Predicting Client Emotions and Therapist Interventions in Psychotherapy Dialogues

Tobias Mayer[*†], **Neha Warikoo**[*†], **Amir Eliassaf** [‡], **Dana Atzil-Slonim**[‡], **Iryna Gurevych**[†]
[†]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI), TU Darmstadt
[‡]Psychotherapy Research Lab (PR Lab)
Department of Psychology, Bar-Ilan University

Abstract

Natural Language Processing (NLP) can advance psychotherapy research by scaling up therapy dialogue analysis as well as by allowing researchers to examine client-therapist interactions in detail. Previous studies have mainly either explored the clients' behavior or the therapists' intervention in dialogues. Yet, modelling conversations from both dialogue participants is crucial to understanding the therapeutic interaction. This study explores speaker contribution-based dialogue acts at the utterance-level; i.e, the therapist - Intervention Prediction (IP) and the client - Emotion Recognition (ER) in psychotherapy using a pan-theoretical schema. We perform experiments with fine-tuned language models and light-weight adapter solutions on a Hebrew dataset. We deploy the results from our ER model predictions in investigating the coherence between client self-reports on emotion and the utterance-level emotions. Our best adapters achieved on-par performance with fully fine-tuned models, at 0.64 and 0.66 micro F1 for IP and ER, respectively. In addition, our analysis identifies ambiguities within categorical clinical coding, which can be used to fine-tune the coding schema. Finally, our results indicate a positive correlation between client self-reports and utterance-level emotions.[1]

1 Introduction

Understanding the therapists' intervention and the clients' emotional response is crucial to developing more effective treatments in psychotherapy (Castonguay et al., 2021). Psychotherapy studies have emphasized the central role of client emotions and therapist interventions in predicting treatment outcomes from psychotherapy dialogues (Greenberg, 2012). However, since these studies are mainly based on client self-reports or human coding, they have been limited in scale and specificity (Imel et al., 2017). This has led to a greater push for research in Natural Language Processing (NLP) for psychotherapy (Aafjes-van Doorn et al., 2021; Shatte et al., 2019).

Recent studies have demonstrated the usefulness of NLP in automatically identifying key processes in psychotherapy, such as emotional processes, by modelling therapy dialogues (Tanana et al., 2015, 2021).

Psychotherapy dialogues like any conversation data can provide meaningful information about the speaker actions when explained for the shortest sentences within a dialogue; i.e, utterance-level. Dialogue Act (DA) classification is commonly used to attribute meaning or intention behind the utterances in a conversation (Searle, 1969; Austin, 1975).

As noted by Stolcke et al. (2000) the task and content related distinctions are important in DA labeling for conversational speech. Speaker roles define dialogue contributions in psychotherapy domain (Park et al., 2019); e.g, clients often express their *emotions* during conversation and the therapists offer various *interventions* such as helping clients to process and regulate their emotions. A sample excerpt of such a dialogue is shown in Table 1. These contributions define the types of dialogue actions characteristic to each speaker.

NLP studies of therapy dialogues tend to focus on identifying either the therapists' interventions (Cummins et al., 2019; Can et al., 2016) or the clients' emotions (Tanana et al., 2021). However, for psychotherapy researchers it is important to provide both, so the interdependence between them can be observed and analysed. Therefore, we design two application-oriented DA classification tasks based on speaker roles; i.e, DA classification for therapist utterances - *Intervention Prediction* (IP) and DA classification for client utterances - *Emotion Recognition* (ER).

[1]Code is available on Github. Models and data may not be made publicly available due to data privacy laws.

[*]Equal contribution.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1463–1477

Speaker Turn	Utterances	Dialogue Act
Client	$u_1{}^C$: I feel pretty awful about the college entrance exam. $u_2{}^C$: I think there is no way I'm going to get in.	Negative Emotion
Therapist	$u_3{}^T$: Can you say a little bit more about this feeling?	Expansion
Client	$u_4{}^C$: I feel so much pressure. $u_5{}^C$: I think that I'm not talented enough in the direction I am aiming for.	Negative Emotion
Therapist	$u_6{}^T$: Sounds like you are listening to these voices inside you that put you down and criticize you.	Interpretation

Table 1: Translated and annotated excerpt of a sample therapy session. Each Speaker-turn can have multiple utterances $u_i{}^p$, where i=utterance_id and p=[Client (C),Therapist (T)]

In contrast to the very few studies which have examined both client and therapist utterances in a dialogue (Gibson et al., 2017; Tanana et al., 2015), our approach is not limited to a specific treatment approach, such as Motivational Interviewing (MI) or Cognitive Behavior Therapy (CBT). We adopt an annotation schema which is relevant for a wide range of treatment approaches (McCullough, 1988). This *pan-theoretical* schema is compliant with the current development to share a common language between different schools of psychotherapy to identify markers of key clinical events (client emotions) that can be addressed using consensual responses (therapist interventions) (Hofmann and Hayes, 2019).

We conduct our experiments on a Modern Hebrew corpus of psychotherapy sessions transcripts consisting of 47K utterances. As Hebrew is a medium resource language, there are only limited pre-trained language models (LM) available (there are no community-wide accepted benchmarks/models) (Seker et al., 2022). Therefore, we leverage current state of the art (SOTA) models for Hebrew with adapters (Pfeiffer et al., 2020) to classify therapist and client utterances into categories of IP and ER. We experiment with adapters because they are flexible in terms of use (language agnostic, plug-and-play with many large language models), extendable (adapter fusion), and computationally scalable.

Our models and pan-theoretical approach not only empower researchers from a broad spectrum of therapeutic schools to investigate crucial psychotherapy processes on a much larger scale, but also unveils opportunities for cross-dataset comparisons. Such comprehensive analysis holds the potential to yield robust conclusions about the moment-by-moment sequences of therapists' interventions and patients' emotional responses that predict positive treatment outcomes. Such insights and measures can be integrated into existing feedback and monitoring systems and allow clinicians and mental health providers to seamlessly monitor clients' mental states without burdening them with completing questionnaires, assist clinicians in diagnosing signs of mental health problems and provide precise and swift interventions. We are currently conducting experiments to identify sequences of therapists' interventions that lead to patients' emotional improvement over time and plan to include these results in our future work.

Alternatively to this main line of experiments, we provide another useful downstream application scenario in this work, where we deploy the results from our ER model to investigate whether *emotional coherence* exists between self-reported client emotions over a session and utterance-level client emotions. Coherence between emotional expression and emotional experience is considered important to the clients' well being. To summarize, the main contributions of this work are:

- To the best of our knowledge, this is the first study proposing a framework to predict both client emotions and therapist interventions according to a *pan-theoretical* schema.
- We provide an easily extendable model for the automated prediction of the therapist interventions and the client emotions which allows scaling up psychotherapy research and detecting interdependence between them for understanding psychotherapy dialogue.
- Our data-driven analysis offers significant insights into ambiguities and challenges in the clinical coding schema that can further improve psychotherapy research.
- Finally, we give one example how to put our model output to practical use by supporting status monitoring of patients with our coherence study between clients' self-reported emotions and predicted emotions.

2 Related Work

2.1 Clinical Psychology and Intervention Prediction

Earlier works employing NLP techniques for intervention related prediction focused on specific psychotherapy approaches, such as CBT, therefore limiting the applicability to this specific type of treatment approach. For example, Flemotomos et al. (2018) used a Linear Support Vector Machine (SVM) for a CBT dataset showing that specific therapist interventions were predictive of session quality. Other studies have shown the usefulness of deep learning models to automatically annotate therapist interventions in an online text-based CBT (Cummins et al., 2019).

However, expert therapists often tend to be flexible and integrate interventions from different approaches (Solomonov et al., 2016), which makes it important to identify interventions that are pan-theoretical and relevant to various treatment approaches.

Two recent studies have used coding schema applicable to more than one treatment approach. Lee et al. (2019) drew their annotation schema of therapist utterances from DA theory and defined five high-level categories for IP. They used SVM and Neural Network-based (NN) models on a corpus of psychotherapy transcripts from various therapeutic approaches. In another work, Sun et al. (2021) created a Chinese dataset of question/answer pairs from an online mental health service platform, where the labels are similar to the Psychotherapy Interactional Coding system (McCullough, 1988) as used in this study. In their experiments they investigated strategy identification by fine-tuning a Chinese version of RoBERTa (Liu et al., 2019).

2.2 Clinical Psychology and Emotion Recognition

Earlier works identifying emotions in psychotherapy by Mergenthaler (1996, 2008) used dictionaries with negative or positive emotions to examine their prevalence in therapy sessions. Psychologically meaningful features from Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015) have been used in such studies. In a recent study Tanana et al. (2021) used BERT with a dictionary-based approach to automatically label clients' and therapists' emotions.

Beyond the use of linguistic features, the significance of dialogue history in understanding emotions was underscored in the early works by Majumder et al. (2019) using Recurrent Neural Networks (RNN) on benchmark ER datasets (Busso et al., 2008). However, its role has not yet been explored in the psychotherapy domain, which we study in this work. Recent works by Ghosal et al. (2020), Li et al. (2021) and Zhu et al. (2021) such as COSMIC, SKAIG and TODKAT all make use of common-sense knowledge graphs such as COMET for ER in dialogues (Bosselut et al., 2019). However, these are not available for Hebrew.

2.3 Utterance Labelling in Psychotherapy

In this section, we discuss psychotherapy studies which explore utterance-level labelling for both therapist and client together. Earlier studies using computerized methods in this field have focused on behavioral coding, particularly Motivational Interviewing Skill Codes (MISC). These pioneering works have paved the way for scaling up psychotherapy research. However, their focus on a coding schema from a specific psychotherapy approach, limits the ability of researchers coming from other evidenced-based psychotherapy approaches. Consistent with the growing effort in the psychotherapy field to adopt a pan-theoretical perspective that would allow clinicians and researchers from different psychotherapy schools to share a common language, the pan-theoretical coding schema used by us can be used by researchers from various therapeutic approaches (such as CBT, MI, psychodynamic, or interpersonal psychotherapy) to explore which therapists' interventions lead to positive emotional response in patients. In contrast, the MISC schema used in prior works is exclusively applicable to motivational interviewing.

Examples of such prior works are Xiao et al. (2016) or Gibson et al. (2017) who used utterance-level embeddings with RNN and LSTM models to predict these MISC labels for the therapist and the client utterances in a context-independent manner. Tanana et al. (2015) and Can et al. (2015) developed the same task as sequence labelling using RNN and linear chain CRF models respectively. A more comprehensive study on client and therapist labelling task by Gibson et al. (2022) used coding labels from MI and CBT. They developed a multi-label and multi-task approach, with turn context achieving the highest combined prediction for behavioral coding. However, this task has not been evaluated with current BERT models for a

low-resource psychotherapy domain setting. We also cannot point to any study which has explored both an utterance-level DA classification based on speaker roles in psychotherapy dialogue and, the role of dialogue context in such tasks. Therefore, we propose different SOTA-based classification baselines to label therapists' interventions and clients' emotions for utterances.

3 Dataset

Participants The data consists of 872 sessions from 68 clients in psychotherapy sessions that took place at a large university outpatient clinic and were collected as part of the regular practice of monitoring clients' progress. Individual psychotherapy consisted of once-weekly sessions. The most common diagnoses were comorbid anxiety, and affective or comorbid disorders.

Therapists and Therapy Clients were treated by 59 therapists. The dominant approach in the clinic is short-term Psychodynamic Psychotherapy (Shedler, 2010; Summers and Barber, 2010), however, the clinic supports a pan-theoretical training paradigm that involves teaching therapists to be attuned to clinically meaningful scenarios and respond to them using evidence-based strategies from various treatment approaches, such as Schema therapy (Young et al., 2003) and CBT (Beck, 1979).

Coding Categories: Therapists' Interventions Therapists' interventions were assessed using an adaptation of the Psychotherapy Interactional Coding system (PIC; McCullough, 1988). The 8 category coding system from PIC was designed to examine therapists' interventions from a wide range of psychotherapy approaches as shown in Table 2.

Coding Categories: Client's Emotions Client emotions were labeled with four categories of emotional valence: *Negative*, *Positive*, *Neutral*: defined as neither negative nor positive emotional valence, and *Mixed*: defined as both negative and positive emotional valence. This categorization of emotions is common across therapeutic approaches (Greenberg, 2012).

Coding Procedure A sub-sample of 196 sessions was coded speech-turn by speech-turn by clinical experts and is referred to as the **872_Gold** set. It consisted of 22798 therapist utterances (L) and 22248 client utterances (M). Coders were 20 trained undergraduate students. Out of the 196 sessions 22 (11%) were coded twice, once by a trained undergraduate annotator and once by a clinical psychology doctoral student. This led to therapists' utterances Cohen's kappa of 0.65 (substantial agreement) and clients' utterances Cohen's kappa of 0.54. Given the natural ambiguity of this task, this is an acceptable level of inter-rater reliability that is consistent with what was achieved in previous studies (Town et al., 2012). The remainder of the 676 un-annotated sessions is referred to as **872_Silver**.

Self reported client emotions The self-reported emotional experience was measured with the *Profile of Mood States* (Cranford et al., 2006) rating scale. The POMS consists of 12 words aggregated to describe current negative (e.g, sad) or positive (e.g., happy) emotional states. Clients were asked to evaluate how they felt during the session on a five-point Likert scale.

4 Methodology

4.1 Task Definition

Formally, given an input sequence of N utterances [u_1^p, u_2^p.....u_N^p], where p= [client C, therapist T], each utterance u_i^p=[$u_{i,1}$, $u_{i,2}$,.....$u_{i,J}$] has J words. Our DA labelling tasks are:
1) IP I_l for therapists' utterances [u_1^T, u_2^T.....u_L^T], where I_l=Intervention labels, L = no. of therapist utterances.
2) ER E_m for clients' utterances [u_1^C, u_2^C.....u_M^C], where E_m=Emotion labels, M = no. of client utterances.
Finally, we calculate Emotion Coherence analysis using:

$$Cohr(\tilde{P}_e, \tilde{E}_e) = Correlation(\tilde{P}_e^s, \tilde{E}_e^s) \quad (1)$$

$$\tilde{E}_e^s = \frac{\#(E_x \subset [s])_e}{\sum_{k \subset [pos,neg,mix,neu]} \#(E_x \subset [s])_k} \quad (2)$$

where e=[pos, neg], $\tilde{E}_e^s$ = normalized session score for predicted e, $\tilde{P}_e$ = normalized session score for client self reports; i.e., POMS for emotion e, and $E_x \subset [s]$ = predicted E_x from session s.

4.2 Model

We formulate the DA labelling tasks as a sentence-classification problem as done previously by Lee and Dernoncourt (2016); Khanpour et al. (2016); Lee et al. (2019). To establish baselines using sentence-level classification on utterances for both

Annotation	Code	Definition
Clarification	CL	Statements which restate or reflect the client's remark.
Interpretation	IT	Explanation of patterns in client's behavior and expansion of clients' understandings.
Support	SP	Restoring the client's sense of well-being through sympathy, empathy, praise or reassurance.
Directive	DR	Statements advising the client to respond in a certain way either during/outside of the session.
Information	IF	Providing information to the client in a teaching manner (does not contain advice).
Expansion	EX	Comments/questions through which therapists gather information and expand the knowledge.
Self-disclosure	SD	Comments where therapists deliberately refer to their personal thoughts and feelings.
Filler	F	Statements that do not fit into the other categories, including words or humming.

Table 2: Coding definition for therapist interventions following the Psychotherapy Interactional Coding system. An example for each category is given in Table 6 in the Appendix.

tasks, we primarily experiment with a) SOTA models for Modern Hebrew, and b) their adapter versions (Pfeiffer et al., 2020). We also conducted minor experiments with a recent few-shot learning approach, i.e., SetFit (Tunstall et al., 2022) to get a better understanding of the difficulties of the task.

In general, DA sentence-classification tasks have been shown to perform well with dialogue context for conversation datasets (Can et al., 2016; Park et al., 2019; Ortega and Vu, 2017). Therefore, we studied the role of dialogue context in psychotherapy dialogue understanding.

We test this in two setups 1) *dialog context-independent* classification (DC_I): only u_i^p is considered by the model, i=current utterance index, and 2) *dialog context-based* classification (DC_B): we describe *dialogue context* $\mathbf{DC_{u_i^p}}$ = [$u_{i\text{-}3}^p$, $u_{i\text{-}2}^p$.....u_i^p], where p=[C, T], i=current utterance index $\subset [4, N]$. Our $DC_{u_i^p}$ size is 4[2] and is independent of the role of the speaker.

Finally, we generate predictions with our best performing model on the **872_Silver** subset to scale the annotation to the full dataset. From this we use the labels to calculate coherence as shown in Equation 1 for both positive and negative emotions.

5 Experimental Setup

For our experiments, we selected four pre-trained models which are capable of handling the Hebrew language 1) XLM-RoBERTa-base model (Conneau et al., 2020), a multilingual LM based on the RoBERTa architecture (Liu et al., 2019) 2) HeBERT (Chriqui and Yahav, 2022), a monolingual BERT model trained on Hebrew data, and 3) AlephBERT (Seker et al., 2022), another monolingual BERT-based model trained on a larger Hebrew vocabulary, and 4) a multilingual T5 model, i.e., mT5 (Xue et al., 2021).

[2]Prior literature used a context size of 3-5. Our decision was also influenced by maintaining a reasonable input token length for the transformer models.

We also experiment with light-weight adapter solutions on the aforementioned models where only a small number of task specific parameters are trained. We use bottleneck adapters (Houlsby et al., 2019) and Mix-and-Match (MAM) adapters (He et al., 2021a) for training. As the pre-trained sentence transformer for SetFit, we use the paraphrase-xlm-r-multilingual-v1 with 2 epochs for the contrastive fine-tuning. All models are implemented in PyTorch using the transformers v4.18.0 library (Wolf et al., 2020) and its adapter-transformers v3.0.0 extension (Pfeiffer et al., 2020).

For the experiments, **872_Gold** is split into training (70%), development (10%) and test (20%) sets. Learning rates (lr) and epochs are determined via hyper-parameter tuning on the development set. The learning rate is set in both DC_I and DC_B setups to 1e-4 for adapters, 1e-3 for SetFit, 2e-6 for XLM, and 3e-5 for the remaining LMs.

The maximum token size per utterance J is set to 128. To account for potential variability in the results, we run each setup as a 10-fold stratified cross validation (CV). We also conducted approximate randomization tests (Dror et al., 2018) to test for significance ($\alpha = 0.05$) between the DC_B and DC_I versions of a model. We further implement partial class balancing to counter the skewed class distribution (Chawla et al., 2002), see Appendix A.2 Figure 2. We follow previous works and report model performance with micro F1 as well as Cohen's kappa (Tanana et al., 2021). Evaluating our results on Cohen's kappa (upper bound of human annotations) and F1 (ground truth by clinical experts) helps contextualize results (human vs model) and characterize the difficulty of the tasks.

To calculate the correlation (Equation 1) and its corresponding significance values, we use the Pearson implementation from the SciPy library (Benesty et al., 2009; Kowalski, 1972; Virtanen et al., 2020).

6 Results & Discussion

6.1 Intervention Prediction

Quantitative results Table 3 showcases our results for IP with the DC_I and DC_B setups respectively. While for some models the difference between both *DC* setups is only marginal, it is statistically significant (P value $< \alpha$), demonstrating that context indeed influences the prediction. For the DC_B setup, AlephBERT fine-tuned shows slightly better results with 0.64 F1 than the adapter version with 0.63 F1, which is in line with previous adapter literature. In the DC_I setup, however, adapters (peaking at 0.60 F1) outperform fine-tuned BERT versions as well as mT5. As also observed by He et al. (2021b), this might be due to the low-resource cross lingual setting where adapters tend to generalize better. Overall, the choice of the pre-trained model does not seem to influence the outcome, since all models perform equally well for this task. We hypothesized that bigger models, like XLM or mT5, may perform better than the smaller, monolingual models, but they perform on par for the DC_I scenario.

Concerning the few-shot learning experiments with SetFit, we could observe that one could obtain a F1 score of almost 0.50 with only 64 training samples per class. However, looking at the class-wise scores, there are major differences between the classes. While some classes like *Expansion* and *Filler* can be learned from few examples, other classes like *Self-disclosure* or *Information* benefit from more training data. Also, providing context in a few-shot scenario does not benefit the performance. We hypothesize that with small amounts of training samples, the context introduces too much heterogeneity in the training data. The key takeaway from these results is that parameter efficient[3] adapter models are competitive alternatives to classify interventions. Furthermore, the seemingly mediocre F1 scores of the models and the upper bound of 0.74 F1 and 0.65 Kappa of human performance, demonstrate the difficulty of this task.

Intervention Analysis Table 4 showcases the class-wise performance of our IP model on the left. *Expansion* (EX) and *Filler* (F) are the most predictable classes with 0.84 and 0.75 F1 using the AB-Adapter model. They are also the most common classes[4] and can be learned from very few examples. By contrast, the common *Interpretation* (IT) label yields a low F1 of 0.46 for the DC_B approach, a behavior which was also observed by Sun et al. (2021) for their *IT* class. This is interesting since one would assume that an interpretation would intuitively depend on prior context of the client.

In general, higher number of class instances does not guarantee higher classification results. For example, *Clarification* (CL) is the second most common class, yet the model only achieves 0.49 F1. In particular, *CL* and *IT* are often confused. Even the human annotators tend to have difficulties discriminating between these two clinically different categories, as illustrated by their moderate agreement during annotation as shown in Figure 1b.

This can be seen in Table 5*b)*, where the model cannot comprehend that the therapist revealed something new the client was unaware of earlier. Therapists use *CL* to reflect the clients' experience, whereas in *IT* therapists interpret the clients' experience and add something new in a way that expands the client's understanding. However, given the high confusion, this raises the question from a data driven point of view whether these two labels necessarily need to be separated or can be merged in the future within a revised clinical coding schema.

Another concept which is confused with *CL* or *IT* is the *Expansion* (EX), see Figure 1a. Both share the fact that they are expressed through questions where the *therapist revisits the client's discourse*. However, taking the few-shot experiments into account, this suggests that the majority of *Expansions* is distinctive and clearly identifiable. This is also indicated by the substantial human agreement of 0.79 Cohen's kappa. Overall, the less represented classes *Information* (IF), *Directive* (DR) and *Self-disclosure* (SD) are confused with the more common classes; i.e., *CL, IT, F*.

6.2 Emotion Recognition

Quantitative results As shown in Table 3, the fine-tuned and adapter AlephBERT perform equally well in recognizing emotions. They attain a slightly better F1 of 0.66 for the DC_B setup compared to 0.63 F1 for DC_I. The cross-lingual model (XLM) reaches a close second, both in the fine-tuned and the adapter approaches for DC_B setup, while HeBERT and mT5 share the bottom among the fine-tuned models. Both are trained on rela-

[3] see Appendix A.2 Table 7.
[4] see Appendix A.2 Figure 2a.

	Context-independent classification (DC_I)				Context-based classification (DC_B)			
	IP		**ER**		**IP**		**ER**	
Model	F1	kappa	F1	kappa	F1	kappa	F1	kappa
Annotators	0.74	0.65	0.73	0.54	0.74	0.65	0.73	0.54
XLM-ft	0.57*	0.44	0.59*	0.30	0.58*	0.47	0.64*	0.42
XLM-Adapter	0.60*	0.48	0.63*	0.40	0.57*	0.45	0.64*	0.43
XLM-MAM	0.60	0.48	0.63	0.40	-	-	-	-
mT5-ft	0.55*	0.43	0.58*	0.28	0.57*	0.45	0.61*	0.38
HB-ft	0.57	0.44	0.58	0.29	-	-	-	-
HBSenti-ft	-	-	-	-	0.61	0.51	0.61	0.37
HB-Adapter	0.59*	0.48	0.63*	0.39	0.60*	0.50	0.57*	0.34
HB-MAM	0.60	0.48	0.63	0.39	-	-	-	-
AB-ft	0.56*	0.44	0.57*	0.34	**0.64***	**0.55**	**0.66***	**0.46**
AB-Adapter	**0.60***	**0.49**	**0.63***	**0.40**	**0.63***	**0.53**	**0.65***	**0.44**
AB-MAM	**0.60**	**0.49**	**0.63**	**0.40**	-	-	-	-
SetFit-64	0.49*	0.38	0.43*	0.23	0.33*	0.21	0.39*	0.17

Table 3: F1 micro and Cohen's kappa results of SetFit (64 training samples per class), mT5, XLM-RoBERTa (XLM), HeBERT-Sentiment (HBSenti), HEBERT (HB), AlephBERT (AB) using fine-tuning (ft) and Adapters for IP and ER on 10 fold CV for the DC_I and DC_B setting. *Significance was tested with approximate randomization tests.

	Intervention Prediction								Emotion Recognition			
Model	CL	IT	SP	DR	IF	EX	SD	F	POS	NEG	NEU	MIX
Annotators	0.50	0.56	0.52	0.68	0.40	0.88	0.12	0.82	0.55	0.72	0.79	0.32
AB-ft	**0.49**	**0.46**	0.33	0.46	**0.49**	0.82	0.24	0.70	**0.44**	**0.67**	**0.75**	0.30
AB-Adapter	**0.49**	0.44	**0.47**	**0.47**	0.44	**0.84**	**0.32**	**0.75**	0.42	0.64	0.75	**0.33**
SetFit-8	0.17	0.20	0.16	0.13	0.12	0.74	0.11	0.54	0.17	0.43	0.39	0.17
SetFit-64	0.35	0.29	0.24	0.26	0.13	0.75	0.11	0.63	0.25	0.48	0.55	0.29

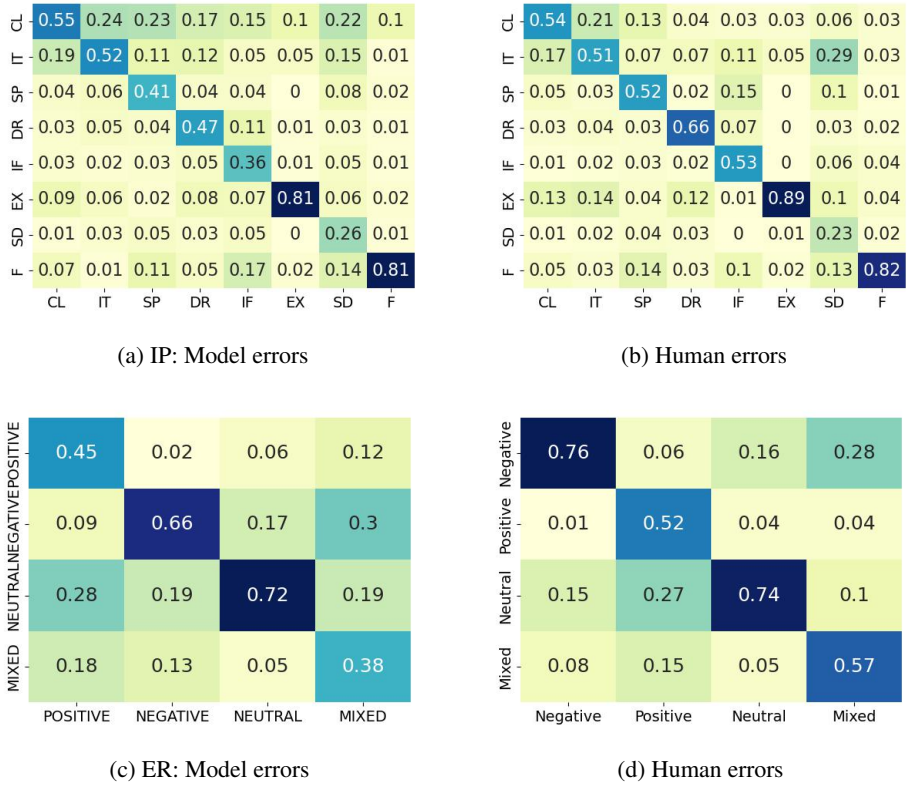
Table 4: Class-wise F1 scores of AlephBERT-ft and Adapter model in the DC_B setup and the few-shot SetFit with 8 and 64 training samples in the DC_I setup.

tively small Hebrew corpora which can explain why AlephBERT performs better.

The adapter models in the DC_I setup perform slightly better than fine-tuned models for ER, similar to the IP results. Both the fine-tuned (0.57-0.59 F1) and adapter models (0.63 F1) score lower for the DC_I setup, compared to the DC_B. This highlights the advantage of using dialogue context with LMs in understanding client utterances and classifying emotions. With respect to the few-shot learning, the context seem to not disturb the model as much as it was the case for IP. This could be due to the lower number of classes.

Emotion Analysis The right side of Table 4 depicts the results from the class-wise evaluation for ER where *Neutral* and *Negative* are predicted with a high F1 of 0.75 and 0.67, respectively, using AlephBERT (ft). To understand the performance of the *Positive* and *Negative* labels we compare the confusion matrices in Figure 1c and Figure 1d. Both *Positive* and *Negative* labels are often confused with *Neutral* followed by *Mixed* by both human annotators and the model. *Neutral* and *Mixed* code definitions present an overlap in positive and negative emotional valence. This kind of ambivalence in the definitions of the two classes causes annotators to subjectively annotate utterances as *Positive* or *Negative*. Human annotators show an agreement of Cohen's kappa 0.57 (moderate) and 0.29 (fair) for the *Neutral* and *Mixed* labels, respectively. The models also pick up on this ambiguity from the annotated data and confuse these codes with *Positive* and *Negative*. Furthermore, the SetFit model with only 64 training samples is on par with the best performing fine-tuned model and human performance (0.32 F1) for the *Mixed* class, indicating that even with more data points the model struggles to find a meaningful pattern for this class. Furthermore, we observe an inherent confusion by the human annotators as shown in Figure 1d. Humans are more likely to label it as *Neutral* or *Negative*. Our model performs slightly better in distinguishing between *Neutral/Mixed* when compared to the human annotators. However, our results in Figure 1c, highlight how this inherent bias causes *Mixed* to be most confused with *Negative*. As these biases are generated at the annotation level, revising the clinical coding might mitigate their effect.

Further analysis of our ER models identifies verbal ambiguity in Hebrew as one of the challenges. As shown in Table 5*g)* the client uses a common Hebrew slang expression to show affection. How-



(a) IP: Model errors

(b) Human errors

(c) ER: Model errors

(d) Human errors

Figure 1: Confusion matrices developed over 872_Gold. a and c showcase AlephBERT-ft predictions and b and d showcase human annotations.

Task	Utterance	Annotation	Prediction
IP	a) It bothered you, but I feel that maybe this was not your dominant emotion	IT	SD
	b) Wait, it sounds to me like this is your internal critical voice that was speaking now	IT	CL
	c) Sounds like what you're asking for is a relationship	CL	IT
	d) So you felt a strong need to write to me, because...	EX	CL
	e) So it felt like you were pushed aside, as if you are a burden?	CL	EX
ER	f) I let go of myself, and say to myself I am allowed to go through difficult times	POSITIVE	NEGATIVE
	g) I'm dying on my dad, he did it to me with good intentions	POSITIVE	NEGATIVE
	h) I did not have a good trip, but in the lab everyone thought that I enjoyed and had fun	NEGATIVE	POSITIVE
	i) yes, we drank beer, ate dinner together	POSITIVE	NEUTRAL

Table 5: Example errors for Intervention Prediction and Emotion Recognition.

ever, its intended effect is missed by the model, probably because of the negative connotation of "dying".

We further observe that despite using BERT-based approaches, our fine-tuned models still lag behind the upper bound of human agreement F1 by a margin of 7% for ER. These results likely derive from the class imbalance, inherent class confusion and a moderate annotator agreement (0.41-0.60) for ER (Table 3).

Emotional Coherence Finally the correlation analysis between the client's self-reported and predicted emotions discover a statistically significant and positive correlation between $\tilde{P_{\text{pos}}}$ and $\tilde{E_{\text{pos}}}$ (0.27, p-value=4.3e-12) and $\tilde{P_{\text{neg}}}$ and $\tilde{E_{\text{neg}}}$ (0.21, p-value=4.1e-8) for the automatically annotated **872_Silver** set. These results validate the ability of our AlephBERT (ft) model to automatically detect genuine emotions from text with specificity. This is the first study to have shown that coherence occurs between self-reported emotional experience and verbal expression of emotions allowing these measures to be integrated into existing feedback systems of mental health providers to seamlessly and non-intrusively monitor the clients' mental state in a higher temporal resolution than regular questionnaires. This result further underscores the usefulness of the evaluated models in detecting key

psychotherapy processes on a larger scale.

7 Conclusion

We evaluated various transformer models for predicting clients' emotions and therapists' interventions, where for the latter we follow current trends in psychotherapy by employing an established *pan-theoretical* schema. Our results indicate that adapter solutions offer a lightweight alternative to fine-tuning. Also, adapters for smaller language models can achieve competitive predictive performance compared to fully fine-tuned models of significantly bigger size, like mT5 (580 million parameters), while having only a fraction of computational cost, see Appendix A.2 Table 7. This is a strong advantage for our application domain as it overcomes the bottleneck of limited computational resources. We also confirm that dialogue context helps in utterance-level dialogue understanding tasks. We encounter challenges with regards to ambiguity in interpreting cultural slang in Hebrew. Our analysis further identifies ambiguities in the coding of IP and ER labels, which causes high confusion in some of the class predictions.

Our models and pan-theoretical approach paves the way for researchers from multiple psychotherapy schools to auto-annotate session data, enabling the examination of pivotal treatment processes on a significantly broader scale. In the future, the conclusions from our data-driven approach can be used to a) advance clinical coding schemes b) study language behavior with treatment-level outcomes to monitor and improve clients' well being; e.g, using emotion coherence analysis. Furthermore, our study lays the groundwork for developing support tools for therapists to provide feedback in higher temporal resolution and guidance on the interventions that lead to positive responses in clients. As next steps, we plan to extend the model to a multimodal level, integrating speech data as well. Furthermore, we will have a closer look at dialogue models and investigate how to efficiently include the grounded dialogue history.

Limitations

In this section, we discuss the limitations of our approach, datasets, and experimental setup. As mentioned in the main text, we work with a Hebrew-language clinical dataset, which poses many challenges. Using a Hebrew-based dataset for NLP in psychotherapy does offer new insights into the psychotherapy based on cultural context, but it also puts limitation on the pre-trained models we can use to develop a baseline for this study.

Psychotherapy dialogue is a conversation dataset. As we mentioned in Section 2.2, there are many SOTA dialogue conversation models like COSMIC, SKAIG and TODKAT which have performed well in ER tasks. However, we haven't experimented with any of these models as they are widely developed on English language-based knowledge graphs. Their implementation for such a study would require a Hebrew-English translation infrastructure, which was beyond the scope of our current work.

It also seems intuitive to study such a dataset with more conversation models. Most of such conversation models are developed for ER. They exploit the emotions labels of the previous utterances along with utterance encoding to capture a global context and then predict current utterance emotion. However, in our case the structure of our psychotherapy conversation is composed of different contributions by each speaker. Since one speaker's role definition is to provide clinical interventions and the other mainly expresses emotions, their successive labels do not complement each other in building a global context. This becomes a limitation in using the full potential of such existing conversation models (especially ones developed for ER). Furthermore, a more technical limitation in our approach is the size of the ante-ceding context window. Prior literature uses a context size between 3 and 5 utterances. Our decision to set the window size to four was also influenced by maintaining a reasonable input token length for the transformer model. As this is an ongoing project, we are currently expanding our research to investigate more variations of integrating context.

The scope of our current task is utterance-level classification of clients' emotions and therapists' interventions. Therefore, each utterance from all 872 sessions is considered as training input. However, such an experimental setup does not account for variability in the same clients' behavior across sessions or different behavior of different clients in this study. Empirical analysis for such setups would require expanding the scope of our study and dataset, which we hope to accommodate in our future work.

Concerning the models' performance, in particular for real-world applications, we are aware that they do not achieve human-like performance and

that false predictions, especially in a medical environment, can have a severe impact. We do neither claim nor recommend that the results from the analysis of the automatically created annotations should be directly used for therapeutic decision making. They should be rather used as a tool which indicates *potentially* important/impactful moments during therapy or *potential* relations between interventions and outcome. These findings can then serve as well-grounded suggestions and hypotheses, but still need to be tested in proper clinical trials.

The results in this work present an empirical analysis based on confidential psychotherapy sessions between clients and therapists without revealing any client information as mandated by the providers of this dataset. This restricts flexibility in data sharing which may be construed as a limitation within the NLP community. However, we request due consideration on part of our readers regarding the protocols of reproducibility, especially concerning datasets which carry ethical implications for dissemenating human opinions conducted in a confidential setup (Ian et al., 2023).

Ethics Statement

The materials were only collected after securing approval from the authors' university ethics committee. Only clients and therapists who gave their consent to participate were included in the study. Clients were told that they could choose to terminate their participation in the study at any time without jeopardizing treatment. All sessions were audiotaped and transcribed according to a protocol ensuring confidentiality and masking of any identifying information, such as names and places. Finally, to ensure privacy due to the sensitive nature of our data, secured servers were used with limited access to develop this study.

Acknowledgements

This work has been funded by the LOEWE initiative (Hesse, Germany) within the DYNAMIC center and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. It was further funded by a grant from the Israel Science Foundation (ISF #2466/21) awarded to Dana Atzil-Slonim.

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.

Aviad Albert, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2013. The hebrew childes corpus: transcription and morphological analysis. *Language resources and evaluation*, 47(4):973–1005.

John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.

Aaron T Beck. 1979. *Cognitive therapy of depression*. Guilford press.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 4762–4779.

Casey L Brown, Natalia Van Doren, Brett Q Ford, Iris B Mauss, Jocelyn W Sze, and Robert W Levenson. 2020. Coherence between subjective experience and physiology in emotion: Individual differences and implications for well-being. *Emotion*, 20(5):818.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Doğan Can, David C Atkins, and Shrikanth S Narayanan. 2015. A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. In *Proceedings of 16th Annual Conference of the International Speech Communication Association, (INTERSPEECH 2015)*, pages 339–343.

Doğan Can, Rebeca A Marín, Panayiotis G Georgiou, Zac E Imel, David C Atkins, and Shrikanth S Narayanan. 2016. "it sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of counseling psychology*, 63(3):343.

Louis G Castonguay, Michael Barkham, Soo Jeong Youn, and Andrew C Page. 2021. Practice-based evidence—findings from routine clinical settings.

N. Chawla, K. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357.

Avihay Chriqui and Inbal Yahav. 2022. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *ArXiv*, abs/2102.01909.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 8440–8451.

James A Cranford, Patrick E Shrout, Masumi Iida, Eshkol Rafaeli, Tiffany Yip, and Niall Bolger. 2006. A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, 32(7):917–929.

Ronan Cummins, Michael P Ewbank, Alan Martin, Valentin Tablan, Ana Catarino, and Andrew D Blackwell. 2019. Tim: A tool for gaining insights into psychotherapy. In *Proceedings of the World Wide Web Conference (TheWebConf 2019)*, pages 3503–3506.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 1383–1392.

Nikolaos Flemotomos, Victor R Martinez, James Gibson, David C Atkins, Torrey Creed, and Shrikanth S Narayanan. 2018. Language features for automated evaluation of cognitive behavior psychotherapy sessions. In *Proceedings of 19th Annual Conference of International Speech Communication Association (INTERSPEECH 2018)*, pages 1908–1912.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics Findings of ACL*, Proceedings of the Association for Computational Linguistics ACL: EMNLP 2020 (ACL-EMNLP 2020), pages 2470–2481.

James Gibson, David C. Atkins, Torrey A. Creed, Zac E. Imel, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2022. Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*, 13:508–518.

James Gibson, Dogan Can, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2017. Attention networks for modeling behaviors in addiction counseling. In *Proceedings of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, pages 3251–3255.

Leslie S Greenberg. 2012. Emotions, the great captains of our lives: Their role in the process of change in psychotherapy. *American Psychologist*, 67(8):697.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021a. Towards a unified view of parameter-efficient transfer learning. *ArXiv*, abs/2110.04366.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021b. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *ArXiv*, abs/2106.03164.

Stefan G Hofmann and Steven C Hayes. 2019. The future of intervention science: Process-based therapy. *Clinical Psychological Science*, 7(1):37–50.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of International Conference on Machine Learning (ICML 2019)*, pages 2790–2799.

Magnusson Ian, Smith Noah A., and Dodge Jesse. 2023. Reproducibility in nlp: What have we learned from the checklist? In *Findings of the Association for Computational Linguistics (ACL), 2023*.

Zac E Imel, Derek D Caperton, Michael Tanana, and David C Atkins. 2017. Technology-enhanced human interaction in psychotherapy. *Journal of counseling psychology*, 64(4):385.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers*, pages 2012–2021.

Charles J Kowalski. 1972. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(1):1–12.

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 6th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019)*, pages 12–23.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 515–520.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 1204–1214.

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*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019)*, 33(01):6818–6825.

L McCullough. 1988. Psychotherapy interaction coding system manual: the pic system. *Soc. Behav. Sci. Doc*, 18.

Erhard Mergenthaler. 1996. Emotion–abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of consulting and clinical psychology*, 64(6):1306.

Erhard Mergenthaler. 2008. Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research*, 18(2):109–126.

Erhard Mergenthaler and Charles Stinson. 1992. Psychotherapy transcription standards. *Psychotherapy research*, 2(2):125–142.

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGdial 2017)*, pages 247–252.

Sungjoon Park, Donghyun Kim, and Alice Oh. 2019. Conversation model fine-tuning for classifying client utterances in counseling dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Volume 1 (Long and Short Papers)*, pages 1448–1459.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (Volume 1: Long Papers)*, pages 46–56.

Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, 49(9):1426–1448.

Jonathan Shedler. 2010. The efficacy of psychodynamic psychotherapy. *American psychologist*, 65(2):98.

David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, Geoffrey C Dunbar, et al. 1998. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *Journal of clinical psychiatry*, 59(20):22–33.

Nili Solomonov, Nadia Kuprian, Sigal Zilcha-Mano, Bernard S Gorman, and Jacques P Barber. 2016. What do psychotherapy experts actually do in their sessions? an analysis of psychotherapy integration in prototypical demonstrations. *Journal of psychotherapy Integration*, 26(2):202.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Richard F Summers and Jacques P Barber. 2010. *Psychodynamic therapy: A guide to evidence-based practice*. Guilford Press.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Proceedings of the 2021 Association for Computational Linguistics: ACL-IJCNLP (IJCNLP 2021)*, pages 1489–1503.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality (CLPsych 2015)*, pages 71–79.

Michael J Tanana, Christina S. Soma, Patty B. Kuo, Nicolas M. Bertagnolli, Aaron Dembe, Brian T. Pace, Vivek Srikumar, David C. Atkins, and Zac E. Imel. 2021. How do you feel? using natural language processing to automatically rate emotion in psychotherapy. *Behavior research methods*, 53:2069–2082.

Joel Michael Town, Leigh McCullough, and Gillian E Hardy. 2012. Therapist interventions in short-term dynamic psychotherapy: Six expert treatments. *British Journal of Guidance & Counselling*, 40(1):31–42.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *CoRR*, abs/2209.11055.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

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 (EMNLP 2020)*, pages 38–45.

Bo Xiao, Dogan Can, James Gibson, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *Proceedings of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, pages 908–912.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*, pages 483–498.

Jeffrey E Young, Janet S Klosko, and Marjorie E Weishaar. 2003. Schema therapy. *New York: Guilford*, 254.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.

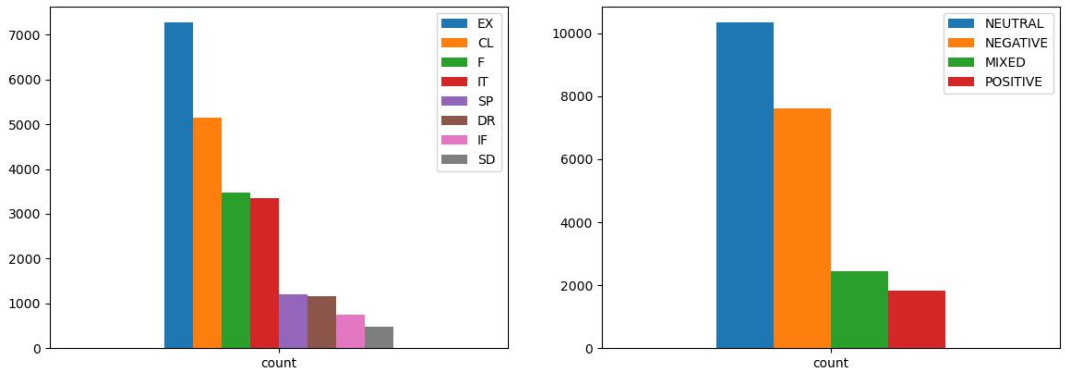
A Appendix

A.1 Data Description

Participants The clients were all above age 18 (Mean age = 39.06, SD = 13.67, range 20–77), most of them were women (58.9%). Clients' diagnoses were based on the Mini International Neuropsychiatric Diagnostic Interview for Axis I Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition diagnoses (MINI 5.0; Sheehan et al., 1998). The interviews were conducted before the actual therapy began, by well-trained independent clinicians. All intake sessions were audiotaped, and a random (25%) of the interviews were sampled and rated again by an independent clinician. The mean kappa value for the Axis diagnoses was excellent (k = .9). Of the clients, (22.9%) had one diagnosis, (20.0%) had two, and (25.7%) had three or more. The most common diagnoses were comorbid anxiety, and affective or comorbid disorders. The most common diagnoses were comorbid anxiety and affective disorders (25.7%), followed by other comorbid disorders (17.1%), anxiety disorders (14.3%), and affective disorders (5.7%). Several clients (31.4%) reported relationship concerns, academic/occupational stress, or other problems that did not meet criteria for any Axis I diagnosis.

Therapists Clients were treated by 59 therapists that were MA or PhD students at different stages of clinical psychology training (1 to 5 years of experience). Therapists received 1 hr of individual supervision and 4 hr of group supervision every week. Individual psychotherapy consisted of once-weekly sessions. Treatment was open-ended, but was often restricted from 9 months to 1 year reflecting the trainee clinicians' program. The therapy was conducted in Modern Hebrew.

Transcription To capture the treatment processes from session to session, and since the transcription process is highly expensive, transcriptions were conducted alternately (i.e., Sessions 2, 4, 6, 8, etc.). In cases where the material was not complete, the next session was transcribed instead. The transcriber team was composed of seven transcribers, all of whom were graduate students in the university's psychology department. The transcribers went through a 1-day training workshop and monthly meetings were held throughout the transcription process to supervise the quality of their work. Their training included specific guidelines on how to handle confidential and sensitive information, and the transcribers were instructed to replace names by pseudonyms and to mask any other identifying information. The transcription protocol followed general guidelines as described in (Mergenthaler and Stinson, 1992; Albert et al.,



(a) Therapists' Intervention

(b) Clients' Emotion

Figure 2: Label distribution of the therapists' intervention and clients' emotion annotation.

Annotation	Code	Example
Clarification	CL	*But usually you describe a lot of friction in your relationships with others.*
Interpretation	IT	*He is taking advantage of you.*
Support	SP	*Yes, it's important for me to say that you can always call me.*
Directive	DR	*Let's try it for the next time.*
Information	IF	*It is not prohibited by law, but is socially controversial.*
Expansion	EX	*And when you met her, was it on your initiative?*
Self-disclosure	SD	*Feeling alone is also tough for me.*
Filler	F	*Hm, okay.*

Table 6: Example for each therapist intervention following the Psychotherapy Interactional Coding system.

2013). The audiotape was transcribed in its entirety and provided a verbatim account of the session. The mean transcribed sessions per dyad was 11.79; SD = 3.08. In total, transcriptions include about five million words, said in over 250,000 utterances. There were 5,895 words in a session, on average.

Collection Procedure The procedures were part of the routine assessment and monitoring process in the clinic. Materials were taken in accordance with the approval of the University Ethics Committee. All sessions were audiotaped and transcribed using a protocol that ensures confidentiality. The participants were only clients and therapists who gave their consent to be included in the study. Clients were informed that they can terminate their participation any time.

A.2 Detailed Results

In this section, we present additional analysis from our study to help better understand the results in Section 6.

Data distribution of therapist and client clinical coding Figure 2 highlights the skewness in data distribution for both therapist and client utterance. *EX* and *CL* dominate therapist intervention labels, refer Figure 2a. We can also observe that *positive* emotions are relatively less in psychotherapy dialogues (refer to Figure 2b), which is intuitive as therapist intervention are aimed to move clients from *negative* to *positive* emotion state.

Computational resources and runtime All experiments were conducted on an in-house computation cluster. All models were trained on at least one NVIDIA Tesla P100 with 16GB of VRAM. Table 7 shows the exact GPU memory occupied as well as the time (in minutes) for training each context-independent model (128 input tokens) with the specifications of parameters reported in Section 5.

Emotional Coherence between self-reports and verbal expression The results on 872_Gold in Table 8 show a positive, statistically significant correlation between $\tilde{P_{\text{pos}}}$ and $\tilde{E_{\text{pos}}}$ (0.29) and $\tilde{P_{\text{neg}}}$ and $\tilde{E_{\text{neg}}}$ (0.24) across all sessions. This result is based on expert annotated emotion labels, and the positive correlation confirms that coherence exists between subjective expression of clients' emotions and verbal expression of emotions even when studied with traditional approaches. These results are consistent with previous studies that have been reported

Model	Time	GPU memory
mT5-ft	585	17361
XLM-ft	173	8354
AB-ft	26	4240
HB-ft	29	4017
XLM-Adapter	14	3479
AB-Adapter	16	2899
HB-Adapter	12	3567

Table 7: GPU memory usage (in MB) and elapsed time (in minutes) for training each model on P100 GPU(s).

	872_Gold	872_Silver
($\tilde{P_{\text{pos}}}$, $\tilde{E_{\text{pos}}}$)	(0.29,7.8e-05)	(**0.27, 4.3e-12**)
($\tilde{P_{\text{neg}}}$, $\tilde{E_{\text{neg}}}$)	(0.24, 0.001)	(**0.21, 4.1e-08**)

Table 8: Session-wide Correlation between POMS and Utterance emotion labels.

coherence across various emotional response systems (e.g., (Brown et al., 2020)), but extend beyond them by showing that coherence also occurs between emotional experience and verbal emotion expression.

Table 8 also depicts a significant positive correlation between $\tilde{P_{\text{pos}}}$ and $\tilde{E_{\text{pos}}}$ (0.27) and $\tilde{P_{\text{neg}}}$ and $\tilde{E_{\text{neg}}}$ (0.21) for 872_Silver. These results validate the performance of the transformer-based ER approach.