Psychotherapy is Not One Thing: Simultaneous Modeling of Different Therapeutic Approaches

Maitrey Mehta[1], **Derek D. Caperton**[2], **Katherine Axford**[3], **Lauren Weitzman**[4], **David Atkins**[5], **Vivek Srikumar**[1], **Zac E. Imel**[3]
[1]School of Computing, University of Utah
[2]Calgary Counselling Centre
[3]Department of Educational Psychology, University of Utah
[4]University Counseling Center, University of Utah
[5]University of Washington

Abstract

There are many different forms of psychotherapy. Itemized inventories of psychotherapeutic interventions provide a mechanism for evaluating the quality of care received by clients and for conducting research on how psychotherapy helps. However, evaluations such as these are slow, expensive, and are rarely used outside of well-funded research studies. Natural language processing research has progressed to allow automating such tasks. Yet, NLP work in this area has been restricted to evaluating a single approach to treatment, when prior research indicates therapists used a wide variety of interventions with their clients, often in the same session. In this paper, we frame this scenario as a multi-label classification task, and develop a group of models aimed at predicting a wide variety of therapist talk-turn level orientations. Our models achieve F1 macro scores of 0.5, with the class F1 ranging from 0.36 to 0.67. We present analyses which offer insights into the capability of such models to capture psychotherapy approaches, and which may complement human judgment.

1 Introduction

A typical psychotherapy session involves a client–therapist dialog with the aim of diagnosing and assuaging a client's mental health condition. Psychotherapists, generally, rely on certain approaches (e.g., Cognitive Behavioral or Interpersonal Therapy) and interventions differ across these approaches.[1]. For example, a therapist might focus on a client's interpersonal relationships, their emotions, or help develop behavioral activities designed to reduce symptoms (or all of the above). A key goal of psychotherapy research is to categorize such approaches and study them to determine the effectiveness of each approach in any given scenario. We refer to this process of categorizing and detecting approaches based on an overarching theory as 'evaluation'.

In this paper, we study an application of Natural Language Processing (NLP) to mental health, and focus on therapists' approach to psychotherapy (Imel et al., 2015). Past NLP research has developed tools for evaluating specific types of interventions like Motivational Interviewing (Cao et al., 2019) or Cognitive Behavioral therapy (Flemotomos et al., 2021). However, psychotherapists differ from each other in the approaches they take. Furthermore, they can also vary in the interventions they use within and between sessions. The lines of work mentioned before assume that a session is comprised of exactly one approach, and consequently do not attempt to automatically evaluate different psychotherapy approaches that may coexist in the same session.

McCarthy and Barber (2009) proposed one multiple-approach evaluation methodology—the *Multitheoretical List of Therapeutic Interventions* (*MULTI*), which is a list of 60 interventions (or, items) against which a psychotherapy session as a whole is evaluated post-session. The *MULTI* items are grouped into eight approaches. Note the *MULTI* is a session-level measure and thereby limited in specificity because it does not record therapist language that informs a given item's presence. Caperton (2021) extend the scheme to the evaluation of therapist monologues, talk-turn by talk-turn, in addition to the session-level evaluation. Such a scheme provides additional detail over time in a session.

Evaluating sessions with the *MULTI* requires a certain amount of time to be set aside post-session. Evaluating talk-turns manually for every session would be even more onerous and inefficient. This calls for a better automatic/semi-automatic method(s) to evaluate talk-turns. These methods serve two advantages: i) reducing the amount of

[1]We use the words 'approach' and 'orientation' interchangeably. Later in this paper, we use 'subscales' to align with practical usage.

effort required in manual classification for research and quality assurance, and ii) creating applications to analyze approaches deemed helpful on out-of-session platforms (e.g., social media).

To that end, we present a neural machine learning model which aims to automate talk-turn level approach annotation. The task is set up in the following fashion: Given a therapist input talk-turn, does the input (or part of the input) correspond to one or more approaches. A talk-turn might only represent one approach, or might have different parts that correspond to different approaches. It is also possible that a therapist talk-turn does not fall within a specific therapeutic approach (e.g., minimal encouragers, small talk, etc.). Examples are shown in Table 1. This problem posits itself perfectly as a multi-label classification task.

The state-of-the-art in natural language processing (NLP) has seen significant improvements with the advent of transformer-based models (Vaswani et al., 2017; Devlin et al., 2019). In this paper, we show the performance of one such pre-trained transformer based language model on three paradigms, and experiment with changing context windows. Our models achieve around 0.5 F1 macro scores with the class F1 ranging from 0.36 to 0.67. Our analyses reveal that while our models mispredict on certain talk-turns during a session, they capture the dominant approaches when viewed from a session-level perspective. Furthermore, we show that certain decisions rely on inter-session context, and even common-sense knowledge which sets up a challenge for current models.

2 Talk-turn Level MULTI-30 Coding

MULTI-60 and MULTI-30. The *Multitheoretical List of Therapeutic Interventions* (*MULTI*) was originally developed as a list of 60 interventions (McCarthy and Barber, 2009). The 60-items belonged to eight different coarse-grained *subscales*, each representing a therapeutic approach. Each item was rated on a 5-point Likert scale for how prevalent the intervention was over the course of a psychotherapy session. The *MULTI*-60 was later re-evaluated through an item reduction procedure to create the more parsimonious *MULTI*-30 (Solomonov et al., 2019), comprised of the same eight subscales. In this work, we use focus on the eight coarse-grained approaches.

Each subscale was defined by a psychotherapeutic theoretical orientation. We describe each subscale briefly here.

1. *Psychodynamic* (**PD**) items focus on addressing nonconscious content from the client's psyche to alleviate distress.

2. *Process-experiential* (**PE**) items emphasize what is happening in the moment during a therapy session with the understanding that what happens in-session mirrors processes in the client's life outside of session.

3. *Interpersonal* (**IP**) items focus on relationship issues with other people in the client's life.

4. *Person-centered* (**PC**) interventions focus on elucidating client experiences and opinions to gain clarity on distress.

5. *Behavioral* (**BT**) items encourage adaptive behavioral activation strategies, assuming that productive actions will produce changes in mental wellbeing.

6. *Cognitive* (**CT**) items address possible distortions or unhelpful patterns in client thinking.

7. *Dialectical-behavioral* (**DBT**) interventions emphasize the client's non-judgment of present experience and the balance between accepting themselves as they are while believing they can be better.

8. *Common factors* (**CF**) items are purportedly transtheoretical and include interventions where the therapist demonstrates encouraging, sympathetic, and attentive listening behaviors.

Data Source. Psychotherapy audio data was collected from a university counseling center at large public school in the western United States. There were 243 unique sessions transcribed, some of which were annotated more than once, totaling to 473 sessions. These sessions were annotated using a talk-turn level version of the MULTI-30 (Caperton, 2021).

Coding Procedure and Reliability. Seven graduate students in mental health fields annotated session content for their varying use of theoretical interventions. Each coder received approximately 18 hours of training during in-person meetings and practiced coding sessions for an additional 36 hours before annotating session data used in this study. To minimize coder drift over time,

Case	Example Talk-turns	Approach(es)
Non-Approach	Okay let's set you up with an appointment.	No Code
	So, All of us trainees get to have a break as well.	No Code
Single Approach	You're scared.	Process-Experiential
	I definitely notice a lot of progress that you've made.	Common Factors
Multiple Approaches	Unfortunately, it's very normal. But I want you to continue practicing that exercise.	Common Factors Behavioral
	When you say that he's better off without you, what do you mean by that? It seems like he still has you.	Person-Centered, Cognitive

Table 1: Examples of talk-turns which have a single, multiple or no approach categories assigned. In the Multiple Approaches examples, colored text snippets correspond to their respective approach categories with the same color.

coders met together with their team leader every two weeks to discuss difficult talk-turns, items, and areas of disagreement.

Coders were tasked with identifying the presence or absence of theory-derived content in therapists' language at every therapist talk-turn (i.e., a string of words or statements uninterrupted by client speech). A given talk-turn could be identified with one, multiple, or no interventions.

Of the 243 unique sessions, 102 were annotated by multiple coders, resulting in 270 codings for interrater analysis. The statement-level interrater reliability of the eight theoretical orientations (subscales) was calculated using Cohen's kappa. Kappa was calculated for every possible coder pair who rated the same session and weighted according to the number of comparisons. Subscale kappa scores ranged from .37 ('fair' reliability; Landis and Koch (1977)) to .63 ('substantial').

The dataset was split by client randomly into train/dev/test sets containing 70%, 15% and 15% of the clients respectively. The splits contain 338, 66, and 76 sessions respectively containing 74k, 14k, and 17k talk-turns in total. Dataset statistics for the training split are presented in Table 2.

Class Name	Counts
No Code	58584
Psychodynamic	1024
Process-Experiential	3865
Interpersonal	1446
Person-centered	4810
Common Factors	5931
Behavioral	1531
Cognitive	1940
Dialectical-Behavioral	1765

Table 2: Training Data Statistics

3 Models

While we want to model the eight subscales (plus the 'No Code' class) conventionally used in literature, we deviate from these eight classes for the implementation. The Behavioral, Cognitive, and Dialectical-Behavioral subscales contain overlapping items (e.g., items 1 and 10 are shared by all three subscales). We break these subscales into four categories such that each of these categories contains mutually exclusive items. Note that the other subscales(Psychodynamic, Interpersonal, etc.) remain the same. Hence, in total, we obtain ten modified model classes (including the 'No Code' class). We refer the reader to Tables 7 and 8 in Appendix A for further details on the breakdown. This method can aid downstream analysis by allowing credit/blame assessment on a smaller set of items.

In our setup, each therapist talk-turn u_i has a corresponding binary label vector y_i. The binary label vector is ten dimensional, one decision each for the nine model classes (i.e., modified subscales) and one additional class indicating the absence of any code (NC). For all our experiments, we consider the RoBERTa-base (Liu et al., 2019b) model as the language model of choice. This model takes in a talk-turn u_i as input to produce contextual representations for its words. We take the pooler output of these contextual representations which gives us

a vector representation h_i for the talk-turn.

$$h_i = Pooler(RoBERTa(u_i)) \quad (1)$$

We consider three modeling paradigms for our experiments.

Stand-Alone (SA) Model. This model is the vanilla multi-label classifier. Talk-turn representations are passed through a linear layer with the number of output nodes equal to the model classes. The result is passed through a sigmoid layer resulting in a vector of presence probabilities for each label $\hat{y_i}$. That is

$$\hat{y_i} = \sigma(w^T h_i + b) \quad (2)$$

where w and b are the weights of the linear layer. For inference, a probability of 0.5 or above indicates label presence for a particular class.

Pipeline Model. A heavily imbalanced dataset can hinder model performance for the under-represented categories. As seen in Table 2, the number of examples with a "No Code" class highly skews the dataset, potentially leading to performance bias towards the class. To alleviate this problem, we define a pipeline model that uses a separate binary classifier to determine whether a talk-turn deserves an orientation category or not. If this binary classifier predicts that the talk-turn supports at least one orientation, then the talk-turn is given to a multi-label model to predict over the nine model classes. The multi-label model will be similar to the one mentioned in the Stand-Alone Model, except nine classes are considered since predicting a "No Code" would be redundant. The multi-label model has the flexibility, nonetheless, to predict an absence of orientation by predicting that none of the codes are present (i.e., a zero vector). Note that two separate RoBERTa models are used for the binary and the multi-label classifiers.

Multi-Task Model. The Pipeline Model trains two separate RoBERTa models — one for the binary classifer and one for the multi-label model. A major drawback of this system is that training two RoBERTa models is computationally expensive and memory-intensive. An alternative method is to share the RoBERTa layer between the two tasks and have two separate linear layers for the respective binary and multi-label classification. This strategy of multi-task or joint learning has shown to be of promise in literature (Liu et al., 2019a; Stickland and Murray, 2019) and allows for better shared representation. The losses for both the tasks are combined as a weighted sum for learning. We consider two variants of the model based on the number of output classes for the multi-label classifier. The **MultiTask**$_{10}$ variant considers all the classes including "No Code" while **MultiTask**$_9$ excludes the "No Code" class. The inference is identical to the Pipeline Model. The Multi-Task and Stand-Alone model paradigm can be thought of as fairly similar architectures. However, the Multi-Task model assigns a higher loss weight to the binary classifier, uses a different optimization metric and utilizes a pipelined inference approach as opposed to the one-shot prediction by the Stand-Alone model.

So far, we explained that we break the conventional eight subscales into nine which have mutually exclusive items. While this approach allows us for better analysis, it is essential to present performance on the original theoretical subscales. To that end, we aggregate binary vector model predictions to the conventional eight *MULTI* subscales during evaluation. The output of the model, a ten-dimensional vector, will be mapped to a nine-dimensional vector(eight subscales plus 'No Code'). We use these nine-dimensional vectors to perform model evaluation. Table 8 is a guide for mapping model classes to the *MULTI* subscales.

4 Results

4.1 Experimental Setup

All the models use the RoBERTa-base implementation in HuggingFace's Transformers library (Wolf et al., 2020) for obtaining contextual representations. We utilize the pooler output as defined by the library which uses the embedding of the classification token passed through a pre-trained linear layer followed by a tanh activation. We use weighted losses to account for class-imbalance in all cases. The loss weight for a label *i* is determined by $1 - \frac{n_i}{n}$, where n_i are the number of talk-turns where label i is coded, and n is the total number of talk-turns in the training data. This choice ensures that rarer classes are given greater importance during learning.

Hyperparameters. All the models use a learning rate of 10^{-5} and the RoBERTa layer is fine-tuned is each case. We use the early stopping mechanism set at 5 epochs to avoid overfitting. The macro-

averaged F1 score on a held-out validation is used to choose the best multi-label classification model. We use macro-averaged F2 score, instead, for the binary classification models since it favors recall on the positive label. This metric is ideal since the multi-label classifier would have the opportunity to correct false positives leaking from the binary classifier. Hyperparameters are tuned based on experimental results on a smaller dataset. All results are averages across three random seeds.

4.2 Results and Discussion

How do our models perform on the dataset? The comparative performance of our models are shown in Table 3. We report the model performance in terms of exact accuracy, micro and macro-averaged F1 scores across the label set, including the No Code (NC) label, and excluding it. We see that all the modeling paradigms perform almost similarly and to our surprise, the Pipeline or the MultiTask models do not produce substantial gains. Furthermore, we investigate the performance of the models on individual approach categories to understand the results further. These are reported in Table 4. We observe that model performances for categories do not deviate substantially between paradigms. By comparing to the number of training examples per label in Table 2, we observe that the performance closely correlates to the amount of data seen by the model.

Does added context help? For the results in Table 3, we consider just the therapist talk-turn and not the context surrounding it, i.e., the client and therapist talk-turns before or after it. We investigate whether adding additional context helps. We consider the following two approaches in addition to the previously shown approach:

1. Client talk-turn immediately preceding the therapist talk-turn in question can help determine the subscale. Take, for example, the Person-Centered subscale items. In these interventions, therapists often paraphrase statements which clients had just made. Hence, we concatenate the previous client **(PrevC)** talk-turn to the therapist talk-turn.
2. We observe from the training data that subscales tend to occur in chunks with the therapist opting for a certain orientation for a period of the session. We experiment with added therapist talk-turn context (**TC**) preceding and following the talk-turn in question.

We choose the MultiTask$_9$ model for this comparison which achieves the best performance. The results are in Table 5. We see that there is a small increase observed when therapist contexts are added. However, these gains are not substantial ($< 2\%$). Client context does not help the performance.
We also show some example predictions of a session snapshot in Table 6.

5 Analysis

In this section, we present analyses on the development set. We choose the best performing MultiTask$_9$ model for our analysis.

Do our models capture the global prevalence of approaches? The *MULTI*, to begin with, was intended to capture approaches at the session level. We investigate whether our models replicate the trends at a session-level. The comparative analysis for a randomly chosen session is shown in Figure 1. We see that despite making mistakes locally, the model captures approaches over therapist talk-turns. In this case, we see that the therapist scarcely uses a Psychodynamic or Interpersonal intervention and the model prediction shows similar behavior. On the other hand, the other subscale interventions are used almost uniformly over the length of the session. The model again captures this pattern.

Which categories are confused with each other? Figure 2a presents which categories tend to co-occur with each other. We observe a category Process-Experiential (PE) co-occurs with Person-Centered (PC) almost every third instance. Similarly, Psychodynamic (PD) approach almost always co-occurs with Process-Experiential (PE). Note that this is not commutative, i.e., PE co-occurs with PD about every fourth instance. Figure 2b shows the same, however, between gold labels and model prediction. Here we ask the question: for a certain category that exists in the gold data, what are the categories predicted by the model? Figures 2a and 2b should be identical if our model is ideal. Studying these figures in conjunction, gives us an idea of where the model confuses predictions the most. For example, a lot of Cognitive (CT) instances get misclassified as Person-Centered (PC), a trend which is not reflected in Figure 2a. We also observe that Psychodynamic (PD) items get significantly mispredicted as Process-Experiential (PE). A large number of approach-labeled instances get classified as 'No Code'. We expected this observa-

Test Labels	Metrics (in %)	SA	Pipeline	MultiTask$_9$	MultiTask$_{10}$
All	Exact Accuracy	76.84	74.63	**78.14**	75.86
	F1$_{Macro}$	48.24	48.52	**49.35**	47.79
	F1$_{Micro}$	**79.06**	75.43	78.63	78.17
Non-NC	F1$_{Macro}$	42.79	43.32	**44.06**	42.32
	F1$_{Micro}$	47.03	46.90	**47.64**	46.26

Table 3: Experimental results for all the classes (top half) and the eight subscales excluding 'No Code' (bottom-half)

Class	Class Abbrv.	SA	Pipeline	MultiTask$_9$	MultiTask$_{10}$
No Code	**NC**	**91.88**	90.07	91.65	91.55
Psychodynamic	**PD**	32.11	32.64	30.65	**32.97**
Process-Experiential	**PE**	67.20	65.30	**67.53**	67.32
Interpersonal	**IP**	33.25	35.21	**38.16**	34.34
Person-centered	**PC**	43.95	**44.86**	43.13	43.77
Common Factors	**CF**	**48.99**	48.87	48.06	47.30
Behavioral	**BT**	41.25	43.00	**43.90**	38.96
Cognitive	**CT**	33.95	33.12	**36.41**	34.26
Dialectical-Behavioral	**DBT**	41.62	43.60	**44.65**	39.67

Table 4: Class-wise F1 Results (in %)

Labels	Metrics	Va	PrevC	TC
All	Acc	78.14	78.34	**78.69**
	F1$_{Macro}$	49.35	49.12	**50.17**
	F1$_{Micro}$	78.63	78.67	**79.00**
Non-NC	F1$_{Macro}$	44.06	43.80	**44.96**
	F1$_{Micro}$	47.64	47.25	**48.00**

Table 5: Comparison of model performance(in %) with added contexts as compared to the MultiTask$_9$ model with just the therapist talk-turn (**Va**). This table shows results for all labels (top half) and the eight subscales excluding 'No Code' (bottom half)

tion given the skew in the training data.

6 Qualitative Analysis

F1 scores and Cohen's kappa scores cannot be compared directly. We analyze some model error examples to assess examples in a fair manner. We selected 22 examples at random with the constraint of selecting different combinations of labels. Out of the 22 examples chosen, five were ones which had an 'NC' gold label and a non-'NC' model prediction, while five had the opposite. The remaining twelve examples were mis-predictions between approach classes. Of the twelve, four were cases in which the talk-turn had a single gold approach and a single model prediction which did not match, while four each were cases in which there were multiple gold approaches but a single model predicted approach, and vice-versa. We made sure that the cases were diverse. We present five of these examples. We consider the best MutliTask$_9$ model which is trained on just the therapist talk-turn (**Va**) for this analysis.

Example 1

"*Interaction with your ex, like that's better for you*"

Human Annotation: NC

Model Prediction: IP

Here the human assessed that the talk-turn was not structured or specific enough to earn a code, despite the presence of interpersonal content. However, the model identified interpersonal language which may or may not be linked to client distress. In this case, the human seems to have been more conservative than the model in applying a code.

Example 2

"*And did you journal? Or keep a log?*"

Human Annotation: BT, CT, DBT

Model Prediction: NC

Here, journaling and log-keeping likely refers to reviewing homework, so the annotator marked an Item 10. This item, subsequently, maps onto three

Speaker	Talk-turn	Gold	SA	Pipeline	MultiTask$_9$
Client	Okay, sounds good, thank you.	-	-	-	-
Therapist	Yeah, so I just want to check in again, see how you're feeling in the room.	PD,PEI	PEI	PEI	PEI,PC
Client	Um I still feel fine, um, yeah I feel pretty good I guess.	-	-	-	-
Therapist	Okay, and that's also okay if you don't feel good, if you feel anxious. I still feel a little anxious as we're getting to know each other.	PEI, PC, CF	PEI, CF	PEI, CF	PEI
Client	Yeah.	-	-	-	-
Therapist	I just want to acknowledge that we have about twenty minutes left in our session. I'm curious is there anything you want to bring up, anywhere you want to start exploring?	CF	CF	CF	CF

Table 6: Example model predictions

subscales (BT, CT, and DBT). The model, in contrast, would not have known the homework context from this statement alone, resembling a case of atheoretical information gathering, hence an NC.

Example 3

"*Yeah and it sound sounds to me like you've already been incredibly patient with him, waiting for him to do those things, and recently he's just been letting you down over and over.*"

Human Annotation: IP

Model Prediction: IP, PC, CF

Both human and model identify clear evidence of client distress linked to an inter-personal relationship. However, the model detects justifiable PC and CF codes, explained by the emotion-added paraphrase and support for the client.

Example 4

"*I would guess that, I mean, that that's a really hard place for her to figure out.*"

Human Annotation: PE, PC

Model Prediction: CF

There is no clear argument for PE with only the context from this talk-turn. The human coder likely saw that the therapist made a paraphrase to justify the PC code. The model's CF coding is likely linked to the phrase 'really hard', which often arises from therapists providing empathic support for their client.

Example 5

"*So how was that experience, this last week of paying attention to your thoughts?*"

Human Annotation: PC, BT, CT, DBT

Model Prediction: PC

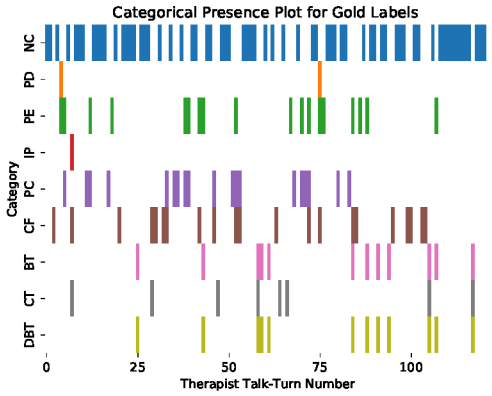
The therapist clearly asks about the client's experience, justifying a PC label. The phrase "last week of paying attention to your thoughts", however, sounds like a homework check-in (Item 10). Similar to example 2, Item 10 triggers three subscales and the human annotation of BT, CT, and DBT subscales seems appropriate and highlights a case which the model does not capture. This is an interesting case of annotation based on common-sense knowledge with which NLP models still struggle.

We should emphasize again that the humans do not annotate eight subscales directly; rather, they annotate based on the 30-item inventory. For instance, in example 2, the human annotator does not annotate the BT, CT, and DBT categories individually. They, instead, might have just annotated a single item (item 10) which maps to the three subscales. Hence, it should not be misconstrued that the human has over-labeled in that scenario.

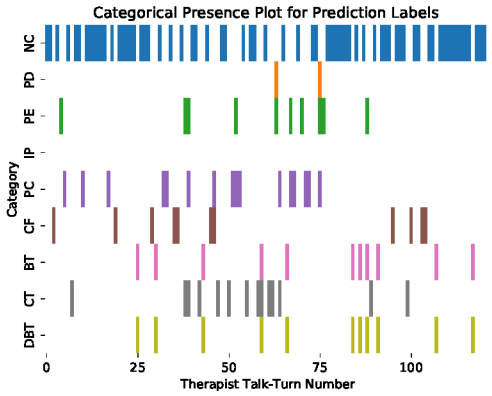
In general, after analyzing the 22 examples, we find that in many such erroneous cases, prior intra-session (short or long range) and even inter-session contextual information might be relevant to determine the correct context. We leave this as a possible direction for future research.

7 Related Work

Artificial Intelligence and its sub-domains are being increasingly discussed as possible sources of

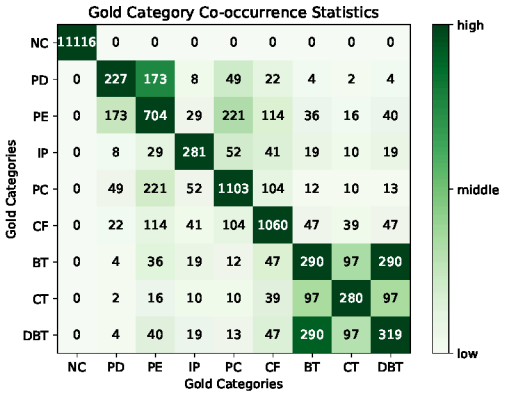


(a)

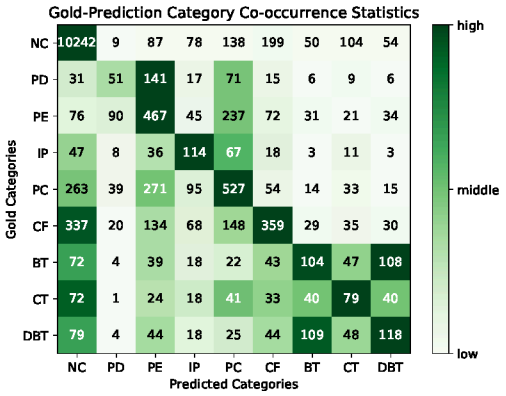


(b)

Figure 1: Predictions over a therapy session. Session proceeds left to right with a colored bar indicating the presence of an approach (or lack thereof) for the respective category. Plot (a) shows the approaches in gold annotations, (b) shows the same for model predictions.



(a)



(b)

Figure 2: Co-occurrence Statistics. Figure (a) describes co-occurrence between approaches in therapist talk-turns in the human-annotated gold data. E.g., out of 227 talk-turns where PD is annotated, 173 talk-turns also had PE annotation. Figure (b) describes co-occurrence between approaches in the gold data and the model predictions. E.g., out of 227 talk-turns where PD is annotated as mentioned in (a), only 51 talk-turns had a PD model prediction. The color gradients are normalized on rows.

improvements in mental health conversations (Lee et al., 2021; Aafjes-van Doorn et al., 2021). Moreover, transcribed therapy data from counselling centres, and public mental health forums have encouraged interest in the NLP community (Goharian et al., 2021; Le Glaz et al., 2021). NLP tools have since been used to help automate Motivational Interviewing (Tanana et al., 2016; Pérez-Rosas et al., 2017), suicide ideation detection (Huang et al., 2014; Sawhney et al., 2018), etc. to name a few. More recently, pre-trained language models have been increasing finding use in various facets like qualitative session content analysis (Grandeit et al., 2020), detecting (Wu et al., 2021) and determining the direction of empathy (Hosseini and Caragea, 2021b,a). Li et al. (2022) use transformer-based pre-trained language models to evaluate interventions from a client perspective. Client talk-turn responses to therapist interventions are evaluated based on 3-class response type and a 5-class experience type adapted from TCCS (Ribeiro et al., 2013). To the best of our knowledge, this will be the first work to automate the *MULTI* subscale assignment of therapist talk-turns.

8 Conclusion

The expanding awareness and need for mental health improvement demands the ubiquity of such resources. Therapeutic evaluation becomes increasingly important as more people leverage mental

health resources. We consider one such evaluation strategy — a talk-turn level adaptation of the *MULTI* — which evaluates therapist orientations. A major downside of such strategies remains their time-intensive nature. In this paper, we propose using pre-trained language models, which have proven to be high performance systems, to automate this evaluation. We experiment across three modeling paradigms using a pre-trained language model — RoBERTa. In addition, we show substantial analyses to understand the results. Our experiments are encouraging, however, we stress that substantial gaps in performance remain. We see this work as a significant stepping stone towards improving therapeutic feedback using NLP tools.

9 Ethics Statement

We note that the gold data used for this project was collected at a university counseling center at a university in the western United States. This induces a demographic bias in the data. It is highly possible that this data is neither representative of the various dialects of the English language spoken around the globe, nor of mental health concerns in the broader population. Our models are built using pre-trained language models, which, by design, are opaque. Consequently, our results are not interpretable.

The data was anonymized to protect information disclosures. Text snippets have been paraphrased by a Psychology graduate to mask stylistic cues.

10 Acknowledgements and COI Declarations

We would like to thank the members of the Utah NLP group, and Utah Laboratory for Psychotherapy Science for their invaluable suggestions through the course of this work. We would also like to thank the anonymous reviewers for their insightful feedback. The authors acknowledge the support of NIH/NIAAA R01 AA018673 and NSF award #1822877 (Cyberlearning).

Conflict of Interest. Drs. Imel and Atkins are co-founders and have minority equity stakes in a technology company – Lyssn.io that is focused on developing computational models that quantify aspects of patient-provider interactions.

References

Katie Aafjes-van Doorn, Céline Kamsteeg, Jordan Bate, and Marc Aafjes. 2021. A Scoping Review of Machine Learning in Psychotherapy Research. *Psychotherapy Research*, 31(1):92–116.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Derek D Caperton. 2021. Development of a Multitheoretical, Statement-level Measure of Psychotherapeutic Interventions. . *Dissertation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Torrey A Creed, David C Atkins, and Shrikanth Narayanan. 2021. Automated Quality Assessment of Cognitive Behavioral Therapy Sessions Through Highly Contextualized Language Representations. *PloS one*, 16(10):e0258639.

Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer, and Rebecca Resnik, editors. 2021. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Association for Computational Linguistics, Online.

Philipp Grandeit, Carolyn Haberkern, Maximiliane Lang, Jens Albrecht, and Robert Lehmann. 2020. Using BERT for Qualitative Content Analysis in Psychosocial Online Counseling. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 11–23.

Mahshid Hosseini and Cornelia Caragea. 2021a. Distilling Knowledge for Empathy Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021b. It Takes Two to Empathize: One to Seek and One to Provide. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13018–13026.

Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. In *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing*

and Communications and Its Associated Workshops, pages 844–849. IEEE.

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.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*, 23(5):e15708.

Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6(9):856–864.

Anqi Li, Jingsong Ma, Lizhi Ma, Pengfei Fang, Hongliang He, and Zhenzhong Lan. 2022. Towards Automated Real-time Evaluation in Text-based Counseling. *arXiv preprint arXiv:2203.03442*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Kevin S McCarthy and Jacques P Barber. 2009. The Multitheoretical List of Therapeutic Interventions (MULTI): Initial Report. *Psychotherapy research*, 19(1):96–113.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J Goggin, and Delwyn Catley. 2017. Predicting Counselor Behaviors in Motivational Interviewing Encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137.

Eugénia Ribeiro, Antonio P Ribeiro, Miguel M Gonçalves, Adam O Horvath, and William B Stiles. 2013. How Collaboration in Therapy Becomes Therapeutic: The Therapeutic Collaboration Coding System. *Psychology and Psychotherapy: Theory, Research and Practice*, 86(3):294–314.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175, Brussels, Belgium. Association for Computational Linguistics.

Nili Solomonov, Kevin S McCarthy, Bernard S Gorman, and Jacques P Barber. 2019. The Multitheoretical List of Therapeutic Interventions–30 Items (MULTI-30). *Psychotherapy Research*, 29(5):565–580.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of substance abuse treatment*, 65:43–50.

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 *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2021. Towards Low-Resource Real-Time Assessment of Empathy in Counselling. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 204–216, Online. Association for Computational Linguistics.

A Subscales, Items and Model Classes

***MULTI*-30 Items**	**Model Classes**
ϕ	No Code
2,6,12,14,15	Psychodynamic
5,7,18,23	Process-Experiential
25,26,27,30	Interpersonal
4,21,22	Person-Centered
3,11,16,17	Common Factors
8,9,19	Behavioral$_{only}$
13,20,24	Cognitive$_{only}$
28,29	Dialectical-Behavioral$_{only}$
1,10	Cognitive-Behavioral$_{shared}$

Table 7: Mapping between model classes and the *MULTI*-30 item codes (Solomonov et al., 2019). We use these classes for model training to facilitate flexibility in at a finer level. The author-defined model classes which are not part of the conventional *MULTI* subscale are highlighted.

***MULTI* Subscales**	***MULTI*-30 Items**	**Constituent Model Classes**
No Code	ϕ	No Code
Psychodynamic	2,6,12,14,15	Psychodynamic
Process-Experiential	5,7,18,23	Process-Experiential
Interpersonal	25,26,27,30	Interpersonal
Person-Centered	4,21,22	Person-Centered
Common Factors	3,11,16,17	Common Factors
Behavioral	8,9,19,1,10	Behavioral$_{only}$,Cognitive-Behavioral$_{shared}$
Cognitive	13,20,24,1,10	Cognitive$_{only}$,Cognitive-Behavioral$_{shared}$
Dialectical-Behavioral	28,29,1,10,8,9,19	Dialectical-Beh.$_{only}$,Cognitive-Beh.$_{shared}$,Beh.$_{only}$

Table 8: The conventional subscales and their constituent *MULTI*-30 items are shown here. Note that the Behavioral, Cognitive and Dialectical-Behavioral subscales (highlighted) have overlapping items. The constituent model classes from Table 7 are shown. Note that all our evaluations are presented on the conventional *MULTI* sub-scales by aggregating performance on their constituent model classes.