Understanding and Increasing Empathy in Online Text-Based Peer Support Systems

SARA SYED, Brown University, USA

The United States is facing a mental health crisis due to the inaccessibility of mental health resources. Online text-based peer support systems provide one potential solution to alleviating the crisis and providing effective support for the masses. Empathy is a critical component to the success of mental health and peer support systems. Thus, understanding how (1) empathy can be measured in text, (2) if peer support naturally convey empathy online, and (3) how peer support can increase their empathetic responses, are critical to the success of these digital platforms. This research seeks to examine the application of EPITOME [40], a deep learning model for measuring empathy in text, to Cheeseburger Therapy, an online Cognitive Behavioral Therapy based peer support system where users seek help by texting a trained peer. This research found that EPITOME can successfully be applied to Cheeseburger Therapy to measure empathy but that it performed weaker in 2 out of 3 of the measurements sub-scales. Secondly, it was found that trained peer supporters convey low amounts of empathy, but higher levels of empathy than non-trained peers. Findings also suggest a potential negative correlation between peer support leveraging open ended questions and the experience of the individual receiving support. From these findings, this research concludes by outlining areas for further investigation to understand potential best practices that may encourage increased empathetic responses by peers.

CCS Concepts: • Human-centered computing → Empirical studies in HCI;

Additional Key Words and Phrases: empathy measurement, online peer support, deep learning

1 INTRODUCTION

In the United States over one fifth of adults are affected by mental health illness [2]. Unfortunately, many barriers such as a lack of trained professionals, long wait times, and stigma prevent individuals from having access to proper mental health care services. As a result, in 2020 over 50% of adults with mental health illnesses in the United States admitted to having not received mental health services within the past year [2]. One of the biggest barriers preventing access to health care is the lack of trained mental health professionals. It is estimated that there are only 30.0 psychologists and 15.6 psychiatrists for every 100,000 people in the United States [5], resulting in patients experiencing long wait times1 and high costs.2 On top of accessibility barriers, stigma prevents some from getting care, with nearly 1/3 of Americans

---

1In many areas the average wait time for mental health care is about five to six weeks [? ]. This had led some states, such as California, to pass laws enforcing health insurers to reduce those wait times [1].

2In a 2021 study, one out of four respondents with mental health illnesses disagreed with the statement “I can afford mental health services” [43].
admitted to having worried about others judging them for seeking mental health care. In the United States, 46% of individuals who die by suicide have been diagnosed with a mental health condition [44]. Thus, there is a dire need for better mental health resources in the United States to aid with saving lives and improving people’s well-being. With this current state, it is clear change is necessary to scale up mental health resources and provide access to all. Peer support and internet accessible treatments offer one approach to solving the issue of insufficient mental health resources by removing the need of one on one time from a trained professional, minimizing wait times, lowering resource costs, and reducing stigma.

Cognitive behavior therapy (CBT) is one form of therapy that has been proven to be an effective treatment for anxiety and depression [47] and has led to a general improvement in individuals’ well-being and quality of life [23]. The basic principle of CBT is that thoughts, feelings, and behaviors are all intertwined, and that by simply re-framing one’s thoughts, one can change their feelings and behaviors. Researchers have investigated the use of peer support models of cognitive behavioral therapy as a way to increase accessibility by removing the need for one-on-one time from a health professional [11], and found that peer support systems add therapeutic value [31]. Peer support has been demonstrated to promote treatment engagement, prevent treatment drop out, increase confidence, improve mental health and well-being, and increase one’s ability to deal with issues related to stigma and discrimination [3, 18]. In recent years, the use of internet-based CBT (iCBT) approaches has become more prevalent, with an increasing number of studies suggesting that guided internet-based CBT can be as effective as face-to-face CBT [4], and often provide a more cost effective and accessible form of treatment [15]. As such, effective online peer support systems could dramatically improve the mental health crisis in the United States by solving many issues related to the lack of accessibility to mental health resources.

Irrespective of the support method used or the qualifications of a therapist, empathy in a therapist-patient relationship is a necessity for effective treatment outcomes [21, 49]. More specifically, it is the patient’s perception of empathy levels that are most strongly associated with successful outcomes [17], as there is certainly a need for patients to feel that their therapist has empathy for them. According to Aaron T. Beck [6], in cognitive behavioral therapy, accurate empathy refers to how well the therapist can go into the client’s world and see and experience life the way the client does. As cognitive behavioral therapy relies on the examination of thoughts, feelings, and behaviors and their relationship to a person’s experiences, empathy can help a peer supporter or therapist better understand both the emotional reaction and the meaning of the experiences of a client [48].

Because of the strong relationship between increased empathy from a therapist and efficacy of treatments, there is a need to accurately understand and measure empathy in online peer support platforms where so far little research has been conducted. The ability to quantify levels of empathy is important for allowing therapists to learn and understand the amount of empathy that they convey to clients. This is extremely important as Orlinsky, Grawe, Parks [34] showed that when patients assessed empathy, it was positively correlated with recovery in 34 out of 37 studies. However, when therapists rated their own empathy, significant correlations between empathy and recovery were only reported in 4 out of 15 studies. This suggests that therapists may have difficulty with evaluating their own levels of empathy in comparison to how they are perceived by their clients [9], highlighting the importance of developing a quantitative approach to measuring empathy.

In this research, I used data from Cheeseburger Therapy (cheeseburgertherapy.org) an online cognitive behavioral and text-based peer support platform in order to seek to understand and quantify levels of empathy exhibited by trained peer support in online platforms. In order to do so, this research examines EPITOME [40], one current proposed algorithm for measuring empathy from text to uncover how our current approaches to measuring text-based empathy affect peer
support. Lastly, this research analyzes the results of the application of EPITOME to Cheeseburger to understand how empathy is conveyed in digital text platforms and identifies potential solutions to increase the amount of empathy peer supporters convey. While this research specifically examines the use case of Cheeseburger Therapy, it seeks to understand more generally how empathy can be measured in text-based platforms and trends in online empathy, as therapy treatments and peer support increasingly move to online platforms.

2 RELATED WORKS

Online peer support platforms may be an answer to the problem of the lack of access and supply of mental health resources in the United States. As a transition to providing support in a digital realm could help create accessibility to mental health resources, there is a need to understand empathy in online contexts in order to maintain efficacy with treatments.

2.1 Understanding Empathy Online

Despite large amounts of previous research regarding empathy, little research has been conducted in an attempt to understand empathy in an online context (also referred to as digital empathy), with most work exclusively examining empathy in face-to-face settings. Some research has investigated digital empathy, but little consensus has been reached. For example, some previous research [36, 37, 50] found empathy to be prevalent in online communities, while other work [25, 45] contradicts these findings, observing declining digital empathy in online discussions. In addition, a majority of this work is from the 2000s or earlier, when online settings were vastly different and less widespread than they are today. Because of the lack of consensus and understanding regarding digital empathy, even as recently as 2021, Osler [35] argued from a phenomenology perspective that empathy can even be conveyed online through text, as so much research focuses on an in person setting. Because of the lack of consensus, there is a need to understand how empathy can be measured online, and how these measurements correlate to our current understandings of traditional empathy.

2.2 Current Approaches to Measuring Empathy

There is no single standardized method for measuring empathy. Empathy has historically been measured through the analysis of a multitude of features such as in body language and tone, or in face-to-face settings. Few approaches have sought to measure empathy exclusively from text or using an automated method.

Empathy can often be perceived by the examination of physical reactions like facial features [16] or body language, which are often not present in online settings or text-based platforms like Cheeseburger Therapy. Other research provides support for the idea that the process involved in emotional empathy is related to automatic, spontaneous somatic reactions [42]. As such, in text based online platforms there poses a need to be able to quantify empathy without the use of these nonverbal reactions.

One of the most commonly used scales for measuring empathy is the Interpersonal Reactivity Index (IRI) [12] which requires respondents to reply on a 5-point scale to the extent to which they agree with certain statements. The IRI measures 4 subsets of empathy: (1) Perspective Taking – one’s inclinations to adopt the psychological point of view of others. (2) Fantasy – one’s abilities to imaginatively identify with the feelings and actions of fictitious characters in books, movies, and plays. (3) Empathic Concern – one’s tendencies to feel for others. And (4) Personal Distress – one’s inclination to feel discomfort from others’ negative experiences. While the IRI is not the only scale used for empathy measurement, it is representative of the fact that in general the most common way that empathy is assessed is via
self-reported trait measures. However, as previously mentioned, therapists may have trouble self-assessing empathy, so this method may not provide the most accurate results. Instead, an automated approach to measuring empathy is needed to prevent reliance on self-assessment and to create scale-able peer support. Many existing empathy measurements also assess an individual’s general ability to empathize (known as dispositional empathy) and not their empathy level within specific contexts (known as situational empathy) such as each individual therapy session. Thus these scales do not provide applicable approaches for measuring empathy with the goal of improving the efficacy of online support sessions.

To the best of my understandings, little previous research has sought to quantify empathy from text [20, 24, 51]. However, most that do, do not provide publicly available datasets or models, making it difficult to study their capabilities in applied settings. Research by Xiao, Bo et al. [51] measures situational empathy from transcripts of therapy sessions and Khanpour, et al. [24] seeks to measure empathy directly from text of comments left on discussion threads in the Lung Cancer discussion board in a Cancer Survivor’ Network (CSN). However, in both empathy measurement approaches only text from the person seeking to provide help are utilized. As such, these models’ definitions of empathy fail to encapsulate an understanding of if the supporter “can go into a clients world” by disregarding the context of the information the person seeking help has shared. Some methods of empathy analysis [20, 51] relied on transcribing therapy sessions to text, as opposed to settings in which participants communicating directly via text. For these reasons, these methods may not be the most applicable to online text-only based platforms.

Thus for scalable, online, text-based peer support platforms, an empathy measurement needs to (1) provide an automated approach for scalability (2) incorporate text as the only parameter (3) examine situational empathy (4) incorporate both the peer providing support and the client’s perspectives.

In this paper I seek to examine EPITOME [40], a deep learning model for measuring empathy in the applied setting of trained peer support. This model provides an automated approach to measure situational empathy exclusively from text and thus has the potential to be seamlessly integrated into online platforms. Its parameters require text from a person seeking advice and a person giving advice, and thus has the potential to be applied to a trained- peer support platform. This research will provide a fuller understanding of how its empathy scores can be translated to meaningful results in situations of continuous back and forth text conversation and explore approaches to increasing empathetic responses by peer support.

3 METHODS

In order to understand our current abilities to algorithmically measure empathy from text to ultimately create more empathic online peer support platforms, I sought to measure the efficacy of using EPITOME to analyze Cheeseburger Therapy, one such online text based peer support platform.

3.1 The Model: EPITOME

To measure empathy from the Cheeseburger Therapy platform, I apply the EPITOME model. EPITOME is a novel recently released algorithm and the only, to my knowledge, deep learning approach that fully automates measuring empathy from text while considering the text of both an individual seeking advice and the text of a peer support who is replying to the seeker. The model takes as inputs sets of two texts (one text from a user seeking help, referred to as the seeker text and a second text from a peer support, referred to as the response text) and measures the empathy level in a peer supporter’s response to a peer’s initial text. The model was trained using EPITOME’s published dataset of Reddit (reddit.com) posts and replies that were taken from threads of 55 mental health focused sub-Reddit groups [41].
The model’s outputted empathy score relies on a framework of measuring 3 communication methods that result in 3 sub-categories of empathetic responses: Emotional Reactions, Interpretations, and Explorations. Within these three categories, the model outputs a score (0) no communication, (1) weak communication, or (2) strong communication, relaying the graded extent to which peer supporters conveyed the communication method in their reply. The seeker’s text is used as context for the peer’s response.

**Emotional Reactions.** The score for emotional reaction is meant to explain the extent to which peer supporters conveyed kindness, warmth, or other emotions in their reply. Other previous models [8] for measuring empathy have also focused on quantifying warm, tender, and compassionate emotional responses, suggesting the importance of this metric for empathetic measurements. No communication (0) signifies that the peer supporter failed to convey such a warm emotional response. A weak communication score (1) indicates that peers supporters lacked specificity regarding what emotions they felt in reply to the user’s post. A strong communication score (2) was one in which a peer supporter stated a specific emotional response to the poster.

**Interpretations.** The measurement of interpretation is meant to quantify a peer supporter’s level of communication and understanding of the other user’s situation or feelings. As empathy relies on a person’s ability to understand other’s perspectives, nearly all previous empathy scales seek to measure a person’s ability to understand or attempt to understand others. Furthermore, Hirsch [22] underscores that conveying this understanding in responses allows the receiver to perceive this response as empathetic. A no communication score (0) indicates a peer supporter failed to convey an understanding of the poster’s feelings or situation. A weak communication score (1) indicates the peer referenced that they had an understanding, whereas a strong communication interpretation (2) specifies clearly what the other’s feelings are.

**Explorations.** The exploration sub-category measures whether the peer supporter took an active role in trying to explore and further their understanding of the initial poster’s emotional reactions and experiences. A key investment in exploration is inline with previous empathy scales. For example, on the MITI scale [32] a clinician who “makes active and repeated efforts to understand the client’s point of view” would score higher for empathy than clinicians who do not actively try to understand the client’s perspective. A no communication score (0) indicates that the peer did not attempt to further their understanding of the thinker’s feelings or situation. Weak communication (1) would be a response in which the peer generically asks the poster for more information, whereas strong communication of exploration (2) is one in which the peer supporter seeks to further their understanding of a specific point of the initial poster’s situation or feelings, and clearer mentions this specificity.

As such, a minimum score would constitute receiving a 0 in all three sections, whereas a maximum empathy response would constitute a 2 in all sections (for a total of 6).

<table>
<thead>
<tr>
<th>Sub-category</th>
<th>no (0)</th>
<th>weak (1)</th>
<th>strong (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional Reactions</td>
<td>“Too bad.”</td>
<td>“Everything will be fine.”</td>
<td>“I am so sorry for your loss.”</td>
</tr>
<tr>
<td>Interpretations</td>
<td>“Too bad.”</td>
<td>“I can understand your situation.”</td>
<td>“I can only imagine how anxious you feel.”</td>
</tr>
<tr>
<td>Explorations</td>
<td>“Too bad.”</td>
<td>“Why is that?”</td>
<td>“Why do you think this made you feel so hurt?”</td>
</tr>
</tbody>
</table>

Table 1. Paraphrased example responses based on our private dataset and their corresponding scores
Lastly, EPITOME distinguishes itself from many other deep learning prediction models by providing a level of explainability by outputting "rationales" for each of the three sub-scales. The rationales are a sub-sequence of words from the response post that formed the explanation behind the outputted sub-scale score. For each pair of a seeker text and a response text, \((S_i, R_i)\), the corresponding rationale is represented as a mask, \(m_i = (m_{i1}, ..., m_{in})\) where \(m_{ij} \in \{0, 1\}\) and 1 represents a word that was part of the rationale for response post \(R_i = r_1, ..., r_n\).

Empathy Score

Rationale

\[ a_i(e(R_i), e(S_i)) \]

Attention

\[ e(S_i) \]

\[ e(R_i) \]

\[ e(R_i[CLS]) \]

\[ e(R_i[SEP]) \]

\[ e(S_i[CLS]) \]

\[ e(S_i[SEP]) \]

S-encoder

R-encoder

Three identical models are trained and used for each sub-scale.

![High-level Architecture of EPITOME](image)

3.1.1 EPITOME Architecture. The following is a high-level overview of the model’s architecture. To see the model’s code go to: https://github.com/behavioral-data/Empathy-Mental-Health. This architecture is illustrated in figure 2.

EPITOME is a multi-task bi-encoder model based on RoBERTa [27], a language model created by Facebook AI that uses an improved pre-training procedure to BERT [14], Google’s state-of-the-art language model. Three separate but identical models are trained in parallel to produce each of the three sub-category scores and their respective rationales. Two independently pre-trained RoBERTa\textsubscript{BASE} transformer encoders are used to separately encode the seeker text, \(e(S_i)\), and response text, \(e(R_i)\), into word embeddings. Special start [CLS] and end [SEP] tokens that were adapted from BERT are used in this encoding process. A single-head attention layer is used over the two encodings to generate a representation of the response text that is seeker text aware. The attention is computed as follows:

\[ a_i(e(R_i), e(S_i)) = \text{softmax}\left(\frac{e(R_i)e(S_i)}{\sqrt{d}}\right)e(S_i), \text{ where } d \text{ is the hidden size } = 768 \]

The final representation of the [CLS] token from the response text is passed through a linear layer to generate the empathy score (0, 1, or 2). Final representations of the individual tokens, \(r_1, ..., r_n\), of the response post, \(R_i\), are passed through a linear layer to generate the boolean predictions of the rationale mask, \(m_i = (m_{i1}, ..., m_{in})\). The sub-sequence of words, \(x_i\) from the rationale can then be computed by \(x_i = m_i \odot R_i\).

\[3]\text{In the original publication of EPITOME, domain-adaptive pre-training was further performed to provide a conversational and mental health context. However, this dataset was not publicly available, so this additional step was not conducted in this research.}
3.2 Dataset Description

3.2.1 The Cheeseburger Therapy Platform. The dataset was sourced from Cheeseburger Therapy (cheeseburgertherapy.org). Cheeseburger Therapy is an online cognitive behavior therapy based peer-support website created by Michael Toomim and Morgan Dixon at the University of Washington and Invisible College. The Cheeseburger Therapy model aims to help users see their trouble in a new light by learning to re-frame their thoughts. The Cheeseburger model allows anyone with an internet accessible device who is seeking support to sign up for a one-hour session where they communicate with a Cheeseburger Therapy trained peer to help them identify common thought patterns that may be troubling to them. Individuals who use the site to seek support are referred to as the thinker, and the Cheeseburger trained peer support are referred to as the helper. The Cheeseburger framework follows the cognitive behavioral assumption that individuals can work through a troubling issue by learning to re-frame their thoughts. Cheeseburger’s mission is to help scale mental health support which they seek to accomplish through online trained peer support. As such, anyone can sign up to become a helper, which just requires completing Cheeseburger’s training process. The process constitutes of completing their manual and going through practice sessions with another peer in training, usually requiring at minimum 20-30 hours.

During a Cheeseburger Therapy session, peer helpers and thinkers communicate exclusively through text. Peer helpers are trained to follow a systematic process to effectively conduct sessions. The process usually starts with the helper asking the thinker to share what issue has been troubling them recently. The helper is then trained to listen and inquire as they ask the thinker to share about the event, thoughts, feelings, and behaviors they have surrounding

![Helper’s Viewpoint](image-url)
this trouble. The model relies on the session focusing on this one 'trouble', which Cheeseburger defines as "any threat to something we value". Throughout the session, helpers and thinkers may take notes in the "Notes section" on the right-hand side of the screen as the participant discloses information. Figure 3 displays an example of a helper's viewpoint during a Cheeseburger session.

The viewpoint for the thinker is near identical, however, it does not display the prompts listed at the bottom of the screen below the "Send" button. Throughout sessions, helpers are given these prompts to help guide the conversation and assist them with following the Cheeseburger structure. As they complete prompts, they are able to check them off and work through the following assistive workflow:

- Ask the user what is troubling to you?
- Restate what the user finds troubling.
- Explain that we will be taking notes.
- Ask what is most troubling about this.
- Ask what event occurred before you thought <Thought>? Restate Event.
- Ask what event happened immediately before you felt <Feelings> and did <Behaviors>? Restate Event.
- Ask what thoughts do you have when <Event>? Restate the thought.
- Ask how did you feel when you thought <Thought>? Restate the feeling.
- What do you feel before you do <Behavior>? Restate the feeling.
- Ask what actions do you take or avoid when you feel <Feelings>? Restate the behavior.
- Ask how does ____ relate to ____?
- Ask if there is anything missing from the summary.
- Choose the one Thought to focus on.
- Label the thought.
- Restate the Feelings and Behaviors and how they connect to the Labelled thought.
- Ask what is a new thought that you could have in response to the Event?
- Ask what feelings come up for you when you think <New Thought>?
- Ask how are you feeling now, compared to when we started?
- Say when <Event> occurs again in the future, remember to practice the thought <New Thought>

Following this structure, the first portion of the session is dedicated to the task of "listening". In line with CBT practices, helpers are trained to use restatements during this listening portion to help create clarity surrounding the thinker's situation and ensure that the helper has an accurate understanding of the thinker's emotions and situation. Once the helper feels they have a strong understanding, they then work with the thinker to identify a common thought pattern in the thinker's thoughts surrounding the trouble. The goal of this task is to allow the thinker to understand how this common thought pattern may be causing them to see their troubling situation in a one-dimensional manner. The goal for the helper and thinker is to create a new thought for the thinker that would allow the thinker to see their situation in a new way, and with time ultimately affect their feelings and behaviors regarding the situation.

3.2.2 Collection, Cleaning, and Statistics. I collected data from 172 Cheeseburger Therapy sessions that took place from mid-November 2021 through May 2022. Of the 172 sessions, some were publicly accessible and collected directly from the Cheeseburger homepage but the majority were private sessions stored directly on the Cheeseburger server. I received access to this data by writing JavaScript queries to obtain the information needed without having access to any additional identifiable information. Of these 172 sessions, 56 were removed as they were either duplicate sessions
or incomplete sessions. In this case, I considered incomplete sessions to be sessions in which the thinker and the helper did not complete the listening portion of identifying thoughts, feelings, and behaviors. However, thinkers and helpers did not have to complete the task of creating a new thought. Of the remaining 116 sessions in the dataset, 16 sessions were "client sessions", sessions in which the helper was an individual who had completed the Cheeseburger training program and the thinker could be any user. The remaining 100 sessions in the dataset were "buddy sessions"; sessions in which two peers who are in training (have not yet completed the Cheeseburger training program) conduct a session together in which one acts as the thinker and the other the helper. In these sessions, peers practice the Cheeseburger methodology as they would as if they were conducting a session with a real user. However, if they need to communicate with one another for any reason outside of the normal session context, for example for assistance, they can do so through the "back-channel". To communicate via the back-channel, peers were taught to simply send their message texts either within [] or (). As back-channel conversation falls outside of the normal session structure, all buddy session data was parsed and the back-channels were removed. From all sessions, I collected data that included the text messages exchange between the helpers and thinkers, the notes that were taken during the session, and coded anonymous IDs that linked buddy session helpers together. After completion of buddy sessions, thinkers and helpers could also complete a form, providing feedback on the session, how successful they found it, and advice for their buddy. Of the 100 buddy sessions, 72 contained meaningful feedback. Figure 4 shows the breakdown of the sessions that remained in the dataset.

![Fig. 4. Buddy vs Client Session Breakdown](image)

On average approximately 123 messages were exchanged per session, with about 61 messages being sent from the helper and 62 from the thinker. Figure 8 outlines the distribution of messages sent by users. To convert Cheeseburger session data into inputs compatible with the EPITOME model, I wrote and used a script to convert the conversations into pairs of thinker’s posts and helper’s replies. For example, referencing figure 6, messages 1) and 2) would be concatenated together as the thinker’s text and 3) and 4) would be concatenated together as the helper’s text. An empathy score would be generated given this information. Message 5) as the thinker’s text and 6) as the helper’s would be another set of inputs, and another empathy score would be generated given this information. Accordingly, an empathy score was calculated for each of the helper’s concatenated replies in a session. Each session had on average 33 data points, pairs of thinker messages and helper messages, for which an empathy score was calculated. When creating these data points and concatenating messages, messages’ texts were left as was and was not cleaned for typos, as the EPITOME model was also trained using a dataset of uncleaned text. It is also a fair assumption that since individuals communicate
exclusively through text, typos in their messages or stylistic decisions may have an impact on the amount of empathy that is relayed.

3.2.3 Ethical Considerations. It is important to note the ethical considerations that were made when collecting this data due to its sensitive nature, as session conversations could contain information or discussion related to mental health or other personal information. This work does not make any mental health or treatment related recommendations and helpers in the previously conducted sessions were instructed in training to never provide any sort of diagnosis, advice, or recommendations to the users. To protect user’s privacy, I submitted an application through Brown University’s IRB board and received approval as well as exempt determination for this project. In order to mitigate risks, no directly personally identifiable information (PII) was collected (such as names, emails) other than information that may have been contained directly in the messages’ text, and no PII was directly linked to session messages or other collected data. Queries were written in such a way that for each buddy sessions, an anonymous randomized ID was created to link

distribution of the total number of messages sent per session

for example, 'I am happy!!!' versus 'I am happy'.

Fig. 5. Breakdown of messages sent per session
buddy sessions completed by the same helper together to track helper’s progress as they completed more practice buddy session, while eliminating the need to use identifiable information of that buddy. Data was also further anonymized in a way that given session data, an attacker could not re-identify the Cheeseburger session’s URL information to access data through their server.

4 ANALYSIS: UNDERSTANDING EMPATHY

4.1 Error Analysis

To determine the accuracy of produced empathy scores, I first quantitatively analyzed the model’s results by taking a random sample of 100 data points. For each data point in the sample, I hand rated them on the emotional reaction, interpretations, and explorations sub-scales. The emotional reactions and explorations sub-scales were found to have an accuracy of 78% and 73% respectively. The accuracy of the interpretations sub-scale was lower, at 65%. While the performance of the model on the Cheeseburger dataset is high, these results are lower than accuracy published by EPITOME on their published Reddit dataset (the dataset used for training the model).

<table>
<thead>
<tr>
<th></th>
<th>Reddit Accuracy</th>
<th>Cheeseburger Therapy Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional Reactions</td>
<td>79.43%</td>
<td>78.00%</td>
</tr>
<tr>
<td>Interpretations</td>
<td>84.04%</td>
<td>65.00%</td>
</tr>
<tr>
<td>Explorations</td>
<td>92.61%</td>
<td>73.00%</td>
</tr>
</tbody>
</table>

Table 2. Accuracy of Results.
From a qualitative analysis of the results, and in line with the findings of the model’s original publication, I observed that the model often over-scored for exploration responses that contained questions but were not aligned with the intent of a specific exploration of the thinker’s feelings or situation. For example, the response “helpful?” received an exploration score of two, despite it failing to explicitly label the thinker’s feelings or the situation that the helper is hoping to inquire more about. Additionally, strong interpretation reactions from the helper were often mislabeled, and instead given a 0 on the interpretation scale. In these cases, an additional point was often rewarded to the empathy sub-scale, when in fact the helper was expressing an understanding of the thinker’s situation or feeling and not an emotional reaction to the thinker’s situations of feelings. However, on occasion, short replies from the helper such as “haha” or “absolutely” were incorrectly judged and given an interpretation score of 2. Interpretation and Exploration sub-categories produced binary results, with the model only ever rewarding 0 or 2 and never grading any of the helper’s responses with a 1 (weak) interpretation or exploration. Factors like the thinker’s message length did not appear to play a major role in the quality of the outputted score.
4.2 Score Distribution

In general, the helper’s responses scored low on the empathy scales, with the average total empathy (the sum of the emotional reactions, interpretations, and explorations scores) for a data point being 1.69 out of 6. This suggests that there is potential for great improvement to increase the amount of empathy Cheeseburger peer supporters convey. While the results suggest low empathy across sessions, this total empathy score is significantly higher than the one reported in the EPITOME original study, where they observed an average total empathy score of 1.09 out of 6. The distribution of results shown in table 6 suggest that Cheeseburger helpers scored a higher percentage of weak and strong emotional reactions scores than the peers that were studied in the original paper’s Reddit dataset and a higher percentage of strong explorations scores. A key difference between helpers from Cheeseburger Therapy versus Reddit is that peer supporters have gone or are going through a training process. This suggests that trained peers have an increased ability to convey empathy through emotional reactions and explorations and the reasoning for these results will be explored further in later sections.

<table>
<thead>
<tr>
<th></th>
<th>Cheeseburger Therapy</th>
<th>Reddit (training data)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>weak</td>
</tr>
<tr>
<td>Emotional Reactions</td>
<td>1636</td>
<td>1897</td>
</tr>
<tr>
<td>Interpretations</td>
<td>3105</td>
<td>0</td>
</tr>
<tr>
<td>Explorations</td>
<td>2630</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3. Number of data points of each score.

<table>
<thead>
<tr>
<th></th>
<th>Cheeseburger Therapy</th>
<th>Reddit (testing data)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>weak</td>
</tr>
<tr>
<td>Emotional Reactions</td>
<td>42.4%</td>
<td>49.1%</td>
</tr>
<tr>
<td>Interpretations</td>
<td>80.4%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Explorations</td>
<td>68.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 4. Distribution of results

This shows optimism and suggest that (i) through some initial training peer supporters do learn higher empathy or (ii) trained peer supporters may have more incentive to be empathetic in their responses than the average internet user who is responding to mental health related forums. Incentives to be more empathetic in responses could be due to factors such as 1) Cheeseburger has peer supporter’s account information and thus may have user’s identifiable information 2) Peer supporters have committed some initial time and effort to training 3) Individuals may be receiving feedback on their session. These results of higher empathy are particularly optimistic, as the Reddit data contains pairs of inputs in which the thinker was specifically seeking out a mental health support forum. In the Cheeseburger Therapy data, however, it can be argued that there is not a particular need for empathy in all replies from the thinker, as not all data inputs were ones in which the thinker was specifically sharing or seeking advice. Many data points included in sessions, for example, contained conversation where the helper and thinker were communicating regarding additional times to speak or WIFI issues.
### 4.3 Average Empathy

For each session, an empathy score was computed for each “back and forth” speaking turn between the thinker and the helper, resulting in multiple scores per session. However, the analysis aims to use the collection of these scores to make claims about the overall empathy conveyed in a session. To understand how empathy scores that were calculated for many individual segments of the session can be translated to provide meaningful results about the efficacy of the entire session, I leveraged methods of previous research by Xiao, Bo, et al. [51]. This research was conducted using data of transcriptions of therapy sessions that used Motivational Interviewing (a counseling approach aimed to assist clients in finding the motivation to make behavior changes) for substance use. The datasets included measurements of both empathy on an utterance level (therapist’s speaking turns) and empathy on a session level. Utterance level data points were coded as either (EMPATHY) or (NO EMPATHY) using the MISC protocol [29] and session level data was coded using the MITI coding system [32] and ranked on a Likert scale from 1-7. The research found a positive correlation between averaging utterance level empathy scores (\( \frac{\text{number of (EMPATHY) utterances in a session}}{\text{total number of utterances in a session}} \)) and session wide empathy. As such, the analysis of this research centers around averaging the empathy scores for each session to understand the impact of session wide empathy.

For each session \( s_i \in S \), the averaged empathy scores for \( s_i \) were defined as:

\[
\begin{align*}
\text{averaged emotional reactions} &= \frac{\sum_{x=1}^{N} \text{emotional reactions score of } d_x}{N} \\
\text{averaged interpretations} &= \frac{\sum_{x=1}^{N} \text{interpretations score of } d_x}{N} \\
\text{averaged explorations} &= \frac{\sum_{x=1}^{N} \text{explorations score of } d_x}{N} \\
\text{averaged total empathy} &= \text{averaged emotional reactions} + \text{averaged interpretations} + \text{averaged explorations}
\end{align*}
\]

The averaged total empathy score per session stayed within a range of approximately 0.85-2.7 (with the maximum possible score being 6). The averaged interpretation score across sessions was quite low, at approximately 0.39 versus 0.66 and 0.64 for averaged emotional reactions and averaged explorations scores respectively.

#### 4.3.1 Helpers do not exhibit consistent levels of empathy across sessions

First, to understand how practice related to helpers’ conveyed averaged empathy scores, when initially collecting data, anonymous helper IDs were added to the buddy session data in order to link sessions together that were conducted by the same helper without revealing the helper’s identity. Subsequently, this analysis was conducted using the data of the 100 buddy sessions, which were conducted by 12 unique helpers.5 Helpers included in this dataset may have had lots of previous experience (helpers who have completed the training program are still able to participate in practice buddy sessions) or they may have been completing their first buddy session. On average helpers completed approximately eight sessions (with the median being 9.5) during the data collection period of November 2021 to May 2022. Figure 9 provides the breakdown of the number of buddy sessions that were completed per helper.

---

5While this may seem like a small sample size, using data from 100 sessions is well above the necessary requirements, where previous research [46] indicates that a sufficient sample size for data composed of interviews need only contain 30 sessions in order to draw meaningful conclusions regarding patterns and properties. The limited number of helpers is also comparable to other Human-Computer interaction (HCI) research. Caine [10] analyzed the sample sizes of research published in the CHI2014, a leading international conference in HCI, and found that the mode sample size was in fact 12.
Fig. 8. Distribution of averaged empathy scores

No trend was observed between averaged empathy scores and the number of sessions a helper had completed. All helpers’ empathy scores varied greatly from session to session, with all helpers who completed over three sessions, having a difference of at least 0.79 from their maximum averaged total empathy score in a session to their minimum averaged total empathy score. As such, no helpers consistently scored at a high empathy level or a low empathy level. Figure 10 shows the changes in three helpers’ empathy scores across buddy sessions they completed. The figure illustrates how all three helpers greatly fluctuated in the amount of empathy they conveyed, regardless of the empathy level they started at. In fact, while scores varied greatly from session to session for a helper, averaged scores across all sessions were similar for nearly all helpers. Of the helpers that completed 3 or more sessions, 9 out of 10 helpers scored between 1.69-1.99 when averaging all their total empathy scores across all sessions they completed. This suggests that helpers do not exhibit consistent levels of empathy across sessions and that helpers do not self-learn empathy over time.

This observation is in line with the finding from EPITOME’s original publication [40] that peer supporters do not self-learn empathy over time. As EPITOME’s research examined responses from untrained peers, this suggests that
Fig. 9. Distribution of the number of buddy sessions a helper completed

Fig. 10. The averaged total empathy scores of three helpers
there is no strong difference between the ability for trained and non-trained peers to self learn empathy. However, when comparing the averaged total empathy scores from the Cheesburger data to that of EPITOME’s Reddit dataset, higher levels of empathy in peer supporters’ responses were observed. Accordingly, this increased empathy is likely not due to a difference in peer supporters’ ability to self-learn empathy with time, but rather due to trained peer supporter’s initial training or a difference in their underlying motivations that provide them a higher baseline of empathy.

4.3.2 Averaged explorations may be negatively correlated with thinkers’ experiences. To examine the role of the averaged empathy scores of a session on the thinker’s experience, the averaged empathy scores were compared to reviews which were collected from 72 buddy sessions. In each of these 72 buddy sessions, after the session was complete, thinkers filled out a feedback form detailing their experience in the buddy session. Figure 11 shows the UI of the thinker’s form. One component of the form was a range slider in which thinkers rated the overall session from “suffering” to “enlightenment”. Evaluations from this component were converted to a value on a quantitative scale from -1 (“suffering”) to 1 (“enlightenment”).

Data indicated that the averaged emotional reactions, and averaged interpretations scores did not have a high correlation with the thinker’s feedback score. However, the averaged exploration score signifies a weak model of a negative correlation between explorations and the thinker’s experience. See figure 12 for the results of the analysis. This was a surprising result, as it may be more expected for the empathy score and its sub-scores to be positively correlated with the thinker’s feedback on their experience. This result may imply that when asked too many questions regarding their experiences or feelings, thinkers either feel that (i) the helper has a lack of understanding or (ii) they are unhappy with the amount they are being inquired about. While this result is in some ways surprising, it is in line with the results from EPITOME’s research in which approximately 28% of replies that received an exploration score of 0 were liked by the person seeking helper. Whereas only approximately 15% of posts that were scored as 1 or 2 on the exploration sub-scale were liked by the person seeking help, indicating that high exploration scores were less often associated with liked posts. While in many ways, all facets of the empathy score may be expected to be positively correlated with the
thinker’s experience, previous research [38] has also indicated that increased use of open-ended questions does not lead to improved evaluation of counseling sessions and may not be appreciated by all evaluators.

To understand the observed weakly inverse relationship between the averaged explorations score and the thinker’s overall feedback score, I next analyzed the conjecture (i) overuse of explorations may lead the thinker to feel they were not understood by the helper. To test this theory, I compared the averaged explorations score to other feedback by the thinker, specifically the radio-box feedback: “I felt heard and understood”. Figure 13 illustrates the comparison of the distribution of averaged explorations for sessions in which the thinker did and did not feel “heard and understood”. A point bi-serial correlation\(^6\) was used to test for statistical significance between the averaged explorations score to the binary value for whether the thinker “felt heard and understood” and a level of 0.05 was used for all tests of significance. The results indicate that the mean of the averaged explorations scores for sessions in which the thinker indicated that they did not feel “heard and understood” (approximately 0.95) was statistically significant and higher than the mean of averaged explorations for sessions in which the thinker indicated that they did feel “heard and understood” (approximately 0.66), as the p-value for a hypothesis assuming no relationship (approximately 0.0019) was well below the significance threshold of 0.05. This confirms a negative correlation between increased averaged explorations and feeling understood. There were, however, some notable outliers in this data. For example there was a session in which

\(^6\)Equivalent to a two sample t-test with equal variances.

Fig. 12. Averaged empathy scores vs thinker’s overall score

(a) Averaged Total Empathy vs Thinker’s Overall Score

(b) Averaged Emotional Reactions vs Thinker’s Overall Score

(c) Averaged Interpretations vs Thinker’s Overall Score

(d) Averaged Explorations vs Thinker’s Overall Score
the thinker indicated they did feel “heard and understood” and the averaged explorations score was 1.25, the second highest explorations score in this dataset.

This finding may have larger implications as it suggests that when leveraging too many explorations, helpers risk causing the thinker to feel a lack of understanding regarding their situation. While it is important for helpers to use questions to explore their understanding of the thinker’s situation and perspective, it may be necessary to limit the quantity of questions in order to ensure the thinker feels the helper understands them.

4.4 Listening Empathy

In the previous sections of this analysis, I utilized the averaged empathy scores of the entire session. However, it is unclear whether conveying empathy throughout all parts of the conversation is necessary, or if there are certain points of the CBT method that benefit more from the use of empathy. As such, I next seek to examine the effect of the level of empathy in the “listening” portion of the Cheeseburger model within conversation.

4.4.1 Why Listening? One intention of extracting the “listening” section was to remove noisy data\(^7\) from the calculated average. I specifically intended to study the listening portion of the sessions because it is a portion of the Cheeseburger session in which exhibiting empathy may be more important for the session’s success, as this is the section where thinkers do most of their sharing regarding their trouble. Exhibiting empathy in the listening section may encourage thinkers to feel more comfortable sharing, and this is crucial for the success of the session, as it is necessary for cognitive behavioral therapists to accurately understand the patient’s thoughts, feelings, and behaviors to enable them to uncover

---

\(^7\)Data points where an empathetic response from the helper was unnecessary as users were not currently engaged in conversation that was part of the Cheeseburger method. For example, when discussing alternative times to speak or issues with their WIFI connectivity.
alternative perspectives and new ways of thinking [30]. As such, analyzing the listening section gives a measurement of if helpers are empathetic in parts of conversation where it may be advantageous, rather than on average.

As taught in the Cheeseburger training and in line with CBT practices, helpers should leverage restatements during this listening portion to help create clarity surrounding the thinker’s situation and ensure that the helper has an accurate understanding of the helper’s emotions and situation. Previous research has shown [19] that therapists who deployed more restatements and asked for clarifications were rated higher on therapeutic empathy by their patients. Given this, I expect 1) on average higher levels of empathy in helpers’ responses in the listening section than the entire session 2) helpers do not self-learn empathy in listening over time and 3) empathy within listing is more strongly correlated than empathy in the session overall with the thinker’s experience.

4.4.2 Cleaning for Listening. To extract the “listening” portions of the conversations, I qualitatively analyzed the buddy sessions and coded each data point within the sessions as either (LISTENING) or (NOT LISTENING).

The goal of leveraging restatements during this listening portion to help create clarity surrounding the thinker’s situation and ensure that the helper has an accurate understanding of the helper’s emotions and situation is reaffirmed to helpers through this workflow as they are encouraged to use restatements through the prompts. Because helpers followed a framework for the session, which was guided by the assistive workflow, (LISTENING) and (NOT LISTENING) parts were easily distinguishable. Listening portions of the conversation usually began after the helper inquired about what was troubling and was coded as starting once the user began to share about any troubling event, thoughts, feelings, or behaviors. Listening portions of the conversation often began after some sort of introduction between the helper and thinker. Listening portions of the session were completed when helpers asked if there was anything missing from the summary and received confirmation that the thinker had nothing else to add, or when the helper and thinker had moved on to labeling the thinker’s troubling thought.

4.4.3 Listening Dataset. Filtering out data points that were coded as (NOT LISTENING) produced the listening dataset. The listening dataset was composed of 1973 data points indicating that on average listening composed approximately 51.1% of the conversation.

4.4.4 Listening Distribution. As hypothesized, selecting for just listening portions of the conversation resulted in higher empathy scores across all sub-categories. In fact, in 83 out of 100 buddy sessions, helpers averaged empathy scores were higher in the listening section than in the session as a whole. This is in line with expectations as peer supporters are strongly encouraged throughout the original training manual and through the assistive workflow to use restatements and explorations. Subsequently, the number of strong interpretation scores increased from the original 19.6% of the entire conversation to composing 21.4% of the listening portion. Strong exploration scores also increased from accounting for 31.9% of data points of the entire conversation to 36.4%. The number of strong emotional reactions increased from 8.5% of all data points to 10.3% of the listening data points.

<table>
<thead>
<tr>
<th></th>
<th>Listening in Cheeseburger Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Emotional Reactions</td>
<td>797 (40.4%)</td>
</tr>
<tr>
<td>Interpretations</td>
<td>1531 (77.6%)</td>
</tr>
<tr>
<td>Explorations</td>
<td>1255 (63.6%)</td>
</tr>
</tbody>
</table>

Table 5. Number of data points of each score.
The breakdown of empathy scores by sub-section, as shown in table 5, has interesting implications. While there were increases in strong scores for all sub-categories, the percentage of strong emotional reactions scores still remained quite low (10.3%). The training manual and the workflow of the listening section specifically encourages the combined use of restatements, which may be expected to correspond to the interpretations score, and open ended follow up questions, which may be expected to correspond to the explorations score. However, few emotional reactions scores were rated as strong, and in the listening workflow helpers are not specifically encouraged to respond with their own feelings. This suggests, that the use of targeted statements in the assistive workflow that encourage emotional reactions may increase the empathy in the helper’s responses.

4.4.5 Helpers similarly did not self learn empathy in the listening section. Findings from analyzing averaged listening empathy scores with the helper’s experience met the hypothesis and no trend was observed in the amount of empathy helpers conveyed in the listening portion of the session with respect to having completed more buddy sessions. In referencing figure 14, the observed patterns of fluctuation in the averaged total listening empathy look similar for these same three helpers when comparing them to the overall averaged total empathy (figure 10). Again, these results are in line with previous research and the expectation that peer supporters will not self-learn empathy in listening over time, but that they do exhibit proportionally higher levels of empathy in the listening section than the session as whole. As such, more training directly in empathy skills are needed to increase empathy, as this will not be self-learnt.

4.4.6 Empathy conveyed in listening did not have a higher correlation to the thinker’s evaluation. Similar results to the averaged overall session scores were found when comparing the average of the listening empathy scores to the thinker’s feedback. This result is shown in figure 15, where observations again indicate a loosely negative correlation between
averaged listening explorations and the thinker’s overall score of the session. Findings from figure 16 indicate a negative correlation between explorations scores and the thinker indicating that they felt “heard and understood”. The mean listening explorations score for sessions in which the thinker did feel “heard and understood” was approximately 0.77 whereas the mean empathy of all sessions in which the thinker did not feel “heard and understood” was approximately 1.08. However, contradicting the initial hypothesis, results from analyzing the listening section indicate that there is not a higher correlation between the averaged listening empathy and the thinker’s feedback, than when average of the entire session was used. Thus using the average of the entire session is an appropriate measure as there is no particular need for higher empathy specifically in the listening portion of the conversation.

5 LIMITATIONS & FUTURE WORK: INCREASING EMPATHY

This research proved that the EPITOME model for empathy measurements can successfully be applied to Cheeseburger Therapy to draw meaningful understandings regarding empathy conveyed online through text and suggests the potential for empathy measurements to be applied to more online peer support platforms or other digital text-based platforms. Through a quantitative and qualitative error analysis, it was found that EPITOME provided particularly high accuracy results (>70%) specifically for the emotional reactions and explorations sub-scales. Future work should investigate the potential causes of binary results that were predicted from the interpretations and explorations scores, which may have been a limiting factor in the results.
Results indicated that trained peer support conveyed higher empathy than that of previously reported untrained peer support, but that averaged levels of total empathy expressed remained low (1.69 out of 6). The highest averaged total empathy of any session was 2.7, which suggests that perhaps aiming for sessions averaged responses to score a 6 (strong emotional reactions, strong interpretations, and strong explorations) may be an unrealistic goal for helpers. This may be because it requires consistent use of lengthy responses that may feel unnatural to the back-and-forth flow of the session’s conversation in order to strongly convey all three sub-scale components. As such, further investigation should examine if increased EPITOME empathy scores linearly correlate to perceived empathy by the thinker.

While averaged levels of total empathy expressed were low, they were still significantly higher than that of untrained peers. This suggests initial training may have been a factor in increased empathy, but that more resources need to be dedicated to training peer supporters specifically in empathy skills, as it was found that peer support do not self-learn empathy over time. In fact, no trend was observed between the level of empathy a helper conveyed and the number of buddy sessions they had completed. A potential confounding factor in this analysis, was that information the thinker shared may also have played a role in a helper’s ability to convey empathy. This work focused on empathy as a behavior, but future work can be conducted to uncover if other factors in the session may have a causal effect on the empathy conveyed by peer supporters. This work also did not consider the demographic breakdown of helpers or thinkers, as data was collected without directly linked PII to protect users. Future work should investigate the perceptions of text-based empathy with respect to different user groups, in order to provide a fuller understanding of digital empathy.

As more resources should be dedicated to train peer supporters specifically in empathy skills, I propose suggested future work to investigate training methods that were uncovered from this research that demonstrate potential for

---

8For example, it may be beneficial for responses from helpers to obtain a 4 out of 6 on the EPITOME scale, but even higher empathy scores may be found to add no additional value. Alternatively, it may be beneficial for responses to have strong scores in 2 out of 3 sub-scales per response, but not necessary for a response to score strong in each of the three sub-scales.
encouraging more empathetic responses from peer supporters. Some previous research [26, 28, 39] has focused on generating high empathy text through automated approaches, but findings from this work aim to identify potential new methods for training helpers to write more empathetic responses. Findings indicated a negative correlation between stronger exploration scores and thinker feedback, so further research should be conducted to understand the exact relationship between use of open-ended questions and the thinker feeling “heard or understood” by the peer support. Identifying the potential for limiting over-use of open ended questions and subsequently how this can be taught to the helper, could have major impacts in treatment outcomes. More research is needed to be conducted to examine the exact relationship in which live prompts, such as the assistive workflow, can affect the levels of empathy conveyed by peer supporters and the resulting impact on the efficacy of the session and the thinker’s experiences. In this research, levels of empathy were observed to be higher in the listening portion of the session, than in the session as a whole. One potential cause for this was due to the prompts that were used in the assistive workflow and specifically encouraged the use of restatements and explorations. This suggests that altering prompts in the assistive workflow may have an impact on the amount of empathy peer supporters convey or provide additional benefits, as previous research has shown guided peer text-chats can promote deeper discussions [33]. Future research should be conducted to investigate altering the assistive workflow to encourage more use of emotional responses and higher levels of specificity, to uncover the impact on both the empathy measurement scores and the thinker’s experience. Appendix A demonstrates a proof-of-concept that details one such alternative assistive workflow that may encourage higher empathy responses from the helper.

Lastly, I propose that more research is conducted into expanding the explainability of the model. In an area as sensitive as mental health support, explainability needs to go beyond the given rationales, providing extra layer of protection in aiding helpers to identify inaccurate predictions. While the EPITOME model uses fine-tuning to provide contextualized language understandings, more research can be conducted to understand how language plays a role in empathy scores and the perceived empathy levels by the thinker. The table in appendix B categorizes words that were most often found in high empathy and low empathy responses.

6 CONCLUSION

This research sought to examine the application of the EPITOME model for measuring empathy in text to Cheeseburger Therapy, a peer support platform. In doing so, I uncovered a greater understanding of empathy conveyed in online peer support and our current methods for measuring digital empathy. The results indicated that (1) empathy conveyed online in text by peer support is low (2) trained peer supporters convey higher levels of empathy than untrained peer support (3) use of explorations was negatively correlated with the thinker’s experience and (4) responses by peers in the listening portions of the conversations contained higher levels of empathy. Lastly, from these findings I identified suggestions for future investigation outlining practices that have the potential to increase empathetic responses by peer support. In the wake of the COVID-19 pandemic, there has been a rise of telemedicine services [7] like BetterHelp (https://www.betterhelp.com/) and Buddio (https://www.buddio.io/) that provide options to text with licensed therapists or peer support. As more individuals are turning to these internet platforms for mental health care, now more than ever before, understanding digital empathy conveyed from text is necessary in order to create empathic online environments and provide effective support.
7 ACKNOWLEDGEMENTS

Thank you to my advisors Zainab Iftikhar and Jeff Huang. Zainab, I have appreciated your support and mentorship, as we have both navigated our first year in the Brown HCI lab. Your research experience was a great guide throughout this tricky process. Jeff, I could have not done this without your invaluable expertise.

Thank you to everyone in the Brown University HCI lab who has provided feedback and support in this process. Thank you to Gabby Hoefer whose guidance really helped me through some of the ups and downs of the research process.

Thank you to Michael Toomim for creating Cheeseburger Therapy and for working on such an empowering mission. I could not have done this research without you and appreciate that you have supported me and allowed me to take this research in any direction.

Thank you to my reader Geoffroy de Clippel, I am truly grateful for your feedback.

Thank you to Tom Doeppener who has served as my advisor since freshman year and helped me to navigate computer science at Brown.

Thank you to Gabby Marcu who gave me an opportunity in high school, sparking my desire to continue research in Human-Computer Interaction at Brown.

Thank you to the Brown squash team and my housemates/best friends for being my support system and the greatest distraction from any academic stress these past four years.

Lastly, and most importantly, thank you to my parents and my sister, Jasmine. In everything in life you have been my backbone. Thank you to my mom, Vatsala, for always being there for me and believing in me. Thank you for always encouraging me to try new things, like writing a thesis.

REFERENCES


A PROOF-OF-CONCEPT: NEW ASSISTIVE WORKFLOW

- Ask the user what is troubling to you?
- Restate what the user finds troubling. Express warmth regarding the user’s trouble.
- Explain that we will be taking notes.
- Ask what is most troubling about this.
- Ask what event occurred before you thought <Thought>? Restate Event.
- Ask what event happened immediately before you felt <Feelings> and did <Behaviors>? Restate Event.
- Ask what thoughts do you have when <Event>? Restate the specific thought.
- Ask how did you feel when you thought <Thought>? Restate the specific feeling. Express warmth, compassion, and/or concern to the user’s feelings.
- What do you feel before you do <Behavior>? Restate the specific feeling.
- Ask what actions do you take or avoid when you feel <Feelings>? Restate the behavior.
- Ask how does ____ relate to ____? only if you are having difficulty connecting their story.
- Ask if there is anything missing from the summary if you feel you are missing the entire picture.
- Choose the one Thought to focus on.
- Label the thought.
- Restate the Feelings and Behaviors and how they connect to the Labelled thought.
- Ask what is a new thought that you could have in response to the Event?
- Ask what feelings come up for you when you think <New Thought>?
- Ask how are you feeling now, compared to when we started?
- Say when <Event> occurs again in the future, remember to practice the thought <New Thought>

This is an example of an altered assistive workflow that incorporates findings of this research that indicate potential to encourage helpers to convey more empathy. Changes examine the goal of: Red = limiting over use of explorations, Orange = encouraging emotional reactions, Blue = encouraging specificity.

B UNIGRAM TABLE

Words were only included in this table if they appeared in the dataset at least five times. There was a higher number of low empathy words.
<table>
<thead>
<tr>
<th>word</th>
<th>averaged empathy</th>
<th>word</th>
<th>averaged empathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>realized</td>
<td>4.45</td>
<td>extroverted</td>
<td>0.00</td>
</tr>
<tr>
<td>reward</td>
<td>4.20</td>
<td>unfairly</td>
<td>0.00</td>
</tr>
<tr>
<td>hah</td>
<td>4.07</td>
<td>tossing</td>
<td>0.00</td>
</tr>
<tr>
<td>hadn’t</td>
<td>3.93</td>
<td>binary</td>
<td>0.00</td>
</tr>
<tr>
<td>thrill</td>
<td>3.53</td>
<td>bland</td>
<td>0.00</td>
</tr>
<tr>
<td>breath</td>
<td>3.50</td>
<td>vibrant</td>
<td>0.00</td>
</tr>
<tr>
<td>contributed</td>
<td>3.33</td>
<td>reducing</td>
<td>0.00</td>
</tr>
<tr>
<td>adore</td>
<td>3.25</td>
<td>examining</td>
<td>0.00</td>
</tr>
<tr>
<td>self-expression</td>
<td>3.25</td>
<td>spiraling</td>
<td>0.00</td>
</tr>
<tr>
<td>triggers</td>
<td>3.25</td>
<td>degraded</td>
<td>0.00</td>
</tr>
<tr>
<td>she’s</td>
<td>3.20</td>
<td>addiction</td>
<td>0.00</td>
</tr>
<tr>
<td>shoes</td>
<td>3.20</td>
<td>categories</td>
<td>0.08</td>
</tr>
<tr>
<td>hopelessness</td>
<td>3.20</td>
<td>convinced</td>
<td>0.11</td>
</tr>
<tr>
<td>gives</td>
<td>3.17</td>
<td>externalizing</td>
<td>0.11</td>
</tr>
<tr>
<td>bother</td>
<td>3.17</td>
<td>spiral</td>
<td>0.17</td>
</tr>
<tr>
<td>walking</td>
<td>3.17</td>
<td>inherently</td>
<td>0.20</td>
</tr>
<tr>
<td>lover</td>
<td>3.14</td>
<td>varies</td>
<td>0.22</td>
</tr>
<tr>
<td>satisfied</td>
<td>3.11</td>
<td>holds</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table 6. High and low empathy words