Parsing the Landscape of AI Based Mental Health Applications: Thoughts and Considerations on Feedback from End Users

Kumari Davis

The Pennsylvania State University
University Park, PA 16802
kzl5736@psu.edu

Abstract

Given the negative impact of the COVID-19 pandemic on peoples' mental health, there has been a recent surge in the development and use of Artificial Intelligence based mental health applications. While these applications aim to augment and support the existing mental health provider network with useful mobile-based interventions, their effectiveness in improving peoples' mental health has not been carefully evaluated. In this paper, we address this research gap by conducting an AI-driven analysis of English language reviews left by human end users for two popular AI-based mental health applications, *Woebot* and *Sanvello*. We develop topic modeling approaches to infer a set of topics from the corpus of user reviews, where each topic represents a distinct usability concern or benefits experienced by end users. We also provide a high-level discussion about potential design flaws that could be used to improve the effectiveness of AI-based mental health applications.

Introduction

Even before the start of the COVID-19 pandemic, mental illness and depression were the second leading cause of disability worldwide (Walker, McGee, and Druss 2015). In the United States alone, 20.78% of adults (approximately 50 million people) experienced a mental illness in the year 2019-2020 (MHA National 2023). On a global scale, 280 million people suffered from depression, whereas 700,000 people died due to suicide in that year (WHO 2021).

Unfortunately, marginalized communities in America are worst hit by mental illnesses such as anxiety, depression, and suicide, e.g., Native American young people experience suicide rates at $\sim$3X that of the national average. Similarly, LGBTQ+ young people are 5X more likely to attempt suicide than their heterosexual counterparts, whereas the rise in mental illnesses among young African American people has been called a "*health crisis*" by the National Institute of Mental Health (NIMH) (Tang et al. 2019). This situation has been further exacerbated since the onset of the COVID-19 pandemic, during which time the prevalence of major depressive disorders and anxiety disorders increased by 28% and 26%, respectively (especially among young women who were forced to work from home during the pandemic) (Santomauro et al. 2021).

Despite the urgent need for medical care, almost 60% of American youth suffering from major depression do not receive any mental health treatments because of two major reasons. First, there is a significant mismatch between supply and demand, e.g., in the United States, there are an estimated 350 individuals for every one mental health provider (MHA National 2023). Second, victims of poor mental health are wary of stigmatizing and discriminatory attitudes against people suffering from mental illnesses, because of which they do not feel comfortable talking about their problems with anyone (Henderson, Evans-Lacko, and Thornicroft 2013). Tragically, a large majority of people experiencing mental illnesses suffer silently because of both these reasons.

To address these pressing challenges, there has been a recent surge in the development and use of Artificial Intelligence (AI) enabled mental health smartphone applications such as *Woebot*, *Wysa*, and *Sanvello*, etc., as shown in Figure 1. These applications are beneficial in two respects: (i) they supplement the existing mental health provider network by acting as the first point of care for people suffering from mental illnesses, i.e., they can be used to manage mental health conditions such as anxiety or depression either on their own—enabling individuals to learn about and self-manage their mental health—or in conjunction with more traditional therapies; and (ii) these applications also help in removing some of the barriers of cost and inhibition experienced by people suffering from mental illnesses, as they find it easier to share their feelings with an AI "machine" as compared to a human mental health provider.

While these AI-enabled applications hold significant potential for the delivery of high-efficacy mental health interventions, there has been extensive prior research that *does not* find any "convincing evidence" for the use of these mobile AI-enabled smartphone applications leading to reduced anxiety, depression, thoughts of suicide, etc (Goldberg et al. 2022). However, most of these evaluative studies focus on quantitative usability metrics, and therefore, these studies are not centered on end-users' experiences of using these AI-enabled applications, and their perceptions about the usefulness (or lack, thereof) of these applications. We strongly argue that human end-user perspectives about the effective-

Figure 1: *Woebot*, *Sanvello*, and *Wysa* - Three highly popular AI-based mental health smartphone applications with a combined user base of over 8 million users worldwide.

ness of these applications are key to uncovering design flaws (if any), which can then be used to inform the design of better mental health applications.

In this paper, we take the first steps towards addressing this research gap by conducting an AI-driven analysis of English language reviews left by human end users for two popular AI-based mental health applications, *Woebot* and *Sanvello*. In particular, this paper makes the following novel contributions: (i) we scrape all (English language) user reviews of *Woebot* and *Sanvello* left by human end-users on Google Play Store. (ii) Given the large corpus of user reviews (∼25K reviews) for these applications, conducting manual qualitative coding and thematic analysis was infeasible. As a result, we developed Latent Dirichlet Allocation (LDA) based topic models (Blei, Ng, and Jordan 2003) on the corpus of user reviews in order to find a set of themes (or "topics") that represent different usability concerns or benefits experienced by human users of the applications. (iii) Based on the themes uncovered via topic modeling, we provide a high-level discussion of pros and cons regarding the usability and effectiveness of these AI-based mental health applications. This work represents a preliminary inquiry into this big research area of understanding user perspectives on the effectiveness of AI-based mental health applications, and we end with a discussion of future work.

Methodology: Analyzing User Reviews

In this section, we provide a detailed description of our methodology for analyzing user reviews left by human users of widely used AI-based mental health smartphone applications.

Inclusion Criteria. For the purposes of this study, we restricted our attention to AI-based smartphone applications that were listed on the Google Play App Store with an explicit "Mental Health" tag, and which had an average user rating of 4.5 and above.

Through this procedure, we identified three AI-based mental health applications with a combined user base of 8 million people worldwide (Figure 1):

- **Woebot** (500K users) is an AI chatbot that uses principles of Cognitive Behavioral Therapy (CBT). Woebot guides users through managing distressing thoughts and feelings. Through periodic check-ins, Woebot prompts users to enter their mood (and details explaining their mood), and responds by suggesting tools, skills, and strategies to help end users. In addition to messaging, users can view a chart of their mood entries over time, and view psychoeducational media. Woebot is not designed to be used in an emergency or to manage psychiatric crises.
- **Wysa** (4.5 million users) is an AI chatbot that is designed to help users with a variety of issues, including depression, anxiety, sleep, issues facing the LGBTQ+ community, etc. When chatting with Wysa, the user is presented with multiple response options as well as a text box to type customized responses. From there, Wysa guides users through cognitive reframing, breathing exercises, and other strategies depending on how they report feeling and what is appropriate for that situation. Conversations are not saved in the app, so there is no login or account required.
- **Sanvello** (3 million users) uses principles of cognitive behavioral therapy (CBT) to help users with symptoms of anxiety, depression, or stress. Users complete a self-assessment questionnaire and receive anxiety, stress, and depression scores. Sanvello uses these scores to track the user's progress over time and to provide personalized activities and exercises. Sanvello also provides a community message board where users may connect with each other by posting, liking, and chatting on each other's posts.

Table 1: Descriptive Statistics of the Scraped *Woebot* & *Sanvello* User Review Datasets

User Rating	Woebot	Sanvello
1	199	358
2	97	202
3	121	303
4	666	1032
5	4050	5145
Grand Total	**5133**	**7040**

For our preliminary analysis reported in this paper, we only focus on analyzing the user reviews of Woebot and Sanvello. We choose to limit our attention on these two applications since there is a dearth of work done on these applications (by contrast, a significant body of work has already been done on Wysa). In future work, we aim to expand our topic modeling based analysis to Wysa and other AI-based mental health applications.

Scraping User Reviews. Next, we created a scraper for gathering all user reviews for Woebot and Sanvello on the Google Play App store. Our scraping module was built using the Google Play Scraper API[1], and it scraped all user reviews written in monolingual English text by users based in the United States. In addition to getting the text-based reviews,

[1]https://pypi.org/project/google-play-scraper/

```
{
    "reviewId": "8a9f8c22-17b7-439c-8057-082c6569c49f",
    "userName": "Mu Pumpelmuse",
    "content": "Fantastic app that has helped me a lot! I
       struggle with anxiety, depression, and OCD, but Woebot
       has been a big help. I've actually found myself
       catching negative thoughts and reframing them on my
       own now, even when I'm not using the app. It's also
       completely free! Two suggestions: maybe add a way to
       save the rewrites of our distortions so we can look
       back at them, and make a dedicated past lesson section
       so it's easy to look over old lessons. The current
       system is clunky there.",
    "score": 5
}
```


Figure 2: Snapshot of user review data scraped from Woebot.

we also scraped the user rating (out of 5 stars) that accompanied the English language review. In total, we scraped 5,133 user reviews for Woebot and 7,040 user reviews for Sanvello. Figure 2 shows an actual 5-star rated review scraped from Woebot.

Table 1 shows that both datasets are heavily skewed towards higher user ratings, e.g., 5-star reviews account for 79% and 73% of total reviews on *Woebot* and *Sanvello*, respectively. On the other hand, 1-star reviews account for ∼4% and ∼5% of total reviews on *Woebot* and *Sanvello*, respectively. This is expected since our inclusion criteria favored highly-rated applications (above 4.5 average user rating) on the Google Play Store.

Dataset Preprocessing. Both our user review datasets contain a lot of noise. Different from a traditional text-based corpus, user reviews on Google Play Store are typically noisy, as they can contain a significant amount of typographical errors, emojis, and/or Internet slangs. In order to ensure that we only train our topic modeling approaches on clean (noise-free) user review data, it is important to pre-process our dataset, as we describe below.

First, we eliminated all user reviews that were less than five characters in length. We argue that reviews that are less than five characters in length can only be used to express base-level emotions (e.g., happiness or sadness expressed by emojis or single words). Instead, we wanted to capture deeper-level topics corresponding to usability concerns or benefits experienced by users of applications.

Second, we removed all newline, single-quote characters along with any emojis (from a pre-defined set of emojis) that occur in the text of user reviews. Third, we tokenize each user review into a set of lowercase word-level tokens. Fourth, to capture frequently occurring phrases of text that occur in user reviews, we created bigram, trigram, and 4-gram models using the tokenized words. Finally, we also remove any stop-words (that occur in the Natural Language Toolkit, or NLTK stopword list (Bird, Klein, and Loper 2009)) and lemmatize our dataset.

Topic Modeling. After pre-processing our dataset, we used topic modeling approaches to extract hidden topics (or themes) from our corpus of user reviews.

In particular, we use Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) to build our topic model, which

Figure 3: Word Clouds associated with uncovered topics from positive reviews of Woebot.

Figure 4: Word Clouds associated with uncovered topics from negative reviews of Woebot.

is a popular form of statistical topic modeling. In LDA, documents (or user reviews, in our case) are represented as a mixture of topics, and each topic is a bunch of words. Those topics reside within a hidden layer (which is called as a latent layer). At a high level, LDA looks at a document (or user review) to determine a set of topics that are likely to have generated that collection of words. Therefore, if a document uses certain words that are contained in a topic, we infer that the document is about that particular topic. For this pa-

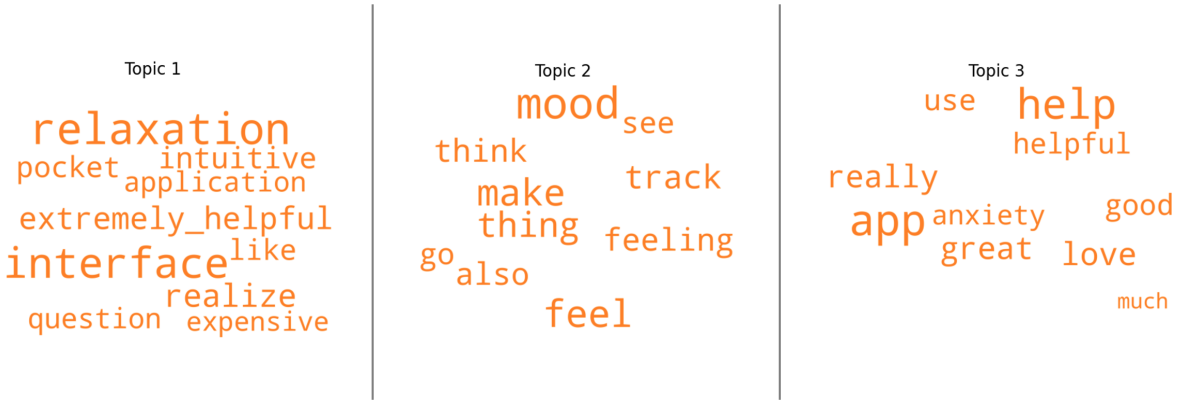

Figure 5: Word Clouds associated with uncovered topics from positive reviews of Sanvello.

Figure 6: Word Clouds associated with uncovered topics from negative reviews of Sanvello.

per, we implemented LDA using spaCy[2] and Gensim[3], two widely used libraries for NLP and topic modeling research.

Results & Discussion

We provide results in two phases. First, we illustrate the results obtained from topic modeling on the user reviews for Woebot. We analyze the uncovered topics to highlight design flaws and considerations for future app improvement. Second, we illustrate the results obtained from topic modeling on the user reviews for Sanvello, and highlight design flaws and considerations for future app improvement.

For both Woebot and Sanvello, we generate two separate LDA models. The first of these LDA models is trained only on positive user reviews (user rating ≥ 4), whereas the second LDA model is trained only on negative user reviews (user rating ≤ 3). In theory, the topic model trained on positive-only user reviews should uncover hidden topics corresponding to useful attributes or positive usability aspects of the mental health applications (as perceived by users). Similarly, the topic model trained on negative-only user reviews should uncover hidden topics corresponding to negative usability aspects of the mental health applications (as perceived by users).

Results on Woebot Reviews. Figures 3 and 4 show the word clouds associated with topics uncovered by our 4-topic LDA model trained on positive and negative Woebot reviews, respectively. In Figure 3, we observe that each topic describes a distinct useful characteristic of Woebot. For example, the word cloud of Topic 1 (Figure 3) seems to reflect Woebot's ability to make users *feel better*, and its *practical* nature. Similarly, the word cloud of Topic 2 (Figure 3) seems to reflect Woebot's ability to provide *Cognitive Behavioral Therapy (CBT)* in a *helpful*, *fun*, and *informative manner*.

Similarly, in Figure 4, we observe that each topic describes a distinct negative characteristic of Woebot. For example, the word cloud of Topic 0 (Figure 4) seems to reflect Woebot's inability to understand long English sentences, whereas the word cloud of Topic 1 reflects that some users find it worthless and useless app.

Results on Sanvello Reviews. Figures 5 and 6 show the word clouds associated with topics uncovered by our LDA model trained on positive and negative Sanvello reviews, respectively. In Figure 5, we observe that each topic describes a distinct useful characteristic of Sanvello. For example, the word cloud of Topic 1 (Figure 5) reflects Sanvello's ability to induce *relaxation*, its *intuitive interface*, and its *extremely helpful* nature.

Similarly, in Figure 6, we observe that each topic describes a distinct negative characteristic of Sanvello. For example, the word cloud of Topic 2 (Figure 6) reflects the fact that users do not appreciate that most useful features of Sanvello are not free, and need to be *paid for*.

Conclusion

We conduct an analysis of user reviews for *Woebot* and *Sanvello*. We develop preliminary LDA-based topic models to infer a set of topics from the corpus of user reviews, where each topic represents a distinct usability concern or benefits experienced by end users. In the future, we hope to expand our analysis to additional well-liked AI-based mental health applications.

[2]https://spacy.io/

[3]https://radimrehurek.com/gensim/

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Goldberg, S. B.; Lam, S. U.; Simonsson, O.; Torous, J.; and Sun, S. 2022. Mobile phone-based interventions for mental health: A systematic meta-review of 14 meta-analyses of randomized controlled trials. *PLOS digital health*, 1(1): e0000002.

Henderson, C.; Evans-Lacko, S.; and Thornicroft, G. 2013. Mental illness stigma, help seeking, and public health programs. *American journal of public health*, 103(5): 777–780.

MHA National. 2023. The State Of Mental Health In America. https://www.mhanational.org/issues/state-mental-health-america.

Santomauro, D. F.; Herrera, A. M. M.; Shadid, J.; Zheng, P.; Ashbaugh, C.; Pigott, D. M.; Abbafati, C.; Adolph, C.; Amlag, J. O.; Aravkin, A. Y.; et al. 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, 398(10312): 1700–1712.

Tang, Y.; Sinha, K. K.; Moen, A.; and Ertekin, N. 2019. Delivering Mental Healthcare to the Underserved Communities: Evaluating the Potential of Social Technologies.

Walker, E. R.; McGee, R. E.; and Druss, B. G. 2015. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. *JAMA psychiatry*, 72(4): 334–341.

WHO. 2021. Depression: Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/depression.