Towards Using AI to Augment Human Support in Digital Mental Healthcare

Prerna Chikersal
School of Computer Science
Carnegie Mellon University, US
prerna@cmu.edu

Gavin Doherty
School of Computer Science & Statistics
Trinity College Dublin, IRL
gavin.doherty@tcd.ie

Anja Thieme
Healthcare Intelligence
Microsoft Research Cambridge, UK
anthie@microsoft.com

Submission to *CHI 2020 workshop on Technology Ecosystems: Rethinking Resources for Mental Health*

Abstract

To address the need for more access to, and increase the effectiveness of, mental health treatment, internet-delivered psychotherapy programs such as iCBT are shown to achieve clinical outcomes comparable to face-to-face therapy. While offering ubiquitous access to healthcare, a key concern of digital therapy programs is to sustain users' engagement with the treatment to attain desired benefits. Research has demonstrated that including a trained 'human supporter' to the digital mental health ecosystem can provide useful guidance and motivation to its users, and lead to more effective outcomes than unsupported interventions. Within this context, we describe early research that makes use of machine learning (ML) approaches to better understand *how* the behaviors of these human supporters may benefit the mental health outcomes of clients; and how such effects could be maximized. We discuss new opportunities for augmenting human support through personalization, and the related ethical challenges.

Author Keywords

Mental health; digital behavioral intervention; ethics; responsible AI; machine learning; NLP; data mining.

CSS Concepts

•Human-centered computing → **Human computer interaction; •Computing methodologies** → *Machine learning;*

Introduction

Our research applies unsupervised machine learning, and statistical and data mining methods to identify what supporter behaviors – delivered as part of an online psychotherapy intervention – correlate with better client mental health outcomes [1]. Our analysis is based on a fully anonymized dataset of 234,735 supporter messages to clients (sent by 3,481 supporters to 54,104 clients) from an established internet-delivered Cognitive Behavioral Therapy (iCBT) program for "depression and anxiety" [7]. This program is one of the most frequently used treatments on the SilverCloud platform (www.silvercloudhealth.com). Accessed online or via mobile, the program presents a self-guided intervention with seven core psycho-educational and psycho-therapeutic modules that are delivered through interactive multi-modal contents and tools [2]. Clients work through the program at their own pace and time, and with regular, personalized feedback messages sent by a trained *human supporter*.

Aiming to better understand what linguistic strategies characterize the feedback messages of supporters whose clients have better mental health outcomes; we employed a set of ML and data mining methods.

Understanding Successful Support Strategies

To better understand what support strategies characterize supporter messages that are correlated with better clinical outcomes for clients, we first needed to identify what constitutes 'successful support messages' based on clinical outcomes. To this end, we computed 4 clinical outcomes by averaging post-message change in PHQ-9 (for depression) [3] and GAD-7 (for anxiety) [4] scores across the messages sent by each supporter (see sidebar). We then use these outcomes as features in *K*-means clustering, through which we obtain k=3 clusters of supporters whose messages are generally linked with either 'high', 'medium' or 'low' improvements in client outcomes. We hypothesize that there are differences in the messages sent by supporters in the 'high' versus 'low' outcome clusters; and that these differences will help us identify what may constitute more effective support strategies.

Linguistic Strategies used in More Successful Messages

As a next step, we want to identify what semantic or linguistic strategies occur significantly more often in the messages of supporters in the 'high' outcome cluster; and this result needs to be consistent independent of differences across clients (*e.g.*, variations in their mental health state, program use). To analyze the content of the supporter messages, we used a lexicon-based text-mining approaches that extract comprehensive text characteristics without risking to identify any text content so as to preserve anonymity.

Amongst others, this included the: (i) use of the NRC Emotion Lexicon [6] to capture the *sentiment* and *emotional tone* of a message by extracting the percentages of positive and negative words, and words related to eight emotion categories (*e.g.,* fear, joy, anger); (ii) extraction of first person plural pronouns (*e.g.*, we, us, our) as indicator of a *supportive alliance*; (iii) uses of *encouraging phrases* (*e.g.,* 'well done', 'good job'); and (iv) use of the Regressive Imagery Dictionary [5] to assess percentages of words related to mental health processes such abstraction (e.g., know, thought), social behavior (*e.g.*, call, say, tell), or temporal references (*e.g.*, when, now, then).

Change & Improvement Rates as Clinical Outcomes

Message-level Change (MC):

The clinical score after a message is highly dependent on the clinical score before that message, as clients that have more severe symptoms before the message also tend to improve more on average after the message. Hence, we measure Message-level Change as the difference between actual change and the expected change given the client score before the message. That is, for each supporter *S* with *NM* messages, compute:

$$\frac{1}{NM}\sum_{r=1}^{NM}(actual\_change_m - expected\_change_m)$$

$$actual\_change_m = score\_before(m) - score\_after(m)$$

$$expected\_change_m = score\_before(m) - \mathbb{E}(score\_after(m)|score\_before(m))$$

Message-level Improvement Rate (MR):

If Message-level Change > 0, then the client improved more than expected post-message, and we label the message as "improved (1)". Otherwise, we label the message as "not improved (0)". For each supporter S with NM messages, we average these labels across all messages to compute this outcome.

Client-level Change (CC):

While MC captures changes in clinical scores across all messages by a supporter, CC normalizes these changes across all clients of the supporter. For each supporter *S*, we first compute the MC for each client of *S* separately using the messages that *S* sent to them. Then, we average the MCs per client across all clients of *S* to get CC. For example, if a supporter sends 6 messages to client A whose change is +1 after each messages and 4 messages to client B whose change is always 0, the MC will be $\frac{6}{6+4}$ = 0.6 while the CC will be $\frac{\left(\frac{6}{6}\right)+\left(\frac{0}{4}\right)}{2}$ = 0.5. Thus, MC can be high even when only a few clients improve, whereas CC will only be high when these improvements are consistent across all/ many clients. This makes CC more robust to a single client's changing situations or symptoms.

Client-level Improvement Rate (CR):

For each supporter *S* with clients *NC*, we first compute the average Message-level Improvement Rate using messages *S* sent to each client separately, and then sum these rates across all clients and divide by the total number of clients.

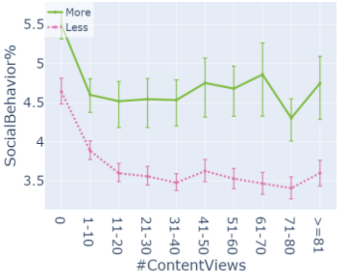

Figure 1. Mean percentage of words in ‘more’ and ‘less’ successful support messages that are associated with Social Behavior (here plotted across at clients with different frequencies of looking at content pages of the iCBT program: ContentViews).

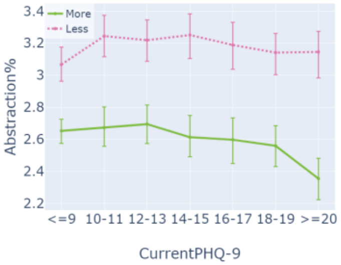

Figure 2. Mean percentage of words in ‘more’ and ‘less’ successful support messages that are associated with Abstraction (here plotted across different client mental health scores of their: CurrentPHQ−9). In other words, independent of variations in clients’ current mental health, more successful message have less words of abstraction associated with them.

Our semantic analysis revealed statistically significant findings that more successful supporter messages consistently used *more positive and less negative words,* and had less occurrences of negative emotions conveying *sadness* and *fear*. Further, more successful messages consistently employed *first person plural pronouns more frequently* than less successful messages, and consistently contained significantly *more encouraging phrases*. For our mental process variables, we found that more successful messages consistently employed *more words associated with social behavior* and *less words associated with abstraction* (see Figures 1 and 2; and [1] for details).

Client Context Specific Support Strategies

The above analysis indicates characteristics of support messages that tend to correlate with improved client outcomes independent of differences in the clients’ specific context (*e.g.*, their mental health, platform use), when treating different context variables in isolation. Next, we want to better understand the more complex relationship that likely exists between the use of support strategies and various client context variables. In other words, how may considering the *combination of multiple context* variables shift how salient a specific support strategy is in messages associated with either ‘high’ or ‘low’ improvements in client outcomes. We believe that identifying such relationship patterns could enable a more effective tailoring of support strategies to specific client contexts.

To identify multi-dimensional context related strategy patterns, we needed to discover associations between multiple client contexts variables and each of the previously identified support strategies (*e.g.*, positive words used, supportive words used). For this, we used the well-known frequent item set mining and association rule learning Apriori algorithm. It first generates a set of frequent items that occur together, and then extracts association rules that explain the relationship between those items. We extracted association rules separately for both our ‘more’ and ‘less’ successful outcome clusters; and then calculated the salience for each of the identified rules as the absolute confidence difference between the two clusters (more salient rules are used more frequently by supporters in either the ‘more’ or ‘less’ successful cluster). We derived the 1584 most salient rules, comprised of 8 support strategies and 66 multi-dimensional contexts.

Upon analyzing these we observed, for example, that support messages that use few words of fear and more first person plural pronouns have high salience in more successful messages when a user hasn’t engaged much with the intervention (e.g., no content views, no content shared). This means that writing messages with less words related to fear, and more first person plural pronouns are strongly associated with more successful support messages, and *this effect is particularly salient in situations where clients are disengaged*. This analysis can enable personalization of human support by informing supporters what strategies would be the best to employ in certain frequently occurring client contexts (*e.g.,* disengaged clients).

Personalized Support in the iCBT-Ecosystem

Previous work has leveraged AI to personalize resources within technology ecosystems for mental health management (*e.g.*, [8]). However, the idea of using AI to augment human support in mental health interventions is fairly novel. While advanced data tools that can identify complex patterns are often seen to generate more accurate and objective insights, it is important that we do not take away from, but help foster, the genuine *human connection* that is formed between supporter and client, and that is crucial to their alliance and positive outcomes. Furthermore, we believe it is important to ensure supporters feel that their input and expertise is valued rather than replaced in favor of data science. Hence, we propose designing interventions that seek to augment human support by enabling iCBT supporters personalize their feedback more effectively for each person.

To develop meaningful AI applications for supporters (or clients) often requires access to rich, personal data about the client. Such information can be sensitive due to the stigma that is often attached to mental illness. Thus, what information would be appropriate to share with researchers and developers of AI systems, and for what purposes? How might this extend outside of the immediate context of the iCBT platform? For example, could passively collected personal data (*e.g.*, location and call logs from smartphones) help convey if, when, and how clients apply learned skills in everyday life? Putting clients and supporters at the nexus of AI-supported decision making within such an ecosystem motivates us to examine the points at which it is most fitting for them to make choices, and what options are available. Examples would include deciding what type of program would be most beneficial to the client, what level of treatment intensity or supporter involvement is recommended, what configuration decisions should be made at the outset of treatment (e.g. is a core program augmented with specific content or additional tools?), as well as ongoing recommendations and feedback from supporters throughout the course of treatment.

References

[1] Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention. *Proc. CHI 2020*.

[2] Gavin Doherty, David Coyle, and John Sharry. 2012. Engagement with online mental health interventions: an exploratory clinical study of a treatment for depression. *Proc. CHI 2012*. 1421–1430.

[3] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine 16*(9), 606–613.

[4] Bernd Löwe, et al. 2008. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care 46*(3), 266–274.

[5] Colin Martindale. 1975. English Regressive Imagery Dictionary (RID).

[6] Saif M. Mohammad and Peter D Turney. 2013. NRC emotion lexicon. National Research Council, Canada.

[7] Derek Richards, Ladislav Timulak, Emma O'Brien, Claire Hayes, Noemi Vigano, John Sharry, and G Doherty. 2015. A randomized controlled trial of an internet-delivered treatment: its potential as a low-intensity community intervention for adults with symptoms of depression. *Behaviour Research and Therapy 75*, 20–31.

[8] David C. Mohr, Kathryn Noth Tomasino, Emily G. Lattie, Hannah L. Palac, et al. 2017. IntelliCare: an eclectic, skills-based app suite for the treatment of depression and anxiety. *Journal of medical Internet research* 19.1 (2017): e10.