attention in mental health domains because they elicit honest self-disclosure about personal experiences and emotions. Although such self-disclosure contents can be useful for gauging mental status, little research has addressed how to automatically assess mental status from self-disclosures to a chatbot. If a chatbot can automatically assess the mental status of users, it can help them improve their mental wellness or facilitate access to mental professionals. In this paper, we examine whether indicators that identify depression from written texts (e.g., social media posts) are also useful for assessing mental status from disclosures to a chatbot. We first ran a study with 30 participants who engaged in daily journaling with a chatbot that prompted them to record their moods and experiences for three weeks. We then divided the participants' self-disclosure data into three groups based on their mental state changes before and after the study: improved vs. deteriorated vs. no change. Comparing the data among the three groups, participants whose mental states deteriorated during the study gradually used fewer positive emotion and concrete words but more negative emotion words when describing their daily experiences and feelings to the chatbot.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **HCI design and evaluation methods** → *User studies.*

KEYWORDS

Intelligent virtual agent, conversational behavior, mental well-being

ACM Reference Format:
Masamune Kawasaki, Naomi Yamashita, Yi-Chieh Lee, and Kayoko Nohara. 2020. Assessing Users' Mental Status from their Journaling Behavior through Chatbots. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423870

1 INTRODUCTION

The application of chatbots in mental health is experiencing considerable growth. Research has shown that people who are deeply distressed prefer confiding their anxieties to chatbots rather than to humans [14, 22] because they fear being rejected or judged [1]. Such anxieties of being stigmatized by others often cause them to avoid reaching out to professionals for proper help [16, 25]. A number of studies have used chatbots for therapy or counseling [4, 31, 32]. For example, a therapy chatbot called "Woebot" utilized a chatbot to explore its feasibility to reduce students' mental health problems and showed that it relieved symptoms of anxiety and depression [12].

Previous research on chatbots in mental healthcare has primarily focused on supporting people who have already been diagnosed with such mood disorders as depression. However, early detection of growing mental health problems is also crucial before symptoms develop for two main reasons: (1) people tend to be insensitive to their own mental status when they are distressed [41], and (2) treatment typically takes longer when recovering from severe symptoms than from mild symptoms. Therefore, it would be useful if a

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423870

chatbot could detect changes in mental status based on daily conversation behavior.

Journaling, which can be conducted with a chatbot, is also common among those who are experiencing mental issues because it provides effective stress relief [2]. Furthermore, people engage in more truthful self-disclosure to chatbots than in face-to-face interviews [23] or through web surveys [19]. Hence, chatbots might encourage self-disclosure through journaling and detect potential health issues.

In this paper, we analyzed the dialogue content that people daily recorded through a chatbot ¥citeYi-Chieh Lee et al.. 30 participants entered daily occurrences and their moods into chatbots for three weeks. In the analysis, we divided the participants' data into three groups based on the changes in their mental status before and after this study and analyzed the differences among the three groups. We hypothesized that we could assess the mental statuses of users from their journaling behavior by indicators that detect depression from such written texts as social media and blog posts. The following are the main contributions of our paper:

- Our findings suggest that we can detect a person whose mental status has deteriorated from self-disclosure data through chatbots.
- We found that people whose mental states have deteriorated during the study gradually used fewer positive emotion and concrete words. They also began to use more negative emotion words when describing their daily experiences and feelings to the chatbot.

2 RELATED WORK AND HYPOTHESES

2.1 Journaling through chatbots

In the practice of journaling, people write about their experiences, feelings, and thoughts [43]. This relatively common practice effectively reduces stress and anxiety [43]. When people get distressed, they tend to write more deeply about their stress that they usually avoid disclosing to others [42].

Given the proliferation of smartphones, researchers have developed chatbots to facilitate user journaling. Studies [10, 12] have described the positive effects of deploying chatbots, including eliciting honest self-disclosure about personal experiences and emotions [23, 35], improving self-awareness, and raising self-reflection. Previous works also concluded that such disclosure is promoted by chatbots. For example, Gale et al. found that when people are interviewed by a virtual agent, they tended to disclose more about their depressed thoughts than when interviewed face-to-face by a human [23].Bhakta et al. showed that especially on sensitive topics such as drugs and gender, people more disclose to a embodied virtual agent than to a human. [5]

If a chatbot can further assess users' mental status from their text input, it may lead them to access mental professionals when help is needed. Indeed, Delahunty et al. [9] suggested that such a feature may facilitate real-time instant crisis support for those suffering from depression by identifying mental health issues from daily written texts through chatbots. However, there is a lack of research addressing if and how such contents can be used for automatically assessing mental status.

2.2 Predicting mental states from written text

Previous works developed methods to estimate the severity of depression by analyzing user's writing. In sociolinguistics, Oxman argued that linguistic analysis can detect groups who are suffering from depression and paranoia [33].

Some works have explored the automated classification of psychological disorders based on such observed differences in communicative behavior [8]. For example, Choudhury et al. found that in a social-media-posting context (e.g., Twitter and Facebook), as a person's depression deepens, the amount of posts and the use of third-person pronouns decreased and the use of negative emotion words and first-person pronouns increased [7]. Resnik et al. [36] analyzed essays and showed that people with mental health problems used fewer positive emotion words and more negative emotion words than people without mental health issues. Gao et al. showed that in social media posts the more depressed a user is, the less concrete their tweets are. For example, the tweet "I played sports" is less concrete than "I played tennis" [13]. Several studies demonstrated that depressed people tend to use more first-person pronouns in their essays than those who are not depressed [28, 38, 39].

2.3 Hypotheses

In this work, we examined the relationship between the tendency found in textual data that people input to chatbots and their mental status. We propose the following hypotheses using indicators that detect depression from written texts in previous researches.

Previous works described how people's text input to social media posts considerably decreases when they get depressed [7]. However, this may be due to the presence of others on social media sites. In other words, people may feel reluctant to write negative things when they know that other people are reading their posts. According to [24], this is true even when posts are anonymous because people worry about getting judgmental comments from others. On the other hand, people can disclose negative thoughts (e.g., anxieties) to a chatbot without worrying about peer pressure or retribution. People might disclose more to a chatbot when they are experiencing mental health issues. Based on the above consideration, we posed the following question:

RQ1: How does the amount of words change over time in chatbot conversations when people are experiencing stress or mental health issues?

Regarding the type of words used, Choudhury et al. [7] found that when people are depressed, they include more first-person pronouns in their social media posts and fewer third-person pronouns. However, in journaling with a chatbot, people naturally talk about themselves because chatbots prompt them to focus on their moods, feelings, and the experiences associated with them. As a result, first-person pronouns are much more likely to appear in conversations with a chatbot than in social media posts. Since the tendency to use pronouns may differ from those found in social media, we also asked the following question:

RQ2: How does the use of first-person pronouns and third-person pronouns change over time in conversations with a chatbot when people are experiencing mental health issues?

In addition to the use of first- and third-person pronouns, previous works reported that the use of negative emotion words in social media posts increased as the degree of depression increased [7]. Similarly, Resnik et al. [36] found that depressed people used fewer positive emotion words in their writing (essays) than the non-depressed. We hypothesize that a similar tendency will be found when people talk with a chatbot:

H1: As people experience more mental health problems, their use of negative emotion words increases and their use of positive words decreases over time in chatbot conversations.

Finally, another previous work argued that the more depressed users are, the fewer concrete words they use in their tweets [13]. Matsumoto et al. showed that since describing negative moods or feelings may fuel negative feelings, people might avoid talking about their negative feelings [27]. This phenomenon could occur in the same way even when prompted by a chatbot. Therefore, we hypothesize that:

H2: As people begin to experience mental health issues, their use of concrete words will gradually decrease in chatbot conversations.

By testing our two hypotheses, we uncover whether we can detect individuals whose mental status gradually deteriorated by analyzing chatbot conversations.

3 METHOD

Chatbot Design

We designed our chatbot using Manychat and Google Dialogflow. We used the former to design the basic flow of the journaling tasks and to monitor whether the participants had completed them. We designed the basic conversation journaling flow with fixed questions and responses. The chatbot generally asked three to eight (M = 5.11) questions during each journaling task. The chatbot primarily behaved like a listener, giving only such simple responses as "I got it," "Okay," or encouraging the user to expand, such as "Do you want to tell me more?" Therefore, the chatbot did not need to completely understand the user's responses to fuel the conversation.

We incorporated Dialogflow into our chatbot so that participants will feel as if they are naturally talking with the chatbot. Accordingly, it utilized natural language processing (NLP) to decide a suitable response. For instance, if a participant input "I felt good today," the chatbot would ask through Dialogflow a follow-up question: "Why did you feel good?" Furthermore, Dialogflow helped deal with some unexpected questions. In such an experiment, participants often posed questions outside the range of the predefined conversational task. In such cases, such entries are sent to Dialogflow, processed, and answered. For instance, a participant might ask about the chatbot's identity ("Where did you go to university?") or talk to it as if it were a human ("Did you finish your breakfast?"). If the system can naturally deal with such simple conversations, participants might have more positive feelings about their chatbot experience. However, if the user asked a question that Manychat or Dialogflow couldn't handle, the chatbot urged participants to rephrase it. The system moved on to a new topic if it got stuck three times in the same chat.

Finally, we designed the chatbot's icon to resemble a handshake instead of a specific gender to maintain a neutral impression of it. We told the participants that their journaling contents will be kept confidential but will be shared and analyzed by the research team. Participants were allowed to access the chatbot any time from 5 p.m. to 12 a.m. We selected 5 p.m. as the start time to ensure that they could easily remember their emotions or experiences on that day for their journaling. Participants could only input one journal entry each day.

Participants

We posted information on social media, websites, and a university's electronic bulletin board to recruit university students. The following are the other participation criteria: (1) 18 years old or above, (2) Kessler Psychological Distress Scale (K6) scores lower than 13 [34], denoting that they did not currently have any serious mental issue. We also informed the participants that the study lasted for three weeks, but they were allowed to drop out at any time.

We recruited 30 participants, 13 males and 17 females. All ranged in age from 20 to 27 (M=23.00). We incorporated our chatbot into Facebook Messenger because they were accustomed to using it. At the end of their three-week journaling with the chatbot, the participants took the K6 test again. The average scores of their K6 tests before and after the study were 5.77 (SD = 3.38, max = 12, min = 0) and 5.58 (SD = 3.71, max = 20, min = 0).

Procedure and Measurements

First, we explained the requirements of the study and installed our chatbot in each user's mobile phone or other devices. We then asked them about their daily activities and their emotions by the chatbot for three weeks every day. After the three-week journaling concluded, we held face-to-face interviews with all of them about their experiences with the chatbot. Each interview lasted for approximately 30 minutes. We recorded all the interviews and transcribed them with the participants' permission. This research was reviewed and approved by our institutional review board (ethics review ID: H31-013).

As discussed earlier, to measure the participants' mental states, they took K6, ([18]). K6 measures mental states by a six-item index and has been used in prior studies measuring several dimensions of psychological distress [17, 34]. On a five-point scale from zero to four, K6 asks a few questions such as whether a participant feels nervous, hopeless, or worthless. A higher score indicates greater psychological distress. Our participants took the K6 test twice, before and after the three-week study.

Although our initial goal was to develop a method that can identify severely distressed participants by the end of the study, the K6 scores of only two participants exceeded 13, which is the cut-off for identifying serious mental illness. Since developing a measure with only two samples is impossible, we focused on the change in the participants' mental status. In other words, we examined whether and how their use of language reflected changes in their mental states. To this end, we divided the participants into three groups based on the changes (improved, deteriorated, or no change) in their mental status before and after this experiment. We grouped those who experienced more mental health issues as participants (n=9) whose K6 scores increased after the experiment and labeled them as Deteriorated group. We defined those whose K6 scores decreased as mentally improved (n =14) and labeled them as Improved group, and those whose K6 scores were unchanged (n = 7) as mentally stable and labeled them as Unchanged group. Note that K6 takes an integer value, ranging from 0 to 24. There was no margin when dividing the participants into the three groups - participants whose K6 score increased or decreased even by one degree was assigned to the Improved or Deteriorated group.

Conversation Logs

We recorded all of the participants' conversations with the chatbot and compared their content among the three groups that were divided by the changes in their mental status before and after this study. Participants generally entered about five messages per day while interacting with the chatbot. With Linguistic Inquiry and Word Count (LIWC) text analysis software [40], which is commonly used in psychological fields, we calculated the following: word counts, positive emotion words, and negative emotion words, first-person pronouns, and third-person pronouns.

Word counts: Participants typically entered a daily average of five (from three to eight) messages in response to chatbot's journaling prompts. We summed up these text entries and counted the number of words each participant entered each day.

Use of positive/negative emotion words: By exploiting the content categories provided by LIWC, we calculated the usage ratio of positive emotion words, e.g., love, nice, and sweet, and negative emotion words, e.g., hurt, ugly, and nasty, to word counts. These emotional words are registered with LIWC. For example, "I love that nice flower," which calculates the rate of positive emotion words (love, nice), is 40% (2/5) and the rate of negative emotion words is 0%.

Use of first- and third-person pronouns: We calculated the first- and third person-person pronouns as the ratio of those numbers to the word counts. For instance, in "I had lunch with her old friend." the first-person pronoun (I) ratio is 14% (1/7), and the third-person-pronoun ratio (her) is 14% (1/7). To determine how much the participants talked about themselves instead of others, we calculated the ratio of first-person pronouns to third-person pronouns (the use of first-person pronouns per third-person pronouns). If this value exceeds 1, first-person pronouns are used more than third-person pronouns.

Use of concrete words: We also calculated the concreteness for each journal entry. We used the lexicon provided by Brysbaert et al. [6] that consists of 37,058 English words rated from 1 (very abstract) to 5 (very concrete). For example, "tennis" (4.43) is more concrete than "sports" (3.79). The journaling-concreteness rating is calculated by averaging the concreteness rating of each word in their conversational logs.

To analyze how the participants' use of language changed depending on fluctuations on their mental states, we extracted their conversational logs and conducted a mixed-model ANOVA. A Tukey HSD was then used for post-hoc analysis with two independent variables: experimental day (21 experiment days) and group (3 groups). The dependent variables were their word counts, positive emotion words, negative emotion words, concrete words, first-person pronouns, third-person pronouns and the ratio of first-person pronouns to third-person pronouns (the use of first-person pronouns per third-person pronouns).

4 RESULTS

First, we explored RQ1, which asked about the relationship between the changes in the mental status of the participants as well as the number of words they used when journaling, we conducted a 3 by 2 mixed-model ANOVA on the word counts. We found a significant main effect of the experimental day on word count (F=6.34, p<.0001, η^2 = 0.18, M=70.48->61.93), which means that the average word counts significantly decreased over time. There were no statistically significant main effects of group, and the interaction effects of the experimental day and group were also non-significant.

Second, we investigated whether the use of first- and third-person pronouns increased or decreased over time in conversations with a chatbot when people began to experience mental health problems (RQ2). Our ANOVA tests revealed a significant main effect on the experimental day (use of first-person pronouns: F=3.20, p<.0001, η^2 = 0.10, M=12.66->14.40, use of third-person pronouns: F=9.10, p<.0001, η^2 = 0.27, M=7.34->8.57), suggesting that their use of first- and third-person pronouns significantly increased over time. However, the main effect of group was not significant. There was also no significant interaction effect of experimental day and group. Furthermore, through the same test, we found a significant main effect of experimental day (F=4.28, p<.0001, η^2 = 0.15, M=2.66->2.50) for the ratio of the use of first-person pronouns to the use of third-person pronouns. This indicates that the increase in the use of third-person pronouns was much larger than that of first-person pronouns. However, the main effects of the group and the interaction effects of the experimental day and group were both non-significant.

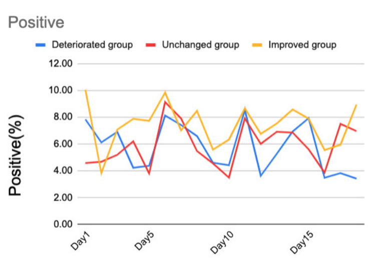

Figure 1: Use of positive words.

Third, we hypothesized in H1 that as people began to suffer health issues, they would start to use more negative emotion words and their use of positive words would decrease over time in the chatbot conversations. We conducted a mixed model ANOVA. Its results showed a significant main effect of the experimental day (F=3.46, p<.0001, η^2 = 0.20, M=6.36->6.16) and group (F=8.21, p<.0003, η^2 = 0.02) on the use of positive emotion words. This means that the use of positive emotion words significantly decreased over time (Figure 1). Regarding the significant effect among the three groups, post-hoc analyses showed that Deteriorated group's use of positive emotion words was significantly lower than Improved group's (p<.002), as was Unchanged group's (p<.0001), (Deteriorated group: M=5.26, SD = 4.25, Unchanged group: M=5.92, SD = 3.93, and Improved group: M=7.50, SD = 3.88), supporting H1. Similarly, we identified a significant main effect of experimental day (F=8.07, p<.0001, η^2 = 0.10, M=2.07->2.24) and the groups (F=8.21, p<.0003, η^2 = 0.03) on the use of negative emotion words, which suggests that their use significantly increased over time (Figure 2). According to post-hoc analyses, Deteriorated group's use of negative emotion words was significantly higher than Unchanged group's (p<.0003) (Deteriorated group: M=2.80, SD = 2.63, Unchanged group: M=2.59, SD = 3.00, and Improved group: M= 1.79 SD = 2.16), again supporting H1. However, the interaction effects of the experimental day and group on the use of positive and negative words were non-significant.

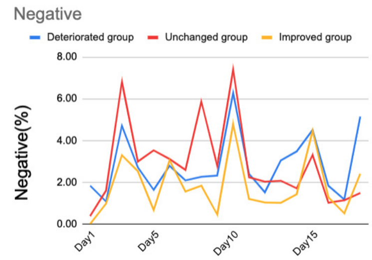

Figure 2: Use of negative words.

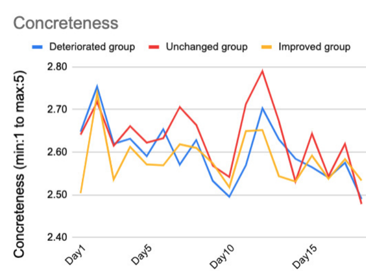

Figure 3: Use of concrete words.

Finally, in H2, we hypothesized that as people experienced more mental health issues, the use of concrete words would decrease over time in the chatbot conversations. Through the mixed model ANOVA, we found a significant main effect of experimental day (F=3.67, p<.0001, η^2 = 0.11, M=2.62->2.57) and groups (F=4.02, p<.05, η^2 = 0.01) on the use of concrete words, which means that their use significantly decreased over time (Figure 3). Post-hoc analyses showed that Deteriorated group's use of concrete words was significantly lower than Unchanged group's ($p<.002$) and Deteriorated group's was higher than Improved group's ($p<.002$), (Deteriorated group: $M=2.60$, $SD = 0.17$, Unchanged group: $M=2.63$, $SD = 0.18$, and Improved group: $M=2.58$, $SD = 0.16$). The interaction effects of the experimental day and group were non-significant. These results are somewhat confusing because the participants' use of concrete words failed to show a consistent trend with their mental status. Overall, although our result supports H2, the use of concrete words decreased not only for the participants whose mental condition deteriorated but also for those whose condition improved.

5 DISCUSSION

The purpose of this study was to investigate how the changes of mental states affect human self-disclosure contents to a chatbot. By using indicators that in previous researches detected depression from written texts, we examined whether similar tendencies are found in chatbot conversations.

Our results showed that participants whose mental status deteriorated gradually used fewer positive emotion and concrete words but used more negative emotion words when describing their daily experiences and feelings to a chatbot. This result is consistent with previous studies. As people begin to experience mental health concerns, their use of negative emotion words increases, and their use of positive words decreases over time [7], which means that H1 and H2 are correct. On the other hand, the number of words decreased and the use of first- and third-person pronouns increased, regardless of the change in mental states before and after the experiment. In this section, we discuss our results in the light of chatbot characteristics and journaling properties.

Why did the number of text entries decrease? Although several studies have identified the benefits of journaling, they have also reported the difficulties of continuing to journal because finding interesting topics to write about is often difficult [42]. Unlike social media, where people can freely choose when to write, in chatbots, people may feel external pressure even if they don't want to write or have nothing to say. Indeed, some participants noted in their post-study interviews that they had many things to write about in the first week, including describing themselves so that the chatbot can understand their feelings better. They complained that it gradually became more difficult to find something to share because their daily lives are highly repetitive.

In social media, people are typically aware of their audience and thus pay attention to what they write [3]. To build common ground with their somewhat broad audience, they would need to explain contextual information so that their audience can understand the contents [26]. In contrast, with a chatbot, after introducing themselves and briefly explaining their situation/background, they tend to drop redundant information in subsequent exchanges, assuming that the chatbot remembers what they previously said. As the Computers Are Social Actors (CASA) paradigm [30] indicates, people tend to apply social norms of human-to-human interaction when interacting with computer agents. Thus, people might avoid giving redundant information to a chatbot. As such, we believe that the intrinsic characteristics of chatbot discouraged our participants to write.

Why did the third-person pronouns increase? Prior research showed that as people become more depressed, their use of first-person pronouns in social media posts tends to increase and their use of third-person pronouns tends to decrease [7]. According to [29], when people are depressed, they become too busy focusing on their own feelings and have little space for others.

In this study, the usage rate of first- and third-person pronouns while journaling with a chatbot increased over time regardless of changes in mental status. This could be explained by the significant drop in word counts. The additional result on the significant increase in the use of third-person pronouns compared to the first-person pronouns is explained by the continuous use of chatbots. In our study, the chatbot asked the participants to report their moods/feelings and what caused them. When participants started using the chatbot, they seemed to talk more about themselves, such as their daily routines and hobbies, as if they are introducing themselves to a stranger. As they got familiar with the chatbot, they seemed to talk more about the key persons (e.g., romantic partners, professors) around them who strongly affected their mood and feelings. They explained conversations they had with such individuals and how they mattered to them. This is consistent with previous findings where one's moods are often strongly influenced by one's relationships with others [15, 20, 37].

Why did the use of concrete words decrease? In terms of the use of concrete words, those whose mental health improved used the fewest concrete words, followed by those whose mental health deteriorated and those who had no change in mental status: Unchanged group > Deteriorated group > Improved group. We speculate that Deteriorated group and Improved group had different reasons for reducing the use of concrete words. Perhaps Deteriorated group decreased them because they tended to avoid talking about the details of their feelings when they are depressed [13]. On the other hand, Improved group also gradually decreased their use of concrete words. Journaling is popular among those who are mentally distressed or unstable [42]. It might be boring for people who are experiencing good mental health. In fact, some participants in Improved group commented in the post-experimental interviews that they sometimes were annoyed when the chatbot repeatedly asked about their mood because that topic was boring. As inferred from this comment, perhaps journaling was a less interesting or engaging activity for those who were experiencing good mental health, reducing their motivation to write about their experiences or feelings.

Limitations and Future Directions

This work has several limitations that should be acknowledged.

Firstly, our sample did not include clinically depressed individuals. A study with clinically depressed individuals may have provided a more valid "ground truth" for understanding how depressed individuals disclose to a chatbot. In addition, we divided the participants into groups based on whether their K6 scores increased or decreased, instead of using the absolute K6 values. Since we were investigating how people's interaction with a chatbot changed as their mental status began to deteriorate *before the symptoms develop*, we recruited participants who were not greatly distressed. Although we found that as they began to experience mental health issues, their writing patterns resembled the other patterns, further study is required to see whether a group of people with severe mental health problems can be detected based on the absolute values of K6.

Secondly, although we used a mixed-model ANOVA to analyze the data, the assumption for the ANOVA, namely the homogeneity of variances and normality tests, did not hold. This limitation may be eliminated by increasing the sample size.

As Future Directions, combining other conversation topics with journaling, including gratitude journaling [11], may uncover relationships between tendencies that appear in that task and the K6 scores. Furthermore, by combining the results of this study with counseling chatbots, perhaps people whose mental status has deteriorated can be led to the treatment phase.

Finally, to further understand the relationships between the chatbot and the results, a study that compares different types of chatbots or a Wizard of Oz study that replaces the chatbot with a real human might be useful.

6 CONCLUSION

We investigated the relationship between the tendencies in textual data that people input to a chatbot and mental status. Our three-week study examined whether we can assess mental status from daily conversation data with chatbots. By using indicators that detect depression from written data, we reached the following conclusions: those whose mental status gradually deteriorated used fewer positive emotion and concrete words but used more negative emotion words when describing their daily experiences and feelings to a chatbot. Our findings further suggest that the indicators for detecting depression from conversations with a chatbot may be different from that of social media posts because people write differently depending on the audience and the media they use.

ACKNOWLEDGMENTS

This work is supported by Grant for Scientific Research (A) 17H00771 from Japan Society for the Promotion of Science (JSPS). We thank Prof. Eiji Aramaki and Shuntaro Yada at Nara Institute of Science and Technology for early feedback in the study. Finally, we are thankful for all the anonymous reviewers' comments and suggestions to help polish this paper.

REFERENCES

[1] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4 (2016), 463–476.

[2] Karen A Baikie and Kay Wilhelm. 2005. Emotional and physical health benefits of expressive writing. *Advances in Psychiatric Treatment* 11, 9 (2005), 338-346.

[3] Natalya N Bazarova and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication* 64, 4 (2014), 635–657.

[4] Samuel Bell, Clara Wood, and Advait Sarkar. 2019. Perceptions of Chatbots in Therapy. *In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, LBW1712.

[5] Roy Bhakta, Maggi Savin-Baden, and Gemma Tombs. 2014. Sharing Secrets with Robots?. *In EdMedia: World Conference on Educational Media and Technology*, 1 (2014), 2295-2301.

[6] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 thousand generally known English Word Lemmas. *Behavior Research Methods* 46, 3 (2014), 904–911.

[7] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 7 (2013), 128–137.

[8] Jeffrey F. Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. 2009. *In Affective Computing and Intelligent Interaction*, 9 (2009).

[9] Fionn Delahunty. Delahunty, Ian D Wood, and Mihael Arcan. 2018. First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot. *In AICS*, 12 (2018), 327–338.

[10] S Divya, V Indumathi, S Ishwarya, M Priyasankari, and S Kalpana Devi. 2018. A self-diagnosis medical chatbot using artificial intelligence. *Journal of Web Development and Web Designing* 3, 1 (2018), 1–7.

[11] Robert A Emmons and Robin Stern. 2013. Gratitude as a psychotherapeutic intervention. *Journal of clinical psychology* 69, 8 (2013), 846–855.

[12] 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 (2017), e19.

[13] Yifan Gao, Yuchong Zhong, Daniel Preotiuc-Pietro, and Junyi Jessy Li. 2019. Predicting and analyzing language specificity in social media posts. *In AAAI* 33, 7 (2019).

[14] John Hart, Jonathan Gratch, and Stacy Marsella. 2013. How virtual reality training can win friends and influence people. *Human Factors in Defence.Ashgate* 21, 1 (2013), 235-249.

[15] Takeshi Hashimoto. 1997. Effects of Interpersonal Relationships on Mental Health: From a Perspective of the Interpersonal Stress Arousal Process Model. *The Japanese Journal of Experimental Social Psychology* 37, 1 (1997), 50-64.

[16] Justin Hunt and Daniel Eisenberg. 2010. Mental health problems and help-seeking behavior among college students. *Journal of adolescent health* 46, 1 (2010), 3–10.

[17] Ronald C Kessler, Jennifer Greif Green, Michael J Gruber, Nancy A Sampson, Evelyn Bromet, Marius Cuitan, Toshi A Furukawa, Oye Gureje, Hristo Hinkov, Chi-Yi Hu, Carmen Lara, Sing Lee, Zeina Mneimneh, Landon Myer, Mark Oakley-Browne, Jose Posada-Villa, Rajesh Sagar, Maria Carmen Viana, and Alan M Zaslavsky. 2010. Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative. *Int J Methods Psychiatr Research* 19, 1 (2010), 4–22.

[18] Ronald C. Kessler, Gavin Andrews, Lisa J Colpe, Hiripi EE, Daniel K Mroczek, Sharon-Lise Normand, Ellen E. Walters, and Alan M Zaslavsky. 2002. Short Screening Scales to Monitor Population Prevalences and Trends in Non-specific Psychological Distress. *Psychological Medicine* 32, 8 (2002), 959-76.

[19] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* 86, 5 (2019), 1-12.

[20] Cynthia G Last, David H Barlow, and Gerald T O'Brien. 1984. Precipitants of agoraphobia: Role of stressful life events. *Psychological Reports* 54, 4 (1984), 567-570.

[21] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Selfdisclosure through a Chatbot. *In Proceedings of the 2020 CHI Conference on HumanFactors in Computing Systems (CHI '20)*, 4 (2020), 1–12.

[22] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37, 8 (2014), 94–100.

[23] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI* 4, 10(2017), 1-9.

[24] Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2016. Anonymity, intimacy and self-disclosure in social media. *In Proceedings of the 2016 CHI Conference on HumanFactors in Computing Systems (CHI '16)*, 5 (2016), 3857–3869. DOI:http://dx.doi.org/10.1145/2858036.2858414

[25] Marina Marcus, Yasamy M Taghi, van Ommeren, Dan Chisholm, and Shekhar Saxena. 2012. Depression: A global public health concern. *WHO Department of Mental Health and Substance Abuse*, 10(2012).

[26] Alice E Marwick and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114-133.

[27] Noboru Matsumoto and Satoshi Mochizuki. Overgeneral autobiographical memory and depression: Review and future directions *Japanese Psychological Review* 55, 1 (2012), 459-483

[28] Marc L Molendijk, Lotte Bamelis, Arnold A P van Emmerik, Arnoud Arntz, Rimke Haringsma, and Philip Spinhoven. 2010. Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behaviour Research and Therapy* 48, 9 (2009), 44–51. doi: 10.1016/j.brat.2009.09.007

[29] Mohammed Al-Mosaiwii and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 1 (2018), 529-542.

[30] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, 4 (1994), 72–78.

[31] Jooyoung Oh, Sooah Jang, Hyunji Kim, Jae-Jin Kim. 2020. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. *International Journal of Medical Informatics Volume* 140, 8 (2020), https://doi.org/10.1016/j.ijmedinf.2020.104171

[32] Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and HoJin Choi. 2017. A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation. *In 2017 18th IEEE International Conference on Mobile Data Management (MDM)*, 371–375.

[33] Thomas E Oxman, Stanley D Rosenberg, and Gary J Tucker. 1982. The language of paranoia. *The American Journal of Psychiatry* 139, 3 (1982), 275–82.

[34] Judith J Prochaska, Hai-Yen Sung, Wendy Max, Yanling Shi, and Michael Ong. 2012. Validity study of the K6 scale as a measure of moderate mental distress based on mental health treatment need and utilization. *International journal of methods in psychiatric research* 21, 2 (2012), 88–97.

[35] Abhilasha Ravichander and Alan W Black. 2018. An Empirical Study of Self-Disclosure in Spoken Dialogue Systems. *In Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 7 (2018), 253–263.

[36] Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*, 10 (2013), 1348–1353.

[37] Karen S Rook. 1984. The negative side of social interaction: Impact on psychological well-being. *Journal of Personality and Social Psychology* 46, 5 (1984), 1097-1108.

[38] Stephanie S Rude, Eva-Maria Gortner, and James W Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18, 12 (2004), 1121-1133.

[39] Denise M Sloan. 2005. It's all about me: self-focused attention and depressed mood. *Cognitive Therapy and Research* 29, 6 (2005), 279–288. doi: 10.1007/s10608-005-0511-1

[40] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 2, 3 (2010), 24–54.

[41] Bea Tiemens, Johan Ormel, Jack A Jenner, Klaas van der Meer, Titus W D P van Os, R H van den Brink, Annet Smit, and Wim van den Brink. 1999. Training primary-care physicians to recognize, diagnose and manage depression: does it improve patient outcomes? *Psychological Medicine* 29, 8 (1999), 833–845.

[42] Philip M Ullrich and Susan K Lutgendorf. 2002. Journaling about stressful events: Effects of cognitive processing and emotional expression. *Annals of Behavioral Medicine* 24, 2 (2002), 244 –250.

[43] Allison Utley and Yvonne Garza. 2011. The therapeutic use of journaling with adolescents. *Journal of Creativity in Mental Health* 6, 1 (2011), 29–41.