2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII) | 979-8-3503-2743-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACII59096.2023.10388176

Therapist Empathy Assessment in Motivational Interviews

Leili Tavabi[1], Trang Tran[1], Brian Borsari[2],
Joannalyn Delacruz[2], Joshua D Woolley[2], Stefan Scherer[1], Mohammad Soleymani[1]
[1]Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA
[2]San Francisco VA Health Care System, University of California San Francisco, San Francisco, CA, USA
{ltavabi,ttran}@ict.usc.edu,
{brian.borsari, joannalyn.delacruz}@va.gov,
josh.woolley@ucsf.edu, {scherer,soleymani}@ict.usc.edu

***Abstract*—The quality and effectiveness of psychotherapy sessions are highly influenced by the therapists' ability to lead the conversation with empathy and acceptance. Manual assessment of the quality of therapy sessions is labor-intensive and difficult to scale. In this paper, we propose a method for estimating session-level therapist empathy ratings for Motivational Interviewing (MI) using therapist language, which has applications in clinical assessment and training. We analyze different stages within therapy sessions to investigate the importance of each stage and its topics of conversation in estimating session-level therapist empathy. We perform experiments on two datasets of MI therapy sessions for alcohol use disorder with session-level empathy scores provided by expert annotators. We achieve average CCC (Concordance Correlation Coefficient) scores of 0.596 and 0.408 for estimating therapist empathy under therapist-dependent and therapist-independent evaluation settings. Our results suggest that therapist responses to client's discussions on activities and experiences around the problematic behavior (in this case, alcohol abuse) along with the therapist's usage of in-depth reflections, are the most significant factors in the perception of therapists' empathy.**

***Index Terms*—Motivational Interviewing, empathy, self-attention, quartiles, natural language understanding**

I. Introduction

Empathy in psychotherapy has been described as: "To sense the client's private world as if it were your own, but without ever losing the 'as if' quality – this is empathy, and this seems essential to therapy" [1]. Empathy is hypothesized to be one of the key ingredients in creating a good therapeutic relationship, which in turn, is the best predictor of success in psychotherapy [2]. Indeed, greater therapist empathy has been linked to better therapeutic outcomes [3], [4] further highlighting the importance of empathy in counseling.

In this work, we focus on empathy in Motivational Interviewing (MI). MI is an evidence-based therapeutic approach for enhancing readiness for behavior change through exploring and resolving ambivalence. MI has been shown to have positive outcomes for multiple disorders, including alcohol use disorder [5]. MI focuses on strengthening personal motivation by eliciting the client's own reasons for change while respecting the client's agency. Empathy is therefore an important pillar in MI. It is crucial for the therapist to empathize and understand the client's reasons and motives to best facilitate behavioral change. Empathy is one of the consistent evaluation metrics for assessing MI sessions' quality. Standardized MI coding systems like the Motivational Interviewing Skill Code 2.5 (MISC) [6] and the Motivational Interviewing Treatment Integrity 3.1 (MITI) [7] both use therapist empathy as an important metric for assessing session quality. However, behavioral coding, i.e., the process of listening to audio recordings to observe therapist behaviors for quality assessment, is highly costly and time-consuming, and therefore, hard to scale. Specifically, obtaining the session quality ratings requires trained third-party coders to review and rate the session following the aforementioned standardized codings, MISC and MITI, on 7-point or 5-point Likert scales, respectively.

The current work focuses on building models that utilize therapist language for estimating the session-level empathy ratings (standardized across datasets). To this aim, we utilize real-world MI therapy datasets for alcohol abuse [8], [9]. Motivated by past work demonstrating the potential relevance of the content of certain temporal segments of MI sessions to outcomes [10], [11], we divide each session into four roughly equal-length sequential segments (quartiles), each representing a different stage of the conversation. We aim to determine whether the language from these segments (often with different topics elicited by the therapists) has different predictive power for session-level empathy estimation. Through our analyses, we show that the language of therapists and clients generally follows a common progression. We conduct multiple experiments to study the importance of the content in each quartile for understanding empathy. We demonstrate that the utterances from the second quartile, which focuses on clients' activities and experiences around alcohol, may be more predictive of session-level empathy.

The main contributions of this paper are as follows.

- We propose and evaluate a regression model for estimating therapist empathy using spoken language. We demonstrate that language encoders pre-trained for emotion recognition provide better results compared with general purpose encoders, demonstrating the significance of affect in therapist empathy.

- We investigate and analyze the therapist and client language across the session quartiles, showing that certain quartiles (i.e., discussion of client activities around the problematic behavior) are more predictive of empathy overall.

II. Background

In MI, the therapist is focused on promoting behavior change by encouraging clients to verbalize their desire for change (change talk). MI theory posits that change talk should have a linear, positive slope over the course of the session as the therapist selectively evokes and reflects change talk [12], [13]. This has been examined by dividing the session into equal parts, though how the session is divided has differed in a variety of studies. In the first systematic examination of client change language over the course of an MI session, Armhein et al. [14] divide MI sessions into ten deciles (each being 1/10 of the session's length), and this approach has been replicated in later work [15]–[18]. These studies also showed that the amount of change talk increased quadratically over the session, suggesting that fewer divisions can capture the same fluctuations of client change language, with later work using session quintiles (1/5 of the session's length) [10].

Quality assessment of therapy sessions can provide valuable insights into how a competent therapist operates, and what kind of therapist-client interactions are productive. Researchers have explored approaches to build automatic systems for quality assessment, for different types of therapy. For example, Xiao et al. [19] trained a model to predict empathy levels (high vs. low) using n-gram language model features from manual and automatically recognized speech (ASR) transcripts in MI sessions, with encouraging results regarding human rating correlation (0.65). In [20], the authors extended these approaches, by integrating the language model features into a Hidden Markov Model (HMM) in order to capture the dynamic interactions between utterances in the MI sessions. They show that the dynamic model improved on the accuracy of empathy level prediction compared to a static model as in [21]. To leverage the semantic aspects beyond word counts (n-grams), researchers have also used Linguistic Inquiry and Word Count (LIWC) [22] features to predict empathy levels. Lord et al. [23] used LIWC features to compute language style synchrony between clients and therapists, finding that higher empathy ratings are correlated with higher synchrony, controlling for therapist reflections. Similarly, Gibson et al. [24] found that these psychologically-motivated LIWC features carry complementary information to standard n-gram features in predicting therapist empathy. This is likely because LIWC features were also found useful in distinguishing change vs. sustain talk in client language [25]–[27].

Researchers have also attempted to automatically code session behavior (MI codes) as an intermediate step for predicting session-level quality metrics. For example, Can et al. [28] used Conditional Random Fields (CRF) in a sequence tagging framework to predict MI codes based on speech, which are then used in estimating session quality measures like empathy. Leveraging the advances in neural network models, a series of more recent work have focused on using word embeddings from non-contextualized representations such as GloVe [29] and word2vec [30], to newer large contextualized models such as BERT [31], for natural language understanding. For example, Gibson et al. [32] modeled MI sessions using a recurrent neural network (RNN) applied to word2vec embeddings in order to obtain utterance-level representations, which are then used to predict empathy levels. In [33], the authors used GLoVe embeddings as well as LIWC features to estimate Cognitive Behavioral Therapy (CBT) session quality as measured by the Cognitive Therapy Rating Scale (CTRS) scores, finding that therapist-related language features have more predictive power than client language. In their follow-up study, Flemotomos et al. [34] expanded their analyses by incorporating highly contextualized representations, i.e., by using BERT-based embedding in their classifiers and achieving consistent performance improvements for session quality assessment over simple n-gram features.

All these empirical studies use the entire therapy sessions for predicting therapy quality metrics. However, clinicians are often interested in the behaviors throughout the session, as it relates to the topics of discussion to facilitate a more fine-grained understanding. Our work addresses this limitation by investigating how language from the temporal segments can be used to estimate empathy. This is done through training a regression model using expert-annotated empathy scores for MI therapy sessions as ground-truth labels.

III. Dataset

In this work, we leverage two clinical datasets of real-world Motivational Interviewing sessions. Our datasets come from MI sessions from two populations: 1) College students (ages 18-23) mandated to take part in MI sessions due to alcohol-related problems [8] and 2) Community-based underage (ages 17-20) heavy drinkers transitioning out of high school who were not immediately planning to enroll in a 4-year college. These participants were non-treatment-seeking volunteers recruited via advertisement and recruitment events held at local high schools, community colleges, etc. [9]. Both populations underwent single MI sessions, which were delivered as face-to-face meetings that take approximately 50-60 minutes and include personalized feedback to promote less risky drinking.

The first dataset contains 219 MI sessions with mandated college students. The sessions include audio files and manual transcriptions. They are coded following the MISC 2.5 guidelines for local utterance-level behaviors, as well as global ratings of empathy and other MI-related measures like therapists' acceptance and MI spirit. 20% of the sessions were randomly selected and double-coded to verify inter-rater reliability. Intraclass correlation coefficients (ICCs; two-way mixed, single measure) were calculated for each variable to determine inter-rater reliability across rater pairs [35]. For this dataset, the ICC scores for therapists' global measures range from 0.47 to 0.78, which is considered "fair" to "excellent" [36].

The second dataset comprises of 81 MI sessions with community-based underage drinkers consisting only of audio files. We used the Google Automatic Speech Recognition (ASR) service for the automatic transcription of the sessions. We manually verified the ASR quality of a subset of sessions and found the transcriptions have a few issues that are minor enough not to affect this work (e.g. missing or inserting disfluencies such as 'uh's and 'um's). The sessions are annotated with utterance-level codes, as well as global ratings of therapist skills like empathy and acceptance following the MITI 3.1 coding system. Similar to the first dataset, 20% of the sessions were randomly selected for double-coding and ICC scores (two-way mixed, single measure) were computed [37]. ICC for this dataset was 0.83, which is considered "excellent" [36].

Since both datasets follow different coding systems on different Likert scales (5- and 7-point), we scale the empathy ratings between 0–1. Fig. 1 demonstrates the histogram across the datasets, and dataset statistics are presented in Table I.

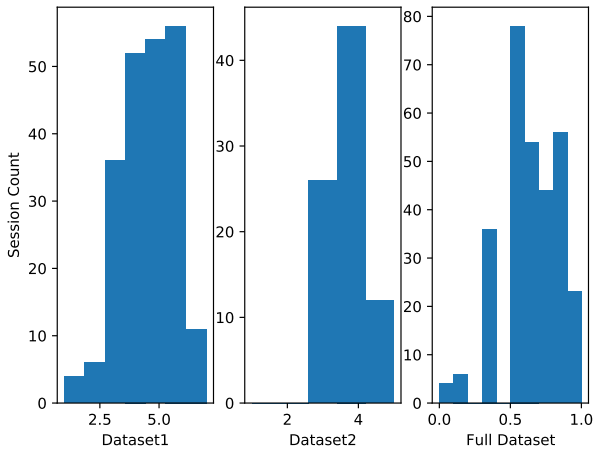

Fig. 1. Histograms of empathy ratings across our two datasets (left, center), and the combination of both datasets (right) with scaled empathy scores between [0-1].

TABLE I
DATASET STATISTICS: SESSION LENGTH (IN MINUTES) AND AVERAGE NUMBER OF TURNS (STANDARD DEVIATION IN PARENTHESES)

	# sess	avg. length (min)	avg. # turns
Dataset1	219	49.9 (13.7)	422 (128.2)
Dataset2	82	54.3 (12.1)	600 (138.5)

IV. Approach

A. Session Segmentation

In this work, we focus on using the therapist's spoken language for estimating the empathy ratings with a regression model. MI therapists are trained to follow a common (but flexible) structure, which inspired our approach to studying language patterns at the quartile level. A typical session roughly progresses as follows: In the first quartile (Q1) the therapist asks about the client's drinking habits and discussing how alcohol fits into their life. The client provides information on their overall patterns of drinking such as number of drinks on a typical drinking day, average number of drinking days per week etc. In the second quartile (Q2), the client discusses the activities they partake in while drinking. These activities mostly revolve around social activities such as parties and drinking games. The therapist provides insights into some of the physiological effects of such drinking behaviors. The third quartile (Q3) focuses on personalized feedback, where the therapist usually provides statistics and quantitative assessment of the drinking behaviors of the client, e.g. how the client compares to other people of their age. In the final quartile (Q4), the client and therapist discuss potential actions corresponding to their plan for change. We provide a preliminary analysis of the common language within session quartiles for more insights into the data. We use class-based tf-idf (Term Frequency - Inverse Document Frequency) analysis to obtain top n-grams per quartile, by exploring individual quartiles combined across all sessions to represent documents. Table II shows the most frequent bi-grams from session quartiles, showing an overall pattern for the session progression.

The decision to divide the sessions into quartiles has multiple practical advantages. First, the quartiles can be mapped onto the format of the sessions of interest into four structural segments described above. Second, a quarter of a session is more interpretable for clinician training and supervision implications than considering a larger (whole session) or smaller (decile) part of a session. Finally, analyzing session language at the quartile level would alleviate the loss of information associated with limitations in a neural network model's context or input sequence length when modeling the entire session.

Since we do not have access to precise annotations of the start or end points of each stage in the sessions (in addition to the fact that the session structure might be more fluid and flexible), we divide the session into four quartiles by time to access utterances within each quartile. We study the therapist language within the quartiles to explain the progression and investigate the importance of each quartile in the estimation of perceived therapist empathy.

B. Model

In this work, we investigate the therapist's in-session language with respect to its perceived empathy level. For encoding the therapist utterances, we leverage the recent advancements on language representation by fine-tuning the pretrained model distil-RoBERTa (distilled Robustly Optimized BERT Pretraining Approach) [38]. We obtain the language representations from the input window and feed the sequence to an initial linear layer for dimensionality reduction, followed by single-layer bidirectional Gated Recurrent Units (GRU). We take the output from the entire sequence and feed them into a multi-head self-attention layer [39] for learning the

TABLE II
MOST COMMON BI-GRAMS ACROSS SPEAKERS AND QUARTILES
*BAC: BLOOD ALCOHOL CONTENT

	Dataset1		Dataset2	
	therapist	client	therapist	client
Q1	heaviest week, questions started, drinks single, single day, maximum number	Thursday Friday, week yeah, Saturday like, sounds right, number drinks	love just, adult make, really opens, related alcohol, life young, hopes dreams	community college, looking job, just hanging, make money, going school
Q2	slurred speech, particular bacs*, outer brain, tolerance heard, emotional center, people associate	slurred speech, flip cup, self conscious, reaction time, drinking game	regret getting, trouble having, standard drink, age 21, measure alcohol	high tolerance, shit faced, funny like, high number, binge drinking
Q3	went estimation, called perceived, thought average, people overestimate, myth alcohol	talk people, easier talk, bad time, nights week, neglected responsibilities	later regretted, alcohol dependence, spend drinking, percent money, lead dangerous	alcohol level, like woke, spend time, heavy drinker, family history
Q4	seal envelope, related problems, alcohol problems, drinkers alcohol, severe consequences	drink unattended, designated driver, good idea, avoid drinking, leave drink	lot today, make hard, really appreciate, keeping track, complete stranger	drinking games, peer pressure, avoid drinking, need help, drug use

relative importance per utterance within the input window. The weighted sequence representations are aggregated into a final representation of the input window by concatenating the mean- and max-pooled hidden states of the entire sequence. These learned representations are passed through a final linear layer for regression.[1]

Using this network architecture, depicted in Fig. 2, we build a regression model for learning continuous empathy ratings for session quartiles. We further aggregate these quartile-level estimations to obtain the session-level empathy, by taking the average across quartiles. In addition to using a base encoder, motivated by the affective nature of empathy, which involves identification with others' emotional experiences, we also train our models with a distil-RoBERTa-emotion encoder [40], which is pre-trained on multiple emotion datasets [41]–[45].

C. Experimental Setup

In this work, we train and evaluate our method on a combined dataset of 301 real-world MI sessions. We extract a fixed window size with the first 64 therapist turns. The choice of window size was based on the average length of session quartiles in terms of the number of therapists' speech turns and the hardware constraints. We use distil-RoBERTa [38] or emotion-distil-RoBERTa [40] encoders for our text representation while fine-tuning the final layer for our task. The dimension of the input vector embeddings is 768, and the hidden dimensions for the initial linear layer and GRU are 512 and 256, respectively, and the dropout rate after self-attention is set to 0.5.

We use 5-fold cross-validation for training and evaluation of the overall dataset, and report the test results by selecting the model with the highest validation performance. We perform both therapist-dependent and therapist-independent cross-validation. In the therapist-dependent cross-validation, the splits are not disjoint by the therapist, meaning sessions from the same therapist can appear in both train and test sets. On the other hand, in therapist-independent cross-validation, sessions from the same therapist do not appear in both training and testing data, preventing the model from learning any therapist-specific patterns or idiosyncrasies. We optimize the network weights using AdamW, with a batch size of 8 and a learning rate of $5e^{-5}$. The small batch size is due to the memory constraints on GPUs given the large input size for each sample. We use the Concordance Correlation Coefficient (CCC), a widely used evaluation metric for regression that measures agreement between the prediction and ground-truth values. Past work demonstrated that using CCC loss results in superior performance in emotion recognition [46]. Therefore, we opt to use a CCC loss rather than a more commonly used Mean Squared Error (MSE) loss.

[1]The code for this model is publicly available at https://github.com/ihp-lab/empathy-recognition-acii-2023

V. RESULTS AND DISCUSSION

A. Base Models vs. Emotion Models

In our first set of experiments, we compare two different encoders for language representations, namely, distil-RoBERTa [38] and distil-emotion-RoBERTa [40], to see if an encoder pre-trained on emotion recognition tasks provides performance improvements for estimating empathy as an affective construct. Our previous experiments showed that the distilled model versions are on-par with their full model counterparts (distilled RoBERTa vs. RoBERTa-base) in terms of performance. This may be due to our limited data size, so we focused on distilled encoders across all experiments.

In Table III, we provide the model performance across different session quartiles, under both therapist-dependent and therapist-independent settings. From these results, we can see that the emotion-RoBERTa outperforms the RoBERTa encoder

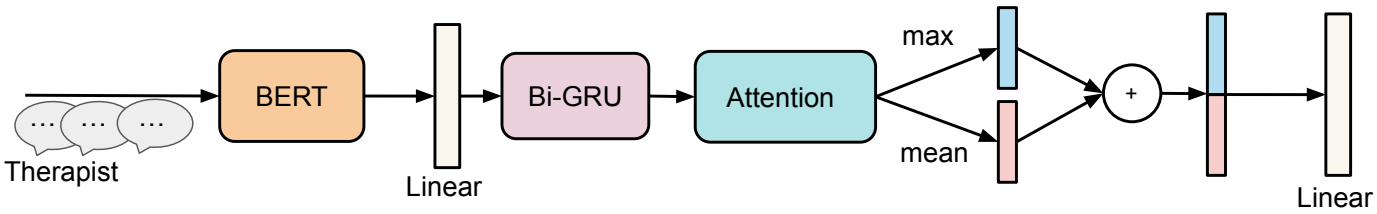

Fig. 2. The model includes an utterance encoder (distil-RoBERTa [38] or emotion-distil-RoBERTa [40]) whose output is projected to a lower-dimensional space by the following linear layer. The sequence of utterance-level representations is then fed to a Bidirectional Gated Recurrent Unit (Bi-GRU) layer. The GRU is followed by a two-head self-attention layer on GRU's hidden states whose output is mean- and max-pooled into the final vector embedding, which is fed to a final linear layer for regression.

on average, leading to higher performance results on the session-level estimations.

The performance gap is more evident in the therapist-independent setting, especially for the first quartile, Q1, (from 0.150 to 0.367) potentially due to affect-related language in that quartile. Comparing the performance across the quartiles, Q2 is consistently more predictive in the therapist-dependent setting, although the pattern is not consistent for the therapist-independent scenario. The session-level performance, obtained by aggregating the predictions across all quartiles, performs best in all but one of the cases, reaching CCC scores of 0.596 and 0.408 for therapist-dependent and independent cases, respectively. As expected, there is a drop in performance in the therapist-independent evaluation setting, since the model can't utilize individual therapist characteristics for recognition. This gap is exacerbated by the real-world nature of our dataset and the large variation in the number of sessions per therapist, which we will further discuss in Section V-C.

TABLE III
PERFORMANCE RESULTS (CCC SCORES) COMPARING DISTIL-ROBERTA VS. EMOTION-DISTIL-ROBERTA ENCODERS, UNDER THERAPIST-INDEPENDENT AND THERAPIST-DEPENDENT SCENARIOS. THE RESULTS INDICATE THE MEAN ACROSS THE CROSS VALIDATION FOLDS, WITH STANDARD DEVIATION IN PARENTHESES.

	therapist-dependent		therapist-independent	
	base	emotion	base	emotion
Q1	0.510 (0.05)	0.488 (0.08)	0.150 (0.09)	0.367 (0.16)
Q2	0.546 (0.03)	0.544 (0.04)	0.264 (0.15)	0.341 (0.12)
Q3	0.436 (0.06)	0.528 (0.08)	0.291 (0.07)	0.310 (0.10)
Q4	0.450 (0.10)	0.470 (0.11)	**0.341 (0.16)**	0.344 (0.17)
sess.	**0.572 (0.04)**	**0.596 (0.06)**	0.320 (0.12)	**0.408 (0.16)**

B. Cross Corpus vs. Within Corpus

In our next set of experiments, we use the emotion encoder due to its superiority, under the therapist-independent configuration for more robust results invariant to individual characteristics. As shown in Table II, the sessions generally follow a similar structure of progression despite some differences in the discussed content. To further analyze this aspect, we explore the overall generalizability of our model predictions across datasets under within-corpus and cross-corpus evaluations.

To this end, we first train and evaluate our model within corpus on our larger dataset (Dataset1). Next, we perform cross-corpus testing by using Dataset1 as the training set and Dataset2 as the validation set with the main goal of identifying whether certain session quartiles are more similar in language. The results are shown in Table IV, with the first column providing the within-corpus results using a 5-fold therapist-independent cross-validation. The remaining columns on the right provide the scores when training on Dataset1 and validating on Dataset2 under different combinations of session quartiles.

TABLE IV
CCC SCORE RESULTS IN WITHIN-CORPUS AND CROSS-CORPUS EXPERIMENTS, THERAPIST-INDEPENDENT SETTING. THE WITHIN-CORPUS RESULTS INCLUDE THE MEAN ACROSS THE CROSS VALIDATION FOLDS, WITH STANDARD DEVIATION IN PARENTHESES. THE CROSS-CORPUS RESULTS ARE OBTAINED FROM TRAINING AND VALIDATION ON THE ENTIRE DATASETS.

Dataset1	Within-corpus	Cross-corpus testing (Dataset2)				
		Q1	Q2	Q3	Q4	sess.
Q1	0.173 (0.18)	0.107	0.178	-0.018	0.067	0.114
Q2	**0.342 (0.18)**	**0.229**	0.188	**0.134**	**0.102**	**0.198**
Q3	0.293 (0.19)	0.037	0.186	-0.007	0.058	0.053
Q4	0.228 (0.27)	0.092	0.191	-0.009	0.069	0.112
sess.	0.299 (0.20)	0.134	**0.194**	0.016	0.090	—

We first compare the within-corpus results, in which the model was trained and tested on Dataset1. Compared to the results on the combined dataset (last column of Table III), the results for Q2 and Q3 are on par across the two experiments, suggesting that these quartiles are most similar in terms of therapist language across datasets. On the other hand, the performance on Q1 and Q4 seems to suffer from significant drops, likely due to the differences in dataset characteristics. Additionally, in the cross-corpus setting, results were consistent when the model was trained on Q2, i.e. the Q2 model transfers well to all test quartiles. Conversely, models trained on any quartile also performed most consistently on Q2 quartiles in test. These findings suggest that there is a language commonality in this specific Q2 quartile across datasets. Recall that Q2 is when the clients are describing their activities around drinking, and the therapist informs them of possible behavioral effects while expressing understanding in a non-judgemental manner, in accordance with MI protocols.

C. Therapist Analysis

One of the challenges of the dataset in the therapist-independent scenario is the high imbalance of the number of sessions across different therapists, ranging from 1 to 44. In Dataset1, we have 219 sessions run by 13 therapists (IDs 1-13), and in Dataset2 we have 82 sessions run by 3 therapists. One session had missing therapist information, so we coded this session as conducted by a separate therapist, i.e., Dataset2 includes therapist IDs 14-17. While the 17 therapists follow the same guidelines across datasets, they have different variances in empathy ratings across sessions. Fig. 3 shows the empathy ratings across different therapists. As described earlier, the two datasets are rated using different Likert scales, which we have scaled to [0,1] for our analysis.

Our research question in this section is: what constitutes a more empathetic therapist, and we explore this question by studying the types of language and MI codes therapists use. MI codes are utterance-level categories following standardized MISC/MITI coding systems. These codes categorize the session utterances into therapist- and client-specific categories. In this analysis, we focus on therapist codes, which include simple/complex reflections, open/closed-ended questions, giving information, facilitation, etc. To this end, we categorize the therapists into groups of high vs. low empathy using their average empathy ratings across sessions. We select a threshold of 0.7, leading to a balanced grouping. We then obtain the normalized usage of each MI code per therapist across different quartiles, by taking the average across sessions. We run Kruskal-Wallis tests across the two groups for different therapist-specific codes including simple/complex reflections, open/closed-ended questions, giving information, etc. Table V demonstrates the significant MI codes that distinguish high vs. low empathy scores across quartiles.

TABLE V
SIGNIFICANT THERAPIST CODES ASSOCIATED WITH EMPATHY; **PVALUE < 0.01; *PVALUE < 0.05; POSITIVE(+) AND NEGATIVE(-) ASSOCIATIONS WITH EMPATHY ARE SHOWN IN PARENTHESES

	Significant therapist codes
Q1	MI-consistent** (+); Open-ended question* (+) Giving information* (-)
Q2	Complex reflection** (+); MI-consistent** (+) MI-inconsistent* (-); Giving information* (-)
Q3	MI-consistent** (+); Complex reflection* (+)
Q4	Complex reflection * (+)

These findings are consistent with what we would expect, based on general MI guidelines. In particular, 'MI-consistent' is a large category that includes codes like 'advice with permission,' 'affirm,' 'emphasize control,' and 'support.' The results show that the 'MI-consistent' category is significantly and positively associated with perceived empathy across most quartiles. 'Complex reflection,' which consists of reflections that add substantial meaning or emphasis to what the client had said, is also significantly and positively associated with therapist empathy. 'Complex reflection,' which can be a good indicator of empathetic understanding, has a stronger association in Q2 (p-value < 0.01), where the client is describing their activities and experiences around drinking alcohol. This further supports why Q2 is a prominent quartile for the estimation of session-level empathy in the majority of our experiments. This result also suggests that effective therapists exhibit deep and empathetic understanding around discussion of alcohol-related activities and experiences.

TABLE VI
SAMPLE DIALOGUE EXCERPTS FROM THE DATASETS INCLUDING CORRESPONDING MI CODES. **T** DENOTES THE THERAPIST AND **C** DENOTES THE CLIENT.

T: ... you can see that on a typical occasion and on the heavier occasion you are getting above that **(Giving information)** **C**: oh yeah i wouldn't i wouldn't consider driving after more than like two drinks basically ... **(Change talk)** **T**: okay so it sounds like the avoiding the driving is pretty important to you **(Complex reflection)**
T: ... that's definitely not a positive effect of drinking you know balance and movement are affected **(Change talk)** **C**: so it sounds like for you that would be really difficult especially if you know you tend to try to stay in control of yourself and situations and having that impairment **(Complex reflection)**
C: yeah i was drinking every day for like four years right and **(Follow/neutral)** **T**: ... it is less for sure and you are functioning so it makes sense that you know you wouldn't think that the number would be as high as it is but when you look at ... **(MI-consistent)**

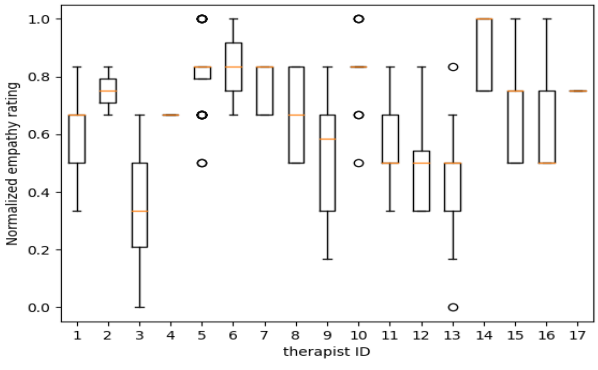

Fig. 3. Empathy ratings across different therapists. Therapist IDs 1-13 are from Dataset1; Therapist IDs 14-17 are from Dataset2.

Other statistically significant codes include 'MI-

inconsistent' and 'Giving information', which are both negatively associated with empathy. 'MI-inconsistent' consists of those actions directly proscribed by MI guidelines (such as giving advice without permission, confronting, directing), which shows to be detrimental to perceived therapist empathy. 'Giving information,' which is the category for when the therapist explains something, educates or provides feedback, is also negatively associated with empathy. This suggests that these types of speech should be used in moderation by therapists. Some example dialogue excerpts are shown in Table VI, noted with the types of MI codes associated with each utterance.

VI. Conclusion

In this paper, we presented a method for computational understanding of therapist empathy in MI therapy sessions. To this end, we utilized therapist language from individual quartiles within the sessions. We developed and evaluated a neural network architecture for estimating empathy ratings per session using therapist language. We conducted experiments and analyses within and across datasets to gather insights into the importance of each quartile within the session progression. Our results indicate that the second quartile of the session, which is commonly focused on discussing clients' experiences around drinking alcohol, may be of higher importance for estimating therapist empathy. We achieved promising results for estimating empathy using pre-trained language encoders with affect-aware representations. Moreover, our analyses show that therapists with higher empathy ratings tend to provide more complex reflections, which are most significant during the second quartiles. This finding provides evidence for the importance of complex reflections for therapist empathy, through demonstrating a deeper understanding of the client by the therapist.

Therapist empathy is key to successful therapy. Modeling and understanding empathy requires effective and efficient means, in order to facilitate therapist training and therapy quality assessment. With this work, we provide evidence for the salience of certain topics (experiences around problematic behaviors) and therapy techniques (reflection), where empathy is most effectively modeled and potentially where it is most important for building an empathetic alliance.

Ethical Impact

The datasets and labels are the result of a secondary analysis from past studies, which were reviewed by their relevant IRB (see the original studies [8], [9].) The original data were recorded with informed consent from the participating clients to be used for research. The original studies allowed for secondary analysis of the audio recordings for training and research purposes, in accordance with the goals of the original study and the participants' consent. In the original studies, the audio data were reviewed and cleared of any identifiable information, such as names and addresses. Therefore, the research presented in this paper was deemed IRB-exempt analysis of secondary data by the USC IRB, which designated the data non-identifiable. Nevertheless, we ensured that data was always transferred and stored by encrypted and password-protected means. When speech data from the second dataset were transcribed by the Google Cloud speech-to-text service, data were transferred through encrypted connections and were only kept in the protected cloud storage for the minimum necessary duration to complete the transcription. We also did not allow the cloud service provider to log the data during transcription for additional protection. We chose Google Cloud Platform due to its superior performance in transcription and its reputation and ability to provide secure and compliant services for storage and analyses of sensitive data. Specifically, Google uses encryption in receiving speech data; It does not claim ownership over the speech data and the resulting transcripts, and does not store or reveal the information when logging is not enabled by the user.[2]

The goals of this study and its developed tools are to provide efficient and scalable means for assisting clinicians in psychotherapy quality assessment and clinical training. This work is in the research phase and future deployment in the real-world setting would be subject to additional experiments for validation through trials.

Acknowledgment

This work was supported by NIAAA grants R01 AA027225, R01 AA017427 and R01 AA12518. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAAA, NIH, Dept. of Veterans Affairs, or the US Government.

References

[1] C. R. Rogers, "The necessary and sufficient conditions of therapeutic personality change." *Journal of consulting psychology*, vol. 21, no. 2, 1957.

[2] A. C. Bohart and L. S. Greenberg, *Empathy and psychotherapy: An introductory overview.* American Psychological Association, 1997.

[3] C. Truax, "Research on certain therapist interpersonal skill in relation to process and outcome," *Handbook of Psychotherapy and Befavioral Change*, 1971.

[4] R. Scheeffer, "Toward effective counseling and psychotherapy," *Arquivos Brasileiros de Psicologia Aplicada*, vol. 23, no. 1, pp. 151–152, 1971.

[5] S. Rubak, A. Sandbæk, T. Lauritzen, and B. Christensen, "Motivational interviewing: a systematic review and meta-analysis," *British journal of general practice*, vol. 55, no. 513, pp. 305–312, 2005.

[6] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc)," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[7] T. Moyers, T. Martin, J. Manuel, W. Miller, and D. Ernst, "Revised global scales: Motivational interviewing treatment integrity 3.1. 1 (miti 3.1. 1)," *Unpublished manuscript, University of New Mexico, Albuquerque, NM*, 2010.

[8] B. Borsari, J. T. Hustad, N. R. Mastroleo, T. O. Tevyaw, N. P. Barnett, C. W. Kahler, E. E. Short, and P. M. Monti, "Addressing alcohol use and problems in mandated college students: a randomized clinical trial using stepped care." *Journal of consulting and clinical psychology*, vol. 80, no. 6, 2012.

[9] S. M. Colby, L. Orchowski, M. Magill, J. G. Murphy, L. A. Brazil, T. R. Apodaca, C. W. Kahler, and N. P. Barnett, "Brief motivational intervention for underage young adult drinkers: Results from a randomized clinical trial," *Alcoholism: clinical and experimental research*, vol. 42, no. 7, pp. 1342–1351, 2018.

[2]https://cloud.google.com/speech-to-text/docs/data-usage-faq

[10] J. Houck, S. Hunter, J. Benson, L. Cochrum, L. Rowell, and E. D'Amico, "Temporal variation in facilitator and client behavior during group motivational interviewing sessions," *Psychology of Addictive Behaviors*, vol. 29, no. 4, pp. 941–949, 2015.

[11] J. Fokas, Kathryn and Houck and B. McCrady, "Inside Alcohol Behavioral Couple Therapy (ABCT): In-session speech trajectories and drinking outcomes," *Journal of Substance Use & Addiction Treatment (JSAT)*, vol. 118, 2020.

[12] W. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford Press, 2013.

[13] W. Miller and G. Rose, "Toward a theory of motivational interviewing," *American Psychologist*, vol. 64, 2009.

[14] P. Amrhein, W. Miller, C. Yahne, M. Palmer, and L. Fulcher, "Client commitment language during motivational interviewing predicts drug use outcomes," *Journal of Consulting and Clinical Psychology*, vol. 71, 2003.

[15] E. Aharonovich, P. Amrhein, A. Bisaga, E. Nunes, and D. Hasin, "Cognition, commitment language, and behavioral change among cocaine-dependent patients," *Psychology of Addictive Behaviors*, vol. 23, 2008.

[16] T. Apodaca, M. Magill, R. Longabaugh, K. Jackson, and P. Monti, "Effect of a Significant Other on Client Change Talk in Motivational Interviewing," *Journal of Consulting and Clinical Psychology*, vol. 81, 2012.

[17] J. Morgenstern, A. Kuerbis, P. Amrhein, L. Hail, K. Lynch, and J. McKay, "Motivational Interviewing: A Pilot Test of Active Ingredients and Mechanisms of Change," *Psychology of Addictive Behaviors*, vol. 26, 2012.

[18] B. Borsari, T. Apodaca, K. Jackson, A. Fernandez, N. Mastroleo, M. Magill, N. Barnett, and K. Carey, "Trajectories of in-session change language in brief motivational interventions with mandated college students," *Journal of consulting and clinical psychology*, vol. 86, 2018.

[19] B. Xiao, D. Can, P. Georgiou, D. Atkins, and S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[20] S. N. Chakravarthula, B. Xiao, Z. E. Imel, D. C. Atkins, and P. G. Georgiou, "Assessing empathy using static and dynamic behavior models based on therapist's language in addiction counseling," in *Interspeech*, 2015.

[21] B. Xiao, Z. Imel, P. Georgiou, D. Atkins, and S. Narayanan, ""Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PloS one*, vol. 10, no. 12, 2015.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[23] S. Lord, E. Sheng, Z. Imel, J. Baer, and D. Atkins, "More Than Reflections: Empathy in Motivational Interviewing Includes Language Style Synchrony Between Therapist and Client," *Behavior Therapy*, vol. 46, 11 2015.

[24] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *Interspeech*, 2015.

[25] L. Tavabi, T. Tran, K. Stefanov, B. Borsari, J. Woolley, S. Scherer, and M. Soleymani, "Analysis of behavior classification in motivational interviewing," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Online: Association for Computational Linguistics, Jun. 2021, pp. 110–115.

[26] C. Aswamenakul, L. Liu, K. B. Carey, J. Woolley, S. Scherer, and B. Borsari, "Multimodal analysis of client behavioral change coding in motivational interviewing," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 356–360.

[27] L. Tavabi, K. Stefanov, L. Zhang, B. Borsari, J. D. Woolley, S. Scherer, and M. Soleymani, "Multimodal automatic coding of client behavior in motivational interviewing," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 406–413.

[28] D. Can, D. C. Atkins, and S. S. Narayanan, "A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[29] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[32] J. Gibson, D. Can, B. Xiao, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. Narayanan, "A deep learning approach to modeling empathy in addiction counseling," *Commitment*, vol. 111, no. 2016, p. 21, 2016.

[33] N. Flemotomos, V. R. Martinez, J. Gibson, D. C. Atkins, T. A. Creed, and S. S. Narayanan, "Language features for automated evaluation of cognitive behavior psychotherapy sessions." in *Interspeech*, 2018, pp. 1908–1912.

[34] N. Flemotomos, V. R. Martinez, Z. Chen, T. A. Creed, D. C. Atkins, and S. Narayanan, "Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations," *PloS one*, vol. 16, no. 10, p. e0258639, 2021.

[35] B. Borsari, T. R. Apodaca, K. M. Jackson, N. R. Mastroleo, M. Magill, N. P. Barnett, and K. B. Carey, "In-session processes of brief motivational interventions in two trials with mandated college students," *Journal of consulting and clinical psychology*, vol. 83, no. 1, pp. 56–67, February 2015.

[36] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." *Psychological assessment*, vol. 6, no. 4, p. 284, 1994.

[37] M. Magill, T. Janssen, N. R. Mastroleo, A. Hoadley, J. Walthers, N. P. Barnett, and S. M. Colby, "Motivational interviewing technical process and moderated relational process with underage young adult heavy drinkers," *Psychology of Addictive Behaviors*, vol. 33, p. 128–138, 2019.

[38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[40] J. Hartmann, "Emotion english distilroberta-base," online, 2022. [Online]. Available: \url{https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/}

[41] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4040–4054.

[42] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 1–17.

[43] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536.

[44] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning." *Journal of personality and social psychology*, vol. 66, no. 2, p. 310, 1994.

[45] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3687–3697.

[46] D. Le, Z. Aldeneh, and E. Mower Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Proc. Interspeech 2017*, 2017, pp. 1108–1112.