A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support

WARNING: This paper contains content related to suicide and self-harm.

Ashish Sharma[♠] **Adam S. Miner**[♣♡] **David C. Atkins**[♢] **Tim Althoff**[♠]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Department of Psychiatry and Behavioral Sciences, Stanford University
[♡]Center for Biomedical Informatics Research, Stanford University
[♢]Department of Psychiatry and Behavioral Sciences, University of Washington
{ashshar,althoff}@cs.washington.edu
asmpsyd@stanford.edu datkins@uw.edu

Abstract

Empathy is critical to successful mental health support. Empathy measurement has predominantly occurred in synchronous, face-to-face settings, and may not translate to asynchronous, text-based contexts. Because millions of people use text-based platforms for mental health support, understanding empathy in these contexts is crucial. In this work, we present a computational approach to understanding how empathy is expressed in online mental health platforms. We develop a novel unifying theoretically-grounded framework for characterizing the communication of empathy in text-based conversations. We collect and share a corpus of 10k (post, response) pairs annotated using this empathy framework with supporting evidence for annotations (rationales). We develop a multi-task RoBERTa-based bi-encoder model for identifying empathy in conversations and extracting rationales underlying its predictions. Experiments demonstrate that our approach can effectively identify empathic conversations. We further apply this model to analyze 235k mental health interactions and show that users do not self-learn empathy over time, revealing opportunities for empathy training and feedback.

1 Introduction

Approximately 20% of people worldwide are suffering from a mental health disorder (Holmes et al., 2018). Still, access to mental health care remains a global challenge with widespread shortages of workforce (Olfson, 2016). Facing limited in-person treatment options and other barriers like stigma (White and Dorman, 2001), millions of people are turning to text-based peer support platforms such as TalkLife (talklife.co) to express emotions, share stigmatized experiences, and receive peer support (Eysenbach et al., 2004). However, while peer supporters on these platforms are motivated and well-intentioned to help others seeking support (henceforth *seeker*), they are untrained and typically unaware of best-practices in therapy.

Figure 1: Our framework of empathic conversations contains three empathy communication mechanisms – *Emotional Reactions*, *Interpretations*, and *Explorations*. We differentiate between no communication, weak communication, and strong communication of these factors. Our computational approach simultaneously identifies these mechanisms and the underlying rationale phrases (highlighted portions). *All examples in this paper have been anonymized using best practices in privacy and security (Matthews et al., 2017).*

In therapy, interacting empathically with seekers is fundamental to success (Bohart et al., 2002; Elliott et al., 2018). The lack of training or feedback to layperson peer supporters results in missed opportunities to offer empathic textual responses. NLP systems that understand conversational empathy could empower peer supporters with feedback and training. However, the current understanding of empathy is limited to traditional face-to-face, speech-based therapy (Gibson et al., 2016; Pérez-Rosas et al., 2017) due to lack of resources and methods for new asynchronous, text-based inter-

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5263–5276,
November 16–20, 2020.

actions (Patel et al., 2019). Also, while previous NLP research has focused predominantly on empathy as reacting with emotions of warmth and compassion (Buechel et al., 2018), a separate but key aspect of empathy is to communicate a cognitive understanding of others (Selman, 1980).

In this work, we present a novel computational approach to understanding how empathy is expressed in text-based, asynchronous mental health conversations. We introduce EPITOME,[1] a conceptual framework for characterizing communication of empathy in conversations that synthesizes and adapts the most prominent empathy scales from speech-based, face-to-face contexts to text-based, asynchronous contexts (§3). EPITOME consists of three communication mechanisms of empathy: *Emotional Reactions*, *Interpretations*, and *Explorations* (Fig. 1).

To facilitate computational modeling of empathy in text, we create a new corpus based on EPITOME. We collect annotations on a dataset of 10k (post, response) pairs from extensively-trained crowdworkers with high inter-rater reliability (§4).[2] We develop a RoBERTa-based bi-encoder model for identifying empathy communication mechanisms in conversations (§5). Our multi-task model simultaneously extracts the underlying supportive evidences, *rationales* (DeYoung et al., 2020), for its predictions (spans of input post; e.g., highlighted portions in Fig. 1) which serve the dual role of (1) explaining the model's decisions, thus minimizing the risk of deploying harmful technologies in sensitive contexts, and (2) enabling rationale-augmented feedback for peer supporters.

We show that our computational approach can effectively identify empathic conversations with underlying rationales (∼80% acc., ∼70% macro-f1) and outperforms popular NLP baselines with a 4-point gain in macro-f1 (§6). We apply our model to a dataset of 235k supportive conversations on TalkLife and demonstrate that empathy is associated with positive feedback from seekers and the forming of relationships. Importantly, our results suggest that most peer supporters do not self-learn empathy with time. This points to critical opportunities for training and feedback for peer supporters to increase the effectiveness of mental health support (Miner et al., 2019; Imel et al., 2015). Specifically, NLP-based tools could give actionable, real-time feedback to improve expressed empathy, and we demonstrate this idea in a small-scale proof-of-concept (§7).

2 Background

2.1 How to measure empathy?

Empathy is a complex multi-dimensional construct with two broad aspects related to emotion and cognition (Davis et al., 1980). The "emotion" aspect relates to the emotional stimulation in reaction to the experiences and feelings expressed by a user. The "cognition" aspect is a more deliberate process of understanding and interpreting the experiences and feelings of the user and communicating that understanding to them (Elliott et al., 2018).

Here, we study *expressed empathy* in text-based mental health support – empathy *expressed* or *communicated* by peer supporters in their textual interactions with seekers (cf. Barrett-Lennard (1981)).[3] Table 1 lists existing empathy scales in psychology and psychotherapy research. Truax and Carkhuff (1967) focus only on communicating cognitive understanding of others while Davis et al. (1980); Watson et al. (2002) also make use of expressing stimulated emotions.

These scales, however, have been designed for in-person interactions and face-to-face therapy, often leveraging audio-visual signals like expressive voice. In contrast, in text-based support, empathy must be expressed using textual response alone. Also, they are designed to operate on long, synchronous conversations and are unsuited for the shorter, asynchronous conversations of our context.

In this work, we adapt these scales to text-based, asynchronous support. We develop a new comprehensive framework for text-based, asynchronous conversations (Table 1; §3), use it to create a new dataset of empathic conversations (§4), a computational approach for identifying empathy (§5; §6), & gaining insights into mental health platforms (§7).

2.2 Computational Approaches for Empathy

Computational research on empathy is based on speech-based settings, exploiting audio signals

[1] **Em**Pathy **I**n **T**ext-based, asynchr**O**nous **ME**ntal health conversations

[2] Our dataset can be accessed from `https://bit.ly/2Rwy2gx`.

[3] Note that *expressed* empathy may differ from the empathy *perceived* by seekers. However, obtaining perceived empathy ratings from seekers is challenging in sensitive contexts and involves ethical risks. Psychotherapy research indicates a strong correlation of expressed empathy with positive outcomes and frequently uses it as a credible alternative (Robert et al., 2011).

		Context	Applicable to text-based peer-support	Communication Mechanisms		
				Emotional Reactions	Interpretations *(Cognitive)*	Explorations *(Cognitive)*
Scales	Truax and Carkhuff (1967)	Face-to-face therapy	✗	✗	✓	✓
	Davis et al. (1980)	Daily human interactions	✗	✓	✓	✗
	Watson et al. (2002)	Face-to-face therapy	✗	✓	✓	✓
Methods	Buechel et al. (2018)	Reaction to news stories	✗	✓	✗	✗
	Rashkin et al. (2019)	Emotionally grounded convs.	✗	✗*	✗*	✗*
	Pérez-Rosas et al. (2017)	Motivational interviewing	✗	✗	✓	✓
	EPITOME	Text-based, asynchronous support	✓	✓	✓	✓

Table 1: EPITOME incorporates both emotional and cognitive aspects of empathy that were previously only studied in face-to-face therapy and never computationally in text-based, asynchronous conversations. *Rashkin et al. (2019) implicitly enable empathic conversations through grounding in emotions instead of communication.

like pitch which are unavailable in text-based platforms (Gibson et al., 2016; Pérez-Rosas et al., 2017). Moreover, previous NLP research has predominantly focused on empathy as reacting with emotions of warmth and compassion (Buechel et al., 2018). For mental health support, however, communicating cognitive understanding of feelings and experiences of others is more valued (Selman, 1980). Recent work also suggests that grounding conversations in emotions implicitly makes them empathic (Rashkin et al., 2019). Research in therapy, however, highlights the importance of expressing empathy in interactions (Truax and Carkhuff, 1967). In this work, we present a computational approach to (1) understanding empathy expressed in textual, asynchronous conversations; (2) address both emotional and cognitive aspects of empathy.

3 Framework of Expressed Empathy

To understand empathy in text-based, asynchronous, peer-to-peer support conversations, we develop EPITOME, a new conceptual framework of expressed empathy (Fig. 1). In close collaboration with clinical psychologists, we adapt and synthesize existing empathy definitions and scales to text-based, asynchronous context. EPITOME consists of three communication mechanisms providing a comprehensive outlook of empathy – *Emotional Reactions*, *Interpretations*, and *Explorations*. For each of these mechanisms, we differentiate between – (0) peers not expressing them at all (*no communication*), (1) peers expressing them to some weak degree (*weak communication*), (2) peers expressing them strongly (*strong communication*).

Here, we describe our framework in detail using the following seeker post as context for all example responses: *I am about to have an anxiety attack.*

Emotional Reactions. Expressing emotions such as warmth, compassion, and concern, experienced by peer supporter after reading seeker's post. Expressing these emotions plays an important role in establishing empathic rapport and support (Robert et al., 2011). A **weak communication** of emotional reactions alludes to these emotions without the emotions being explicitly labeled (e.g., *Everything will be fine*). On the other hand, **strong communication** specifies the experienced emotions (e.g., *I feel really sad for you*).

Interpretations. Communicating an understanding of feelings and experiences inferred from the seeker's post. Such a cognitive understanding in responses is helpful in increasing awareness of hidden feelings and experiences, and essential for developing alliance between the seeker and peer supporter (Watson, 2007). A **weak communication** of interpretations contains a mention of the understanding (e.g., *I understand how you feel*) while a **strong communication** specifies the inferred feeling or experience (e.g., *This must be terrifying*) or communicates understanding through descriptions of similar experiences (e.g., *I also have anxiety attacks at times which makes me really terrified*).

Explorations. Improving understanding of the seeker by exploring the feelings and experiences not stated in the post. Showing an active interest in what the seeker is experiencing and feeling and probing gently is another important aspect of empathy (Miller et al., 2003; Robert et al., 2011). A **weak exploration** is generic (e.g., *What happened?*) while a **strong exploration** is specific and labels the seeker's experiences and feelings which the peer supporter wants to explore (e.g., *Are you feeling alone right now?*).

Consistent with existing scales, responses that

only give advice (*Try talking to friends*), only provide factual information (*mindful meditation overcomes anxiety*), or are offensive or abusive (*shut the f**k up*)[4] are not empathic and are characterized as no communication of empathy in our framework.

4 Data Collection

To facilitate computational methods for empathy, we collect data based on EPITOME.

4.1 Data Source

We use conversations on the following two online support platforms as our data source:

(1) TalkLife. TalkLife (talklife.co) is the largest global peer-to-peer mental health support network (talklife.co/about). It enables seekers to have textual interactions with peer supporters through conversational threads. The dataset contains 6.4M threads and 18M interactions (seeker post, response post pairs) on TalkLife between May 2012 to Jan 2019.

(2) Mental Health Subreddits. Reddit (reddit.com) hosts a number of sub-communities aka *subreddits* (e.g., *r/depression*). We use threads posted on 55 mental health focused subreddits (Sharma et al. (2018)). This publicly accessible dataset contains 1.6M threads and 8M interactions on Reddit between Jan 2015 to Jan 2019.

We use the entire dataset for in-domain pretraining (§5) and annotate a subset of 10k interactions on empathy. We further analyze empathy on a carefully filtered dataset of 235k mental health interactions on TalkLife (§7).

4.2 Annotation Task and Process

Empathy is conceptually nuanced and linguistically diverse so annotating it accurately is difficult in short-term crowdwork approaches. This is also reflected in prior work that found it challenging to annotate therapeutic constructs (Lee et al., 2019). To ensure high inter-rater reliability, we designed a novel training-based annotation process.

Crowdworkers Recruiting and Training. We recruited and trained eight crowdworkers on identifying empathy mechanisms in EPITOME. We leveraged Upwork (upwork.com), a freelancing platform that allowed us to hire and work interactively with crowdworkers. Each crowdworker was trained through a series of phone calls (30 minutes to 1 hour in total) and manual/automated feedback on 50-100 posts. Refer Appendix A for more details.

	Data Source	No	Weak	Strong	Total
Emotional Reactions	TalkLife	3656	2945	461	7062
	Reddit	2034	899	148	3081
Interpretations	TalkLife	5533	178	1351	7062
	Reddit	1645	115	1321	3081
Explorations	TalkLife	5137	767	1158	7062
	Reddit	2600	104	377	3081

Table 2: Statistics of the collected empathy dataset. The crowdworkers were trained on EPITOME through a series of phone calls and manual/automated feedback on sample posts to ensure high quality annotations.

Annotating Empathy. In our annotation task, crowdworkers were shown a pair of (seeker post, response post) and were asked to identify the presence of the three communication mechanisms in EPITOME (Emotional Reactions, Interpretations, and Explorations), one at a time. For each mechanism, crowdworkers annotated whether the response post contained no communication, weak communication, or strong communication of empathy in the context of the seeker post.

Highlighting Rationales. Along with the categorical annotations, crowdworkers were also asked to highlight portions of the response post that formed rationale behind their annotation. E.g, in the post "*That must be terrible! I'm here for you*", the portion "*That must be terrible*" is the rationale for it being a strong communication of interpretations.

Data Quality. Overall, our corpus has an average inter-annotator agreement of 0.6865 (average over pairwise Cohen's κ of all pairs of crowdworkers; each pair annotated >50 posts in common) which is higher than previously reported values for the annotation of empathy in face-to-face therapy (∼0.60 in Pérez-Rosas et al., 2017; Lord et al., 2015).[5] Our ground-truth corpus contains 10,143 (seeker post, response post) pairs with annotated empathy labels from trained crowdworkers (Table 2).

Privacy and Ethics. The TalkLife dataset was sourced with license and consent from the TalkLife platform. All personally identifiable information (user and platform identifiers) in both the datasets were removed. This study was approved by University of Washington's Institutional Review Board.

[4] Our approach is focused on supporting peers who are trying to help seekers. This is different from toxic language identification tasks. Such content can be independently flagged using existing techniques (e.g., perspectiveapi.com)

[5] We achieve an inter-annotator agreement of 0.69 for emotional reactions, 0.61 for interpretations, and 0.76 for explorations.

In addition, we tried to minimize the risks of annotating mental health related content by providing crisis management resources to our annotators, following Sap et al. (2020). This work does not make any treatment recommendations or diagnostic claims.

5 Model

With our collected dataset, we develop a computational approach for understanding empathy.

5.1 Problem Definition

Let $\mathbf{S}_i = s_{i1}, ..., s_{im}$ be a seeker post and $\mathbf{R}_i = r_{i1}, ..., r_{in}$ be a corresponding response post. For the pair $(\mathbf{S}_i, \mathbf{R}_i)$, we want to perform two tasks:

Task 1: Empathy Identification. Identify how empathic $\mathbf{R}_i$ is in the context of $\mathbf{S}_i$. For each of the three communication mechanisms in EPITOME (Emotional Reactions, Interpretations, Explorations), we want to identify their level of communication (l_i) in $\mathbf{R}_i$ – no communication (0), weak communication (1), or strong communication (2).

Task 2: Rationale Extraction. Extract rationales underlying the identified level $l_i \in \{$no, weak, strong$\}$ of each of the three communication mechanism in EPITOME. The extracted rationale is a subsequence of words x_i in $\mathbf{R}_i$. We represent this subsequence as a mask $m_i = (m_{i1}, ..., m_{in})$ over the words in $\mathbf{R}_i$, where $m_{ij} \in \{0, 1\}$ is a boolean variable: 1 – rationale word; 0 – non-rationale word. Correspondingly, $x_i = m_i \odot \mathbf{R}_i$.

5.2 Bi-Encoder Model with Attention

We propose a multi-task bi-encoder model based on RoBERTa (Liu et al., 2019) for identifying empathy and extracting rationales (Fig. 2). We multi-task over the two tasks of empathy identification and rationale extraction and train three independent but identical architectures for the three empathy communication mechanisms in EPITOME (§3). The bi-encoder architecture (Humeau et al., 2019) facilitates a joint modeling of $(\mathbf{S}_i, \mathbf{R}_i)$ pairs. Moreover, the use of attention helps in providing context from the seeker post, $\mathbf{S}_i$. We find that such an approach is more effective than methods that concatenate $\mathbf{S}_i$ with $\mathbf{R}_i$ with a `[SEP]` token to form a single input sequence (§6).

Two Encoders. Our model uses two independently pre-trained transformer encoders from RoBERTa$_{\text{BASE}}$ – S-Encoder & R-Encoder – for encoding seeker post and response post respectively.

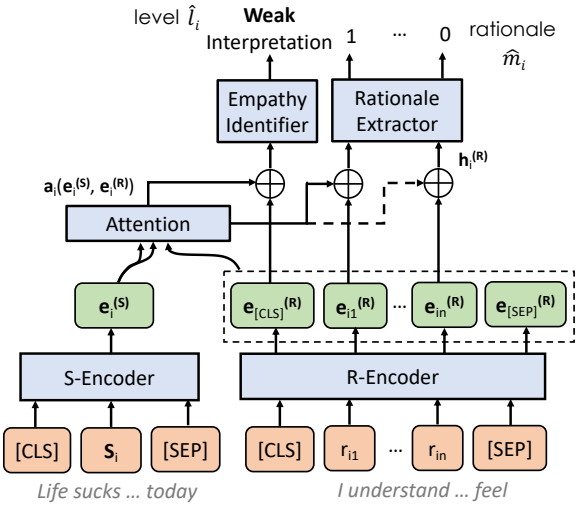


Figure 2: We use two independently pre-trained RoBERTa-based encoders for encoding seeker post and response post respectively. We leverage attention between them for generating seeker-context aware representation of the response post, used to perform the two tasks of empathy identification and rationale extraction.

S-Encoder encodes context from the seeker post whereas R-Encoder is responsible for understanding empathy in the response post.

$$\mathbf{e}_i^{(\mathbf{S})} = \text{S-Encoder}(\texttt{[CLS]}, \mathbf{S}_i, \texttt{[SEP]}) \quad (1)$$

$$\mathbf{e}_i^{(\mathbf{R})} = \text{R-Encoder}(\texttt{[CLS]}, \mathbf{R}_i, \texttt{[SEP]}) \quad (2)$$

where `[CLS]` and `[SEP]` are special start and end tokens adapted from BERT (Devlin et al., 2019).

Domain-Adaptive Pre-training. Both the S-Encoder and R-Encoder are initialized using the weights learned by RoBERTa$_{\text{BASE}}$. We further perform a domain-adaptive pre-training (Gururangan et al., 2020) of the two encoders to adapt to conversational and mental health context. For this additional pre-training of the two encoders, we use the datasets of 6.4M seeker posts (182M tokens) and 18M response (279M tokens) posts respectively sourced from TalkLife (§4). We use the masked language modeling task for pre-training (3 epochs, batch size = 8).

Attention Layer. We use a single-head attention over the two encodings for generating seeker-context aware representation of the response post. Using the terminology of transformers (Vaswani et al., 2017), our query is the response post encoding $\mathbf{e}_i^{(\mathbf{R})}$, and the keys and the values are the seeker post encoding $\mathbf{e}_i^{(\mathbf{S})}$. Our attention is computed as:

$$\mathbf{a}_i(\mathbf{e}_i^{(\mathbf{R})}, \mathbf{e}_i^{(\mathbf{S})}) = \text{softmax}\left(\frac{\mathbf{e}_i^{(\mathbf{R})}\mathbf{e}_i^{(\mathbf{S})}}{\sqrt{d}}\right)\mathbf{e}_i^{(\mathbf{S})} \quad (3)$$

	Model	Emotional Reactions		Interpretations		Explorations	
		acc.	f1	acc.	f1	acc.	f1
TalkLife	Log. Reg.	58.02	51.58	55.53	41.19	63.23	51.97
	RNN	69.09	54.02	82.25	47.94	73.40	28.22
	HRED	78.91	48.70	79.26	29.48	73.40	28.22
	BERT	76.98	70.31	85.06	62.24	85.87	71.56
	GPT-2	76.89	70.76	80.00	58.43	83.25	65.65
	DialoGPT	76.71	70.42	85.67	66.60	83.95	66.34
	RoBERTa	78.28	71.06	86.25	62.69	85.79	71.83
	Our Model	**79.93**	**74.29**	**87.50**	**67.46**	**86.92**	**73.47**
Reddit	Log. Reg.	41.69	42.69	70.58	49.77	67.08	46.63
	RNN	71.63	42.85	76.21	51.76	85.58	30.74
	HRED	71.11	44.10	79.65	54.16	85.58	30.74
	BERT	72.13	50.41	82.16	61.20	89.35	56.54
	GPT-2	76.69	71.65	82.32	62.27	88.25	58.28
	DialoGPT	66.07	51.16	81.85	**68.95**	89.65	70.65
	RoBERTa	76.99	70.35	82.16	61.38	90.58	63.41
	Our Model	**79.43**	**74.46**	**84.04**	62.60	**92.61**	**72.58**

Table 3: Empathy identification task results. We observe substantial gains over baselines with our seeker-context aware, mult-tasking approach.

	Model	Emotional Reactions		Interpretations		Explorations	
		T-f1	IOU	T-f1	IOU	T-f1	IOU
TalkLife	Log. Reg.	47.44	63.27	46.92	32.97	47.18	62.25
	RNN	62.80	58.22	67.26	57.31	63.29	64.65
	HRED	60.56	55.01	64.26	70.92	61.54	70.85
	BERT	61.29	51.20	61.06	67.33	62.50	64.80
	GPT-2	47.39	51.27	64.06	81.12	**66.71**	78.21
	DialoGPT	66.24	61.24	64.05	79.64	57.95	76.95
	RoBERTa	59.12	63.82	60.08	84.85	60.05	78.21
	Our Model	**68.49**	**66.82**	**67.81**	**85.76**	64.56	**83.19**
Reddit	Log. Reg.	43.26	61.27	49.85	31.31	48.21	70.36
	RNN	45.54	43.94	48.22	51.35	65.11	78.27
	HRED	46.34	45.65	48.88	52.12	66.66	80.33
	BERT.	51.06	54.81	48.38	50.75	67.91	71.00
	GPT-2	51.44	57.10	54.53	52.38	73.39	82.89
	DialoGPT	51.83	49.37	54.43	55.85	**73.43**	**85.20**
	RoBERTa	51.89	58.31	55.62	54.60	69.76	83.33
	Our Model	**53.57**	**64.83**	**57.40**	**55.90**	71.56	84.48

Table 4: Rationale extraction task results. We evaluate both at the level of tokens (T-f1) and spans (IOU-f1).

where $d = 768$ (hidden size in RoBERTa$_{\text{BASE}}$). We sum the encoded response $\mathbf{e}_i^{(\mathbf{R})}$ with its representation transformed through attention $\mathbf{a}_i(\mathbf{e}_i^{(\mathbf{R})}, \mathbf{e}_i^{(\mathbf{S})})$ to obtain a residual mapping (He et al., 2016) – $\mathbf{h}_i^{(\mathbf{R})}$, which forms the final seeker-context aware representation of the response post.

Empathy Identification. For the task of identifying empathy, we use the final representation of the `[CLS]` token in the response post ($\mathbf{h}_i^{(\mathbf{R})}$[`[CLS]`]) and pass it through a linear layer to get the predictions of the empathy level $\hat{l}_i$ (0, 1, or 2) of each empathy communication mechanism. Note that we train three independent models for the three communication mechanisms in EPITOME (§3).

Extracting Rationales. For extracting rationales y_i underlying the predictions, we use final representations of the individual tokens in $\mathbf{R}_i$ ($\mathbf{h}_i^{(\mathbf{R})}[r_{i1}, ..., r_{in}]$) and pass them through a linear layer for making boolean predictions, $\hat{m}_i$.

Loss Function. We use cross-entropy between the true and predicted labels as the loss functions of our two tasks. The overall loss of our multi-task architecture is: $\mathcal{L} = \lambda_{\mathbf{EI}} * \mathcal{L}_{\mathbf{EI}} + \lambda_{\mathbf{RE}} * \mathcal{L}_{\mathbf{RE}}$.

Experimental Setup. We split both datasets into train, dev, and test sets (75:5:20). We train our model for 4 epochs using a learning rate of 2e$-$5, batch size of 32, $\lambda_{\mathbf{EI}} = 1$, and $\lambda_{\mathbf{RE}} = 0.5$ (Refer Appendix B for fine-tuning details).

6 Results

Next, we analyze how effectively we can identify empathy with underlying rationales using our computational approach.

6.1 Overall Results

We compare the performance of our approach with a range of models popularly used in related tasks (e.g., sentiment classification, conversation analysis). We quantify how challenging identifying empathy with underlying rationales is, how well do existing models perform, and what performance is achieved by our proposed approach.

Baselines. Our baselines are: **1.** Log. reg. (logistic regression over tf.idf vectors); **2.** RNN (two-layer recurrent neural network); **3.** HRED (hierarchical recurrent encoder-decoder, often used for modeling conversations (Sordoni et al., 2015)); **4.** BERT$_{\text{BASE}}$ (Devlin et al., 2019); **5.** GPT-2 (typically used for language generation (Radford et al., 2019)); **6.** DialoGPT (GPT-2 adapted to asynchronous conversations (Zhang et al., 2020)); and **7.** RoBERTa$_{\text{BASE}}$ (Liu et al., 2019).

Empathy Identification Task. Table 3 reports the accuracy and macro-f1 scores of the three communication mechanisms (random baseline for each is 33% accurate; three levels). Log. reg., RNN, and HRED struggle to identify empathy with noticeably low macro-f1 scores indicative of failures to distinguish between the three levels of communication. Among the baseline transformer architectures, we obtain best performance using RoBERTa but observe substantial gains over them with our approach

	Model	Emotional Reactions				Interpretations				Explorations			
		identification		rationale		identification		rationale		identification		rationale	
		acc.	f1	T-f1	IOU	acc.	f1	T-f1	IOU	acc.	f1	T-f1	IOU
TalkLife	**Our Model**	**79.93**	**74.29**	**68.49**	**66.82**	**87.50**	**67.46**	67.81	**85.76**	**86.92**	**73.47**	**64.56**	**83.19**
	-attention	79.00	73.02	59.59	63.49	87.41	66.97	67.12	79.20	84.86	63.45	59.42	73.82
	-seeker post	79.37	73.52	61.08	62.58	86.04	63.23	65.56	77.23	86.16	70.80	60.05	81.87
	-rationales	79.12	71.21	–*	–*	87.01	66.71	–*	–*	86.38	72.14	–*	–*
	-pre-training	78.95	73.41	60.34	62.91	87.31	65.86	**69.03**	84.95	86.21	70.54	64.53	80.19
Reddit	**Our Model**	**79.43**	**74.46**	**53.57**	**64.83**	**84.04**	**62.60**	57.40	**55.90**	**92.61**	**72.58**	**71.56**	**84.48**
	-attention	75.51	52.66	51.79	59.83	83.26	62.25	54.90	52.79	91.98	64.75	68.81	81.91
	-seeker post	79.15	71.47	45.87	58.56	83.57	62.41	55.59	55.51	91.67	64.59	68.73	81.56
	-rationales	78.50	73.21	–*	–*	83.26	62.13	–*	–*	91.51	64.44	–*	–*
	-pre-training	76.97	69.03	51.58	57.35	82.32	61.38	**57.61**	55.34	91.99	65.26	70.44	81.71

Table 5: Ablation results. Most of our gains are due to context provided through attention and seeker post; higher gains for the rationale extraction task. *Note that rationales cannot be predicted after removing them from training.

(+1.73 acc., +4.02 macro-f1 over RoBERTa). We analyze the sources of these gains in §6.2.

Rationale Extraction Task. We perform both token level and span level evaluation for this task. We use two metrics, commonly used in discrete rationale extraction tasks (DeYoung et al., 2020): **1.** T-f1 (token level f1); **2.** IOU-f1 (intersection over union overlap of predicted spans with ground truth spans; threshold of 0.5 on the overlap for finding true positives and the corresponding f1). We find that GPT-2 and DialoGPT perform better than BERT and RoBERTa likely due to appropriateness to the related task of generating free-text rationales (Table 4). Our approach obtains gains of +2.58 T-f1 and +6.45 IOU-f1 over DialoGPT, potentially due to the use of attention and seeker post (§6.2).

6.2 Ablation Study

We next analyze the components and training strategies in our approach through an ablation study.

No Attention. Instead of using attention, we concatenate the seeker post encoding ($\mathbf{e}_i^{(\mathbf{S})}$) with the response post encoding ($\mathbf{e}_i^{(\mathbf{R})}$) and use the concatenated representation as input to the linear layer.

No Seeker Post. We train without the S-Encoder, i.e., by only encoding from the R-Encoder.

No Rationales. We set $\lambda_{\mathbf{RE}}$ to 0 and only train on the empathy identification task.

No Domain-Adaptive Pre-training. We initialize by only using model weights from RoBERTa$_{\text{BASE}}$.

Results. Our most significant gains come from using attention and the seeker post (Table 5) which greatly benefits the rationale extraction task (+4.88 T-f1, +5.74 IOU-f1). Also, using rationales and pre-training only leads to small performance improvements.

6.3 Error Analysis

We qualitatively analyze the sources of our errors. For the empathy identification task, we found that the model sometimes failed to identify short expressions of emotions in responses that otherwise contained a lot of instructions (e.g., *Sorry to hear that! Try doing ...*). Also, certain responses trying to universalize the situation (e.g., *You are not alone*) got incorrectly identified as strong interpretations. Furthermore, a source of error for explorations was confusions due to questions that were not an exploration of seeker's feelings or experiences (e.g., offers to talk - *Do you want to talk?*).

For the rationale extraction task, the model accurately extracts phrases that are commonly used for expressing empathy (e.g., *this might be tough*), but also sometimes incorrectly extracts single words from sentences indicating errors in disambiguating word usage (e.g., 'what' in *Tell them what you need* gets extracted as an exploration, likely because *what* is really common in explorations as in *what happened?*).

7 Model-based Insights into Mental Health Platforms

We apply our model to study how empathy impacts online peer-to-peer support dynamics. To only focus on conversations related to significant mental health challenges and filter out common social media interactions (e.g., *Merry Christmas*), we carefully select 235k mental health related interactions on TalkLife using a seeker-reported indicator.[6]

[6]We focus analyzing TalkLife alone as Reddit lacks rich publicly available signals such as seeker liking the response.

Seeker Post	Original Response	Re-written Response
I cannot do anything without getting blamed today. This day is getting worse and worse.	Days end, tomorrow is a fresh start.	**I'm sorry that today sucks**, but tomorrow is a fresh start.
An hour ago i was happy an hour later i'm sad. Am i getting mad now?	Try mindful meditation which can control anxiety	**That's something I've struggled with too**, and **it really pains me to hear that you're dealing with the same thing. Have you considered trying meditation?** I've found it to be very helpful.

Table 6: Example re-written responses with our model-based feedback. Participants increased empathy from 0.8 to 3.0. **blue** = Strong emo. reactions, **light red**/**dark red** = Weak/Strong Interpretations, **green** = Strong explorations.

We investigate (1) the levels of empathy on the platform, its variation over time, and examine the relationship of empathy with (2) conversation outcomes, (3) relationship forming, and (4) gender.

(1) Peer supporters do not self-learn empathy over time. Overall, we observe that empathy expressed by peer supporters on the platform is low (avg. total score[7] of 1.09 out of 6). In addition, we find that the emotional reactivity of users decreases over time (36% decrease over three years) and their levels of interpretations and explorations remain practically constant (Fig. 3a). Further analyzing whether a user individually improves, worsens, or remains constant in their expression of empathy, we find that 69% users either worsen or stay constant in their empathy expressions and only 10% improve by at least one point (i.e. one level in our framework). This is also reflected in prior work on therapy that shows that without deliberate practice and specific feedback, even trained therapists often diminish in skills over time (Goldberg et al., 2016). We find this trend robust to potential confounding factors (new users, user dropout) and users of different groups (low vs. high activity users, moderators; Appendix C). This indicates that most users do not self-learn empathy and highlights the need of providing them feedback.

(2) High empathy interactions are received positively by seekers. We analyze the correlation of empathic conversations with positive feedback, concretely with seeker "liking" the post. We find that strong communications of empathy are received with 45% more likes by seekers than no communication (Fig. 3b). Strong explorations get 44% less likes but receive 47% more replies than no explorations, leading to higher engagement.

(3) Relationship forming more likely after empathic conversations. Psychology research emphasizes the importance of empathy in forming alliance and relationship with seekers (Watson, 2007). Here, we operationalize relationship forming as seeker "following" the peer supporter after a conversation (within 24hrs). We find that seekers are 79% more likely to follow peer supporters after an empathic conversation (total score of 1+ vs. 0) than after a non-empathic one (Fig. 3c).

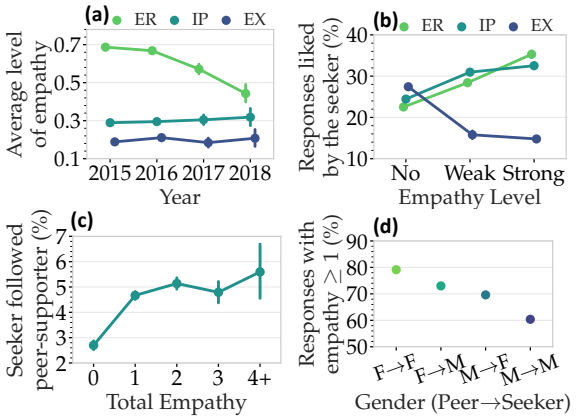


Figure 3: (a) Peer-supporters do not self-learn empathy over time. Only users who joined in 2015 were included but similar trends hold for other user groups; (b) Stronger communications of emotional reactions and interpretations are received positively by seekers. Stronger explorations get 47% more replies; (c) A lot more seekers follow peers after empathic interactions; (d) Females are more empathic towards females.

(4) Females are more empathic with females than males are with males. Previous work has shown that seekers identifying as females receive more support in online communities (Wang and Jurgens, 2018). Here, we ask if empathic interactions are affected by the self-reported gender of seekers and peer supporters. We find that female peer supporters are 32% more empathic towards female seekers than males are towards male seekers (Fig. 3d). Also, females are 6% more empathic towards males than males are towards females.

Implications for empathy-based feedback. These results suggest that our approach not only successfully measures empathy according to a

[7]Total empathy score is obtained by adding the level of communication across the three mechanisms.

principled framework (§3), but that the measured empathy components are important to online supportive conversations as indicated by the positive reactions from seekers and meaningful reflections of social theories. However, peer supporters on the platform express empathy rarely and this does not improve over time. This points to critical opportunities for empathy-based feedback to peer supporters for making their interactions with seekers more effective. Here, we demonstrate the potential of feedback in a simple proof-of-concept. When providing three participants (none are co-authors) simple feedback (Appendix D) based on EPITOME and our best-performing model, they were able to increase empathy in responses from 0.8 to 3.0 (total empathy across the three mechanisms). Table 6 shows two such examples of re-written responses that improve in communicating cognitive understanding (*today sucks*) and are also better with emotional reactions (*I'm sorry, it pains me*) and explorations (*Have you considered trying mindful meditation?*).

8 Further Related Work

Previous work in NLP for mental health has focused on analysis of effective conversation strategies (Althoff et al., 2016; Pérez-Rosas et al., 2019; Zhang and Danescu-Niculescu-Mizil, 2020), identification of therapeutic actions (Lee et al., 2019), and language development of counselors (Zhang et al., 2019). Researchers have also analyzed linguistic accommodation (Sharma and De Choudhury, 2018), cognitive restructuring (Pruksachatkun et al., 2019), and self-disclosure (Yang et al., 2019). We extend these studies and analyze empathy which is key in counseling and mental health support. Previous research has analyzed empathy in health communities (Khanpour et al., 2017), face-to-face therapy (Gibson et al., 2016), motivational interviewing (Pérez-Rosas et al., 2017), and emotionally-grounded conversations (Rashkin et al., 2019). Small-scale studies on manually annotated datasets have also been conducted (Morris and Picard, 2012; Lord et al., 2015). Our work develops a computational method for identifying empathy in text-based, asynchronous mental health support which is grounded in psychology and psychotherapy research and provides deeper insights into mental health platforms. Recent work has also developed proof-of-concept prototypes, such as ClientBot (Huang et al., 2020), for training users in counseling. Our approach is aimed towards developing empathy-based feedback and training systems for peer supporters (consistent with calls to action for improved treatment access and training (Miner et al., 2019; Imel et al., 2015; Kazdin and Rabbitt, 2013)).

9 Conclusion

We developed a new framework, dataset, and computational method for understanding expressed empathy in text-based, asynchronous conversations on mental health platforms. Our computational approach effectively identifies empathy with underlying rationales. Moreover, the identified components are found to be important to mental health platforms and helpful in improving peer-to-peer support through model-based feedback.

Acknowledgments

We would like to thank TalkLife and Jamie Druitt for their support and for providing us access to a TalkLife dataset. We also thank the members of UW Behavioral Data Science group, UW NLP group, Zac E. Imel, and the anonymous reviewers for their feedback on this work. A.S. and T.A. were supported in part by NSF grant IIS-1901386, Bill & Melinda Gates Foundation (INV-004841), an Adobe Data Science Research Award, the Allen Institute for Artificial Intelligence, and a Microsoft AI for Accessibility grant. A.S.M. was supported by grants from the National Institutes of Health, National Center for Advancing Translational Science, Clinical and Translational Science Award (KL2TR001083 and UL1TR001085) and the Stanford Human-Centered AI Institute. D.C.A. was supported in part by a NIAAA K award (K02 AA023814).

Conflict of Interest Disclosure. Dr. Atkins is a co-founder with equity stake in a technology company, Lyssn.io, focused on tools to support training, supervision, and quality assurance of psychotherapy and counseling.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *TACL*, 4:463–476.

Godfrey T Barrett-Lennard. 1981. The empathy cycle: Refinement of a nuclear concept. *Journal of counseling psychology*, 28(2):91.

Arthur C Bohart, Robert Elliott, Leslie S Greenberg, and Jeanne C Watson. 2002. Empathy. *J. C. Norcross (Ed.), Psychotherapy relationships that work: Therapist contributions and responsiveness to patients*.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *EMNLP*.

Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. *Journal of Personality and Social Psychology*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399.

Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj*, 328(7449):1166.

James Gibson, Doğan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Interspeech*.

Simon B Goldberg, Tony Rousmaniere, Scott D Miller, Jason Whipple, Stevan Lars Nielsen, William T Hoyt, and Bruce E Wampold. 2016. Do psychotherapists improve with time and experience? a longitudinal analysis of outcomes in a clinical setting. *Journal of counseling psychology*, 63(1):1.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Emily A Holmes, Ata Ghaderi, Catherine J Harmer, Paul G Ramchandani, Pim Cuijpers, Anthony P Morrison, Jonathan P Roiser, Claudi LH Bockting, Rory C O'Connor, Roz Shafran, et al. 2018. The lancet psychiatry commission on psychological treatments research in tomorrow's science. *The Lancet Psychiatry*, 5(3):237–286.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *CoRR abs/1905.01969. External Links: Link Cited by*, 2:2–2.

Zac E Imel, Mark Steyvers, and David C Atkins. 2015. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, 52(1):19.

Alan E Kazdin and Sarah M Rabbitt. 2013. Novel models for delivering mental health services and reducing the burdens of mental illness. *Clinical Psychological Science*, 1(2):170–191.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *IJCNLP)*, pages 246–251.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathleen McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 12–23.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. 2017. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *CHI*.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in psychiatry*, 10.

Robert R Morris and Rosalind Picard. 2012. Crowdsourcing collective emotional intelligence. *arXiv preprint arXiv:1204.3481*.

Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*, 35(6):983–990.

Sundip Patel, Alexis Pelletier-Bui, Stephanie Smith, Michael B Roberts, Hope Kilgannon, Stephen Trzeciak, and Brian W Roberts. 2019. Curricula for empathy and compassion training in medical education: A systematic review. *PloS one*, 14(8).

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *ACL*.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *ACL*.

Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *CHI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*.

Elliot Robert, Arthur C Bohart, JC Watson, and LS Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43–49.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Robert L Selman. 1980. *Growth of interpersonal understanding*. Academic Press.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *CHI*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *CIKM*.

CB Truax and RR Carkhuff. 1967. Modern applications in psychology. *Toward effective counseling and psychotherapy: Training and practice. Hawthorne, NY, US: Aldine Publishing Co*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *EMNLP*.

Jeanne C Watson. 2007. Facilitating empathy. *European Psychotherapy*, 7(1):59–65.

Jeanne C Watson, Rhonda N Goldman, and Margaret S Warner. 2002. *Client-centered and experiential psychotherapy in the 21st century: Advances in theory, research, and practice*. PCCS Books.

Marsha White and Steve M Dorman. 2001. Receiving social support online: implications for health education. *Health education research*, 16(6):693–707.

Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *CHI*.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *ACL*.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding your voice: The linguistic development of mental health counselors. In *ACL*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

A Data Collection Details

A.1 Annotation Instructions

For each (seeker post, response post) pair, the annotators were asked the following four questions:

1. **(Mental Health Related)** Is the seeker talking about a mental health related issue or situation in his/her post?[8]
   - Yes
   - No
2. **(Emotional Reactions)** Does the response express or allude to warmth, compassion, concern or similar feelings of the responder towards the seeker?
   - No
   - Yes, the response alludes to these feelings but the feelings are not explicitly expressed
   - Yes, the response has an explicit mention of these feelings
3. **(Interpretations)** Does the response communicate an understanding of the seeker's experiences and feelings? In what manner?
   - No
   - Yes, the response communicates an understanding of the seeker's experiences and/or feelings

   If the answer to the above question was "Yes", the annotators were further asked to annotate one or more of the following:
   - The response contains conjectures or speculations about the seeker's experiences and/or feelings
   - The responder has reflected back on similar experiences of their own or others
   - The responder has also described similar experiences of their own or others
   - The response contains paraphrases of the seeker's experiences and/or feelings
4. **(Explorations)** Does the response make an attempt to explore the seeker's experiences and feelings?
   - No
   - Yes, but the exploration is generic
   - Yes, and the exploration is specific

   The detailed instructions can be found at https://mhannotate-test.cs.washington.edu/annotate/readme.html.

[8] We use this question for filtering non-mental related posts from the data collection process

A.2 Interactive Training of Crowdworkers

The crowdworkers on Upwork were initially provided with our entire annotation instructions and an interactive training system[9] containing ten examples. After this initially automated training, we scheduled a 1hour long phone call with them to discuss our annotation instructions and annotation interface. During the phone call, crowdworkers also asked questions on the annotation guidelines which greatly helped in addressing potential ambiguities. After the phone call, we assigned them 20 tasks each (randomly chosen; different for each crowdworker). We manually evaluated the annotations on those 20 tasks. Based on the evaluation, we either decided to discontinue with the crowdworker (there were two such crowdworkers) or we provided them further manual feedback. Throughout the process, crowdworkers actively asked questions through the chat feature on Upwork. After the initial training phase, we also did spot checks on quality (at least two times for each crowdworker; $\geq$ 20 posts each) to provide them further feedback.[10]

B Reproducibility

B.1 Implementation Details

Code. Our codes are based on the huggingface library (https://huggingface.co/). We make them publicly available at https://github.com/behavioral-data/Empathy-Mental-Health.

Seed Value. For all our experiments, we used the seed value of 12.

B.2 Hyperparameter Fine-tuning

We searched through the following space of hyperparameters for fine-tuning our model:

- learning rate = 1e-5, 2e-5, 5e-5, 1e-4, 5e-4
- $\lambda_{\mathbf{EI}}$ = 1
- $\lambda_{\mathbf{RE}}$ = 0.1, 0.2, 0.5, 1

[9] This system contained prompts of manually written feedback for both correct and incorrect annotations.

[10] Crowdworkers only needed minor feedback on these posts.

B.3 Runtime Analysis

Domain-Adaptive Pre-training Time. We conducted domain-adaptive pre-training on four RTX 2080 Ti GPUs. Pre-training S-Encoder took around 22 hours. Pre-training R-Encoder took around 38 hours. Both are pre-trained for three epochs.

Model Training Time. We trained our model on one RTX 2080 Ti GPU. The training approximately takes five minutes. Our model is trained for four epochs.

B.4 Train, Dev, Test Splits

We split both the datasets into train, dev, and test sets in the ratio of 75:5:20. Table 7 contains the statistics of the train, dev, and test splits.

B.5 Number of Parameters

The total number of parameters of our model = 2 * number of parameters of RoBERTa$_{\text{BASE}}$ + parameters in the linear layers $\approx$ 2*125M + 2 * .5M = 251M

B.6 Reddit dataset

The entire Reddit dataset can be accessed through its archive on Google BigQuery at https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2015_05?pli=1

C Potential confounding factors in analysis of variation of empathy over time

We note that such an analysis can be affected by several confounding factors such as old vs. new users, user dropout, and low activity of several users. To account for these factors, we stratify users by the year in which they started supporting on the platform (2015, 2016, 2017) and analyze the average levels of empathy during subsequent years in each stratum. We further filter users with < 10 posts and only consider users who stay on the platform for at least a year.

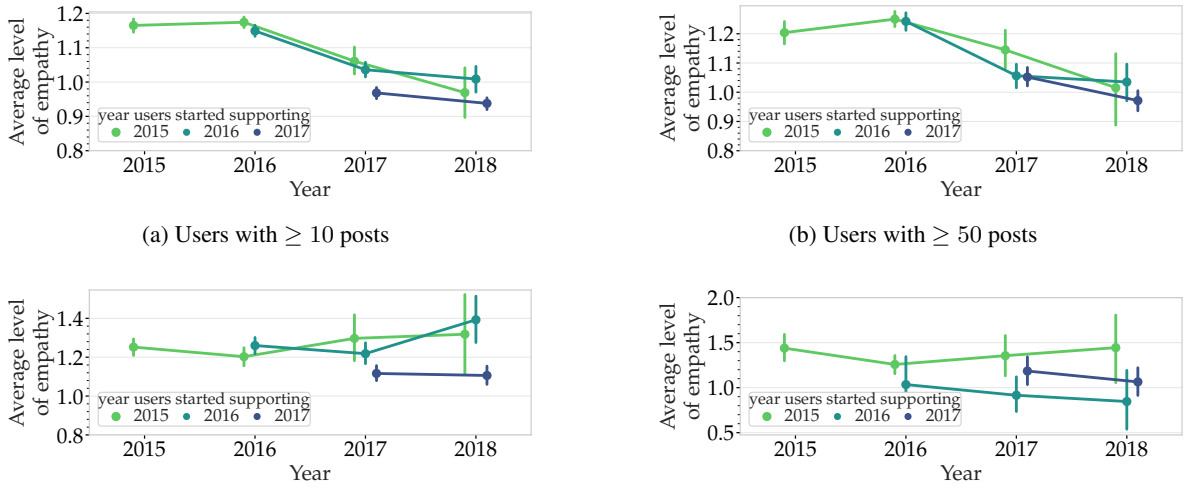
In addition, we analyze various user groups but observe similar trends (Fig. 4).

D Proof-of-Concept Details: model-based feedback for making responses empathic

We work with three computer science students with no training in counseling and give them (seeker post, response post) pairs identified low in empathy by our approach (total empathy score ≤ 1). We show them – (1) the levels of empathy predicted by our model, (2) extracted rationales, (3) a templated feedback explaining where the response lacks and how it can be made more empathic (based on the predicted levels, extracted rationales, definitions and examples in EPITOME). A sample feedback is shown below:

- **Seeker Post:** I'm hurt so much that I don't really have feelings anymore
- **Response Post:** Yeah, I felt it once
- **Feedback:**
  1. The response communicates an understanding of the seeker's post to a weak degree in the portion "I felt it once". The communication can be made stronger by talking about the seeker's feelings or experiences that you interpret after reading the post. Typically, they are expressed by saying "This must be terrible", "I know you are in a tough situation".
  2. It also lacks expressions of emotions of warmth, compassion, or concern and also does not attempt to explore the seeker's emotions or feelings. Typically, they are expressed by saying "I am feeling sorry for you", "What makes you feel depressed?"

We ask them to rewrite the response post making use of the templated feedback. Overall, the participants were comfortable to rewrite the responses with an average difficulty of 1.92 out of 5 (most difficult is 5) and found the feedback useful in the rewriting process with an average usefulness rating of 3.5 out of 5 (highly useful is 5).



(a) Users with $\geq$ 10 posts

(b) Users with $\geq$ 50 posts

(c) Users with $<$ 10 posts

(d) Moderators

Figure 4: Empathy over time analysis of various user groups. We find similar trends across multiple groups.

	Data Source	Train			Dev			Test		
		No	Weak	Strong	No	Weak	Strong	No	Weak	Strong
Emotional Reactions	TalkLife	52.02%	41.55%	6.43%	49.44%	44.66%	5.90%	52.28%	41.27%	6.45%
	Reddit	65.80%	29.52%	4.68%	66.87%	26.88%	6.25%	66.98%	27.39%	5.63%
Interpretations	TalkLife	78.39%	3.33%	18.28%	77.20%	4.00%	18.80%	79.26%	2.69%	18.04%
	Reddit	54.59%	3.63%	41.77%	48.12%	4.37%	47.5%	48.83%	3.91%	47.26%
Explorations	TalkLife	72.87%	10.56%	16.57%	73.88%	10.11%	16.01%	73.40%	11.09%	15.51%
	Reddit	83.41%	3.80%	12.79%	89.94%	62.89%	9.44%	85.60%	3.13%	11.27%

Table 7: Train/Dev/Test Splits.