Psychotherapy

Leveraging Natural Language Processing to Study Emotional Coherence in Psychotherapy

Dana Atzil-Slonim, Amir Eliassaf, Neha Warikoo, Adar Paz, Shira Haimovitz, Tobias Mayer, and Iryna Gurevych

Online First Publication, January 18, 2024. https://dx.doi.org/10.1037/pst0000517

Psychotherapy

ISSN: 0033-3204

https://doi.org/10.1037/pst0000517

Leveraging Natural Language Processing to Study Emotional Coherence in Psychotherapy

Dana Atzil-Slonim[1], Amir Eliassaf[1], Neha Warikoo[2], Adar Paz[1], Shira Haimovitz[1], Tobias Mayer[2], and Iryna Gurevych[2]

[1] Department of Psychology, Bar-Ilan University

[2] Department of Computer Science, Technical University of Darmstadt

The association between emotional experience and expression, known as emotional coherence, is considered important for individual functioning. Recent advances in natural language processing (NLP) make it possible to automatically recognize verbally expressed emotions in psychotherapy dialogues and to explore emotional coherence with larger samples and finer granularity than previously. The present study used state-of-the-art emotion recognition models to automatically label clients' emotions at the utterance level, employed these labeled data to examine the coherence between verbally expressed emotions and self-reported emotions, and examined the associations between emotional coherence and clients' improvement in functioning throughout treatment. The data comprised 872 transcribed sessions from 68 clients. Clients self-reported their functioning before each session and their emotions after each. A subsample of 196 sessions were manually coded. A transformer-based approach was used to automatically label the remaining data for a total of 139,061 utterances. Multilevel modeling was used to assess emotional coherence and determine whether it was associated with changes in clients' functioning throughout treatment. The emotion recognition model demonstrated moderate performance. The findings indicated a significant association between verbally expressed emotions and self-reported emotions. Coherence in clients' negative emotions was associated with improvement in functioning. The results suggest an association between clients' subjective experience and their verbal expression of emotions and underscore the importance of this coherence to functioning. NLP may uncover crucial emotional processes in psychotherapy.

Clinical Impact Statement

Question: The present study examined the coherence between clients' verbal expressions of emotions and their subjective experience of emotions and whether this coherence was associated with an improvement in clients' functioning. ***Findings:*** The findings demonstrate the usefulness of computerized text analytic techniques to automatically annotate clients' emotions. The results confirm the association between clients' subjective experience and their verbal expression of emotions. ***Meaning:*** The findings highlight the relevance of emotional coherence for clients' functioning, especially with regard to negative emotions. ***Next Steps:*** Automatic emotion recognition models can be integrated into existing feedback systems to provide an indication of the levels of emotional coherence in psychotherapy sessions and allow therapists to adapt their interventions accordingly.

Keywords: emotional coherence, machine learning, natural language processing, emotion recognition, psychotherapy process outcome

Supplemental materials: https://doi.org/10.1037/pst0000517.supp

Dana Atzil-Slonim https://orcid.org/0000-0001-6958-1200

This study was funded by a grant from the Israel Science Foundation (ISF #2466/21) awarded to Dana Atzil-Slonim.

Dana Atzil-Slonim played a lead role in conceptualization, data curation, funding acquisition, investigation, methodology, writing–original draft, and writing–review and editing. Amir Eliassaf played a supporting role in project administration, writing–original draft, and writing–review and editing. Neha Warikoo played a lead role in formal analysis and a supporting role in writing–review and editing. Adar Paz played a supporting role in writing–review and editing and an equal role in formal analysis. Shira Haimovitz played a supporting role in writing–original draft and writing–review and editing. Tobias Mayer played a supporting role in formal analysis and writing–review and editing. Iryna Gurevych played a supporting role in formal analysis, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Dana Atzil-Slonim, Department of Psychology, Bar-Ilan University, Ramat-Gan 5290002, Israel. Email: dana.slonim@gmail.com

Emotions are experienced and expressed through a range of response systems such as individuals' subjective experiences, thoughts, behavior, and physiology (Greenberg, 2012). Theories of emotions posit that coherence across different emotional responses promotes better functioning (e.g., Ekman, 1992; Levenson, 2003). Emotional coherence is defined as the coordination, or association, between different emotional responses as the emotion unfolds over time (Mauss et al., 2005). Many studies have shown that individuals differ in their degree of coherence across emotional responses (e.g., Brown et al., 2020) and that a lack of emotional coherence tends to be associated with lower psychological functioning (e.g., Leonhardt et al., 2018; Mauss et al., 2011). Previous studies on emotional coherence have primarily relied on laboratory-generated emotional stimuli to assess emotional responses, often by comparing clinical to nonclinical populations at one or only a few time points (e.g., Hastings et al., 2009; Mauss et al., 2011; Negrao et al., 2005; Wagner et al., 2003). Hence, these studies are limited in their ability to determine how different emotional responses change together over time in a genuine real-life interactions such as in psychotherapy. These studies have focused on the relationship between certain aspects of emotional responses such as between individuals' self-reports of their emotions and physiological responses (e.g., Brown et al., 2020; Lohani et al., 2018) or facial expressions (Lohani et al., 2018). However, the coherence between subjectively experienced emotions and verbally expressed emotions remains unexplored, even though there is significant theoretical backing for its importance in the psychotherapy literature (e.g., Greenberg, 2012; Lane et al., 2022).

Various psychotherapy theories have highlighted the importance of emotional coherence and suggest that individuals' ability to progress beyond their raw emotional experiences toward being able to verbally express them is crucial to psychological functioning (e.g., Fosha, 2001; Greenberg, 2012; Lane et al., 2022). According to these theories, emotional coherence is part of emotional processing, a broader term that encompasses a range of emotional processes such as the ability to experience emotions, become aware of one's emotions, differentiate between emotions, provide meaning to one's experience, and resolve discrepancies between felt emotions and expressed emotions (Pascual-Leone, 2018). One primary objective of psychotherapy is to provide clients with the opportunity to align their emotional experiences with the words that describe them, which is expected to lead to better functioning (Fosha, 2001; Greenberg, 2012). When individuals are able to express their emotions in a way that matches their internal emotional state, the better they are able to communicate and handle their emotions (Greenberg, 2012; Gross et al., 2000; Levenson, 2003). In contrast, when individuals experience emotions but are unable to express them, or express emotions without recognizing how they feel, their ability to adaptively communicate and handle their emotions is impaired (Greenberg, 2012; Lane et al., 2022). Despite the strong theoretical support for the importance of emotional coherence between subjectively experienced emotions and verbally expressed emotions, there is surprisingly little empirical research on this topic in psychotherapy literature.

To date, psychotherapy studies have largely focused on a given aspect of emotions such as emotional experience (e.g., Fisher et al., 2016; Pos et al., 2009) or emotional expression (Mergenthaler, 2008) and its association with treatment outcomes; however, the coherence between these emotional responses and the association of such coherence with clients' functioning has yet to be explored.

The lack of research on emotional coherence in psychotherapy may have to do, at least in part, with traditional methods of assessing emotions in psychotherapy. Many studies have used self-reports to assess clients' subjective experience of emotions in psychotherapy sessions (e.g., Bar-Kalifa & Sened, 2020; Fisher et al., 2016). While self-reports are easy to obtain and allow access to the subjective experience of emotions, they rely on retrospective reporting, which does not capture the fluctuating nature of emotions from moment-to-moment. To tap emotional responses within psychotherapy sessions, several studies have utilized observers' ratings of clients' emotions (e.g., Kramer et al., 2015; Pos et al., 2017). These studies provide a rich and detailed perspective of emotional processes in psychotherapy. Nevertheless, since observational human coding is very labor-intensive and expensive to implement, these studies have typically focused on a small sample of clients and are conducted at limited time points. Recent developments in artificial intelligence (AI), machine learning (ML), and natural language processing (NLP) that can automatically capture utterance-level verbally expressed emotions make it possible to study emotional processes in psychotherapy at a higher scale and specificity (Aafjes-van Doorn et al., 2021; Delgadillo & Atzil-Slonim, 2023). Given the central role of language in psychotherapy, NLP methods are particularly pertinent to analyzing psychotherapy sessions.

There is a wide range of NLP techniques that can be used to accurately identify emotions in text. These are commonly referred to as sentiment analysis or emotion recognition (ER; for review, see Nandwani & Verma, 2021). Earlier works that used text mining to automatically identify emotions in psychotherapy sessions were based on dictionaries of negative and positive emotion words (e.g., Mergenthaler, 2008; Tausczik & Pennebaker, 2010). These dictionary-based methods are simple to implement and can accurately identify emotional words. However, they rely on a preexisting dictionary or lexicon, which may not always be available for different languages or may not cover all possible words or phrases. Furthermore, despite the fact that text data are sequential and context playing a critical role in fully understanding a sentence, these dictionary-based methods are unable to capture the context surrounding words or the connections between words, thereby diminishing their effectiveness.

In recent years, deep learning models have become the dominant method for emotion recognition (Nandwani & Verma, 2021). Early works on NLP for emotion recognition used sequential language models such as recurrent neural networks (RNNs; e.g., Majumder et al., 2019) that analyze short word sequences (Gers et al., 2000). More recent transformer-based architecture studies use each word bidirectionally in the context of an entire sentence, which yields more robust word representations. Transformer-based language models, such as the bidirectional encoder representation from transformers (BERT; Devlin et al., 2018), are pretrained on huge data sets of unannotated text by randomly masking some of the words and training the model to predict them. This allows the model to learn the underlying structure of the language and the context in which words appear. After pretraining, the model can be fine-tuned for a specific task on a smaller labeled data set, such as emotion recognition in natural psychotherapy text by updating its parameters to optimize a task-specific objective. The versatility and capability of transformer-based language models have resulted in their widespread

application across a variety of research fields, including mental health (Delgadillo & Atzil-Slonim, 2023). Studies have reported the potential usefulness of these methods in analyzing psychotherapy data (e.g., Cao et al., 2019; Ewbank et al., 2021). In one such study, Tanana et al. (2021) showed that deep learning models outperformed dictionary-based models in accurately detecting verbally expressed clients' emotions (with a kappa of 0.31 vs. 0.25). Using these advanced methods to automatically annotate large data sets with emotion labels at the utterance level can serve to explore emotional processes that have previously been neglected, such as the coherence between verbally expressed emotions and self-reported emotions and whether this coherence is associated with clients' functioning.

The present study had three objectives:

Aim 1: Use state-of-the-art language models for emotion recognition to automatically label clients' utterance-level emotions during psychotherapy conversations. Based on recent studies that have used deep learning models to automatically label psychotherapy data (e.g., Ewbank et al., 2021), we expected that the model's performance would be comparable to human interrater reliability (prerequisite Hypothesis 1).

Aim 2: Use these labeled data to examine the emotional coherence between verbally expressed emotions and self-reported emotions (Hypothesis 2). Based on previous studies that have reported emotional coherence in other modalities (e.g., Hastings et al., 2009), as well as psychotherapy theories on emotional coherence (e.g., Greenberg, 2012; Lane et al., 2015), we expected to find emotional coherence between verbally expressed emotions and self-reported emotions for both positive and negative emotions.

Aim 3: Examine whether emotional coherence would be associated with greater improvement in functioning throughout treatment. Based on previous findings on the association between emotional coherence and functioning (e.g., Leonhardt et al., 2018), we anticipated that higher emotional coherence would be associated with improvement in clients' functioning throughout treatment (Hypothesis 3).

Method

This study was conducted in the community research clinic of Bar-Ilan University, Israel, and approved by the associated institutional review board. Data were collected naturalistically at a large university outpatient clinic, as part of the clinic's regular practice of monitoring clients' progress. Clients were asked to sign consent forms and were told they could choose to terminate their participation in the study at any time without jeopardizing treatment. Due to the sensitive nature of the textual data, secured servers with limited access were used to develop this study. The questionnaires and labeled data, materials, and analysis code for this study can be accessed from the first author upon request.

Participants and Treatment

Clients

The data were drawn from a pool of 180 clients who were in individual psychotherapy between August 2014 and August 2016 and had given their consent to participate in the study. Thirty-four clients (18.88%) dropped out (deciding one-sidedly to end treatment before the planned termination date). Clients were selected if they had complete data including audio recordings that were used for the transcriptions and session-by-session questionnaires. Clients were excluded based on the Mini International Neuropsychiatric Diagnostic Interview for Axis I *Diagnostic and Statistical Manual of Mental Disorders*, *Fourth Edition* (MINI; Sheehan et al., 1998), if they were diagnosed as severely disturbed, either because of a current crisis, past severe trauma or associated posttraumatic stress disorder, a past or present psychotic or manic diagnosis, and/or current substance abuse. Based on these criteria, we excluded 77 clients (42%). Thus, of the total sample, the data for 68 clients (38%) who met the inclusion criteria were transcribed, for a total of 872 transcribed sessions. The clients were all above age 18 (M_{age} = 39.06, SD = 13.67, range 20–77), and most were women (58.9%). Of the clients, 92% were native Hebrew speakers and 92% were born in Israel. Of the clients, 53.5% had at least a bachelor's degree; 53.5% were single and 8.9% were in a committed relationship but unmarried; 23.2% were married and 14.2% were divorced or widowed. Intake interviews were conducted by experienced independent clinicians before the actual therapy began. 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 I 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 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 the criteria for any Axis I diagnosis.

Therapists and Therapy

The clients were treated by 52 therapists. All were MA or PhD students at different stages of their clinical psychology training (1–5 years of experience). All the therapists were native Hebrew speakers. Clients were assigned to therapists in an ecologically valid manner reflecting therapist availability and caseload. Most therapists treated one client each, but some (eight) treated two. Each therapist received 1 hr of individual supervision and 4 hr of group supervision on a weekly basis. All therapy sessions were audiotaped for use in supervision with senior clinicians. The individual and group supervision focused heavily on the review of the audiotaped case material and the appropriateness of the therapists' interventions. The supervisors were senior clinicians. Individual psychotherapy consisted of once-weekly sessions. The dominant approach in the clinic is short-term psychodynamic psychotherapy (Shedler, 2010; Summers & 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 by integrating evidence-based strategies from various treatment approaches, such as schema therapy (Young et al., 2005) and cognitive behavioral therapy (Beck, 1979). Treatment was open-ended in length but was often restricted to 9 months–1 year, reflecting the trainee clinicians' program and the university calendar. On average, treatment spanned 37 sessions (SD = 23.99, range = 18–157). The language of therapy was modern Hebrew.

Instruments and Data Collection

The Outcome Rating Scale (ORS)

The ORS is a four-item Visual Analog scale developed as a brief alternative to the Outcome Questionnaire–45 (OQ-45). It assesses change in three areas of client functioning that are widely considered to be valid indicators of progress in treatment: functioning, interpersonal relationships, and social role performance (Miller et al., 2003). Respondents complete the ORS before each therapy session by rating four statements on a Visual Analog scale anchored at its respective extremes by the words low and high. This scale yields four separate scores between 0 and 10 (for a total score between 0 and 40), with higher scores indicating better functioning. According to the ORS manual, a score of 24 represents the threshold for clinical status. The Reliable Change Index for the ORS is 5; thus, cases with a gain score of 5 and above are classified as improved. The ORS showed excellent internal consistency in the current sample ($\alpha = .95$).

Profile of Mood States (POMS; McNair et al., 1992)

The POMS is a widely used instrument that assesses mood variables. For the purpose of this study, we used an abbreviated version of the measure, which was adapted for intensive repeated measurements (Cranford et al., 2006) and consists of 12 words that describe current emotional states. The Negative Affect scale includes sadness (two items), anxiety (two items), and anger (two items). The Positive Affect scale includes contentment (two items), vigor (two items), and calmness (two items). Examples of feelings on the POMS are "anxious," "sad," "angry," "happy," "lively," and "calm." The clients were asked to evaluate how they felt during the session on a 5-point Likert scale ranging from "Not at all" to "Extremely." The POMS has been tested on college students and was found to be both valid and reliable (Guadagnoli & Mor, 1989). In line with previous studies that have implemented this measure (e.g., Sened et al., 2017), an aggregated total score of positive and negative affect was used in this study. The POMS showed excellent internal consistency in our sample for both the negative ($\alpha = .9$) and positive ($\alpha = .94$) scales.

Transcription

To capture the evolution of treatment from session to session, and since transcription is highly expensive, transcriptions were made every other session (i.e., Sessions 2, 4, 6, 8, etc.). When the material was incomplete (e.g., as a result of low recording quality or failure to complete questionnaires for a specific session), 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, where the transcribers were instructed to replace names with pseudonyms and to mask any other identifying information. The transcription protocol followed general guidelines as prescribed by Mergenthaler and Stinson (1992) as well as Albert et al. (2013). The audiotapes were transcribed in their entirety and provided verbatim accounts of the sessions. An average of 11.79 sessions were transcribed per client ($SD = 3.08$). Each transcript incorporated metadata such as the client's code, which allowed the client data to be linked across sessions and facilitated hierarchical analysis. The transcriptions totaled approximately 5 million words. On average, there were 5,842 words in a session, of which 4,525 (77%; $SD = 1407.07$; range 416–8,176) were client utterances and 1,317 (23%; $SD = 728.12$; range 160–6,048) were therapist utterances.

Emotion Coding

A subsample of 196 sessions was coded speech-turn by speech-turn to identify the emotional valence (positive, negative, mixed, neutral). This categorization of emotions is common in many studies (e.g., Greenberg, 2012; Tanana et al., 2021). Twenty undergraduate students trained by a clinician during a semester course served as the coders. Their training consisted of six class meetings where they were introduced to the guidelines for coding and given the opportunity to label sessions and receive feedback from a clinician on the accuracy of their labeling. Afterward, they began to label independently. Naïve labelers who received relatively short training were used because previous studies suggest that they are viable alternatives for identifying basic aspects of emotions such as valence; furthermore, they require less training than expert coders (e.g., Tanana et al., 2021; Waldinger et al., 2004). Out of the 196 sessions, 22 (11%) were coded twice, once by a trained undergraduate annotator and once by a doctoral student in clinical psychology. This led to a moderate Cohen's Kappa of 0.54 (0.59 for negative emotions and 0.52 for positive emotions). This average interrater reliability is similar to interrater reliabilities reported in previous studies (e.g., Tanana et al., 2021). In what follows, these 196 sessions annotated by human coders is termed the "gold data set," which comprised 22,248 client utterances. The remainder of the nonannotated sessions are referred to below as the "silver data set," which comprised 116,813 client utterances.

Then, the verbally expressed emotions were calculated on the silver data set. To neutralize the effect of the total number of utterances in a session, we calculated the relative proportion of each emotion. For example, to calculate the negative verbally expressed emotions in a given session, the proportion of negative utterances out of the total number of utterances in the session was calculated.

Procedure

The procedure was part of the routine battery in the clinic. All sessions were audiotaped and transcribed according to a protocol ensuring confidentiality and the masking of any identifying information such as names and places. Finally, to guarantee privacy given the sensitive nature of the data, only secured servers with privileged access were employed.

The session questionnaires were electronically completed by the participants using computers located in the clinic rooms and software that time-stamped their responses. The clients completed the ORS before each therapy session and the POMS at the end of each session.

Data Analysis Strategy

The first step was to automatically label the client emotions at the utterance level. The data set had two speakers; namely, the client (C) and the therapist (T). An *utterance* was defined as the shortest continuous unit of speech in a dialogue marked by a pause or change in speaker at the beginning or at the end. Formally, given an input sequence of N utterances $[u_1^p, u_2^p, \ldots, u_N^p]$ spoken by party p = [C, T], where each utterance $u_i^p = [u_{i1}, u_{i2}, \ldots, u_{iT}]$ has T words u_{ij}, the task was to label emotions only for client utterances $[u_1^C, u_2^C, \ldots, u_N^C]$.

To develop the emotion recognition labels at utterance level, we fine-tuned various BERT-based language models and their corresponding lightweight adapter solutions (Houlsby et al., 2019). Compared to fully fine-tuned models, adapter models only incorporate a few task-specific parameters for each new task. The BERT-based experiments were conducted on three pretrained language models, each capable of handling Hebrew text, that is, (a) XLM-RoBERTa (Conneau et al., 2019), a multilingual language model based on the RoBERTa architecture (Liu et al., 2019), (b) HeBERT (Chriqui & Yahav, 2021), a monolingual BERT model trained on Hebrew data, and (c) AlephBERT (Seker et al., 2022), another monolingual BERT-based model trained on a large Hebrew vocabulary of 52K tokens optimized via masked-token prediction. Corresponding variants of these BERT models with lightweight adapter solutions focused on a small number of task-specific parameters for training using bottleneck adapters (Houlsby et al., 2019) and mix-and-match (MAM) adapters (He et al., 2021). A complete list of the experiments for emotion recognition can be found in Table 1.

To evaluate the performance of the approaches on the emotion recognition task, we first trained and tested the models on the expert-labeled gold data set. Each model was trained on the labeled data over 10-fold cross-validation to account for the variability in the results. Hyper-parameter tuning was conducted on the development data set (10% of the gold data set per fold). The learning rate was set to 2e–6 with L2 regularization (decay = 0.0001). The maximum token size T per utterance was set to 128 and partial class balance was implemented given the skewed emotion label distribution (Chawla et al., 2002). In line with previous emotion recognition studies in psychotherapy (e.g., Tanana et al., 2021), we used the F1 micro to evaluate the results from the trained model. The F1 score calculates the harmonic mean of precision and recall and provides a single value of a model's performance (Hossin & Sulaiman, 2015). Precision is defined as the proportion of predicted instances that are truly relevant, and recall is the proportion of relevant instances that are correctly predicted (Powers, 2020).

Table 1
F1-Micro and Cohen's Kappa Evaluation of the Different BERT-Based Models for the Emotion Recognition Task on the Gold Data Set

Model	F1	Kappa
Annotators	0.73	0.54
XLM-ft	0.64	0.42
XLM-adapter	0.64	0.43
HB-senti-ft	0.61	0.37
HB-adapter	0.57	0.34
AB-ft	**0.66**	**0.46**
AB-adapter	0.65	0.44

Note. Cohen (1960) suggested interpreting Kappa results as follows: 0.01–0.20 none to slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement. The best model performance is highlighted in bold. XLM = XLM-RoBERTa; HB = HebrewBERT; AB = AlephBERT; BERT = bidirectional encoder representation from transformers

As presented in Table 1, the results from the experiments on the gold data set indicated that the AlephBERT-based emotion recognition model performed better than the other models in recognizing client emotions. We concluded that this model could be used for emotion recognition on unseen data sets with similar properties. Therefore, we adapted our gold data set-trained AlephBERT emotion recognition model to develop client emotion labels for the silver data set.

Next, we examined emotional coherence and its association with clients' improvement in functioning. The data set had a hierarchical structure, with session ratings nested within clients and clients nested within therapists. Thus, we used a multilevel model (MLM; Raudenbush & Bryk, 2002), with sessions at Level 1 and clients at Level 2.[1]

To test Hypothesis 2, our first model estimated the association between clients' self-reported emotions (SREs) and their verbally expressed emotions (VEEs), separately for positive and negative emotions.

Model 1

Level 1:

$$\text{SRE.NEG}_{\text{SC}} = \beta_{0c} + \beta_{1c} \times \text{VEE.NEG}_{\text{SC}} + e_{\text{sc}}. \quad (1)$$

or

$$\text{SRE.POS}_{\text{SC}} = \beta_{0c} + \beta_{1c} \times \text{VEE.POS}_{\text{SC}} + e_{\text{sc}}. \quad (2)$$

$$(e_{\text{sc}}) \sim N[(0, \sigma^2)]. \quad (3)$$

Level 2:

$$\beta_{0c} = \gamma_{00} + u_{0c}; \quad (4)$$

$$\beta_{1c} = \gamma_{10} + u_{1c}. \quad (5)$$

$$(u_{0d}, u_{1d}) \sim N\left[(00), \begin{pmatrix} \tau_{00} & \tau_{10} \\ \tau_{01} & \tau_{11} \end{pmatrix}\right]. \quad (6)$$

In this multivariate multilevel equation, the self-reported emotions of Client c in Session s were predicted by the following: the sample's average (i.e., the intercept γ_{00}), the client's verbally expressed emotions per session (i.e., the slope γ_{10}), the deviations of each client from the average intercepts and the slope (i.e., the Level-2 random effect for the intercept and the slope, u_{0c} and u_{1c}),

[1] When we attempted to estimate three-level models (i.e., taking therapist effects into account), the models did not converge. This is likely the result of the low rate of clients treated by the same therapist in the sample (most therapists treated only one client).

and the Level-1 residual terms quantifying the session's deviation from these effects (i.e., Level-1 random effect, e_{sc}).

To test Hypothesis 3 that emotional coherence amplifies the positive outcome trajectory across treatment, separately for negative and positive emotions, we first calculated the emotional coherence as the Pearson's correlation between verbal-expressed emotion and the self-reported emotion (Coherence.NEG_{0c}/Coherence.POS_{0c})[2] Next, we utilized a growth model to examine the linear trajectory of the session-level ORS throughout treatment. Finally, we added the interaction coefficient between the coherence and the ORS trajectory. This was done separately for positive and negative emotions.[3]

Model 2

Level 1:

$$\begin{aligned} ORS_{SC} = {} & \beta_{0c} + \beta_{1c} \times Session\_Number_{sc} + \beta_{2c} \\ & \times Coherence.NEG_{0c} + \beta_{3c} \times Session\_Number_{sc} \\ & \times Coherence.NEG_{0c} + e_{sc}. \end{aligned} \quad (7)$$

or

$$\begin{aligned} ORS_{SC} = {} & \beta_{0c} + \beta_{1c} \times Session\_Number_{sc} + \beta_{2c} \\ & \times Coherence.POS_{0c} + \beta_{3c} \times Session\_Number_{sc} \\ & \times Coherence.POS_{0c} + e_{sc}. \end{aligned} \quad (8)$$

$$(e_{sc}) \sim N[(0, \sigma^2)]. \quad (9)$$

Level 2:

$$\beta_{0c} = \gamma_{00} + u_{0c}; \quad (10)$$

$$\beta_{1c} = \gamma_{10}; \beta_{2c} = \gamma_{20}; \beta_{3c} = \gamma_{30}. \quad (11)$$

$$(u_{0c}) \sim N[(0, \tau_{00}^2)]. \quad (12)$$

Results

The hypotheses were a priori but not preregistered.

Automatic Labeling of Emotions (Prerequisite Hypothesis 1)

The complete set of results on the cross-validated gold data set using the emotion recognition approach based on BERT language models are presented in Table 1. The fine-tuned Hebrew language-based transformer model, viz. the AlephBERT-ft, was the best-performing approach out of the set of all baselines trained for this task. AlephBERT-ft scored a moderate F1 of 0.66 and had a Cohen's Kappa of 0.46, which was slightly lower but comparable to human performance on this task (human interrater reliability: F1 = 0.73; Kappa = 0.54). We then adapted this model to develop utterance-level emotion labels for the silver data set.

Assessment of Emotional Coherence (Hypothesis 2)

The descriptive statistics for the variables are presented in Table 2. The results of the first model are presented in Table 3. Consistent with Hypothesis 2, we found positive associations between self-reported emotions and verbally expressed emotions for both positive and negative emotions. In other words, clients' self-reported emotions (either positive or negative) were positively correlated with verbal expressions. The cross-valence associations indicated negative correlations between self-reported emotions and verbally expressed emotions for both positive and negative emotions (for the cross-valence associations, see Supplemental Material, Table 5). We ran an additional post hoc analysis to test for differences in clients' verbal expression and self-reported emotions between negative and positive emotions. We ran two MLM intercept-only models on the difference between positive and negative levels (positive–negative) in self-reported and verbally expressed emotions. On average, clients reported greater levels of positive than negative emotions (*Est.* = 4.14, $p < .001$), but there were greater proportions of verbally expressed negative than positive emotions (*Est.* = −0.28, $p < .001$).

The Association Between Emotional Coherence and Functioning Trajectory Across Treatment (Hypothesis 3)

The results of the second model are presented in Table 4. We found an average positive trajectory for functioning (ORS measures) throughout treatment. In addition, and in line with our hypothesis, this trajectory was moderated by the coherence of negative emotions. In other words, clients who had higher emotional coherence in negative emotions showed more improvement in functioning throughout treatment. However, this association was not found when examining the coherence of positive emotions.

To probe this moderation finding in negative emotions, we plotted (Figure 1) the simple associations between changes in ORS and session number (i.e., ORS trajectory throughout treatment). The results indicated that clients marked by higher coherence levels (+1 *SD*) were characterized by steeper ORS trajectories (*Est.* = 0.21, $p < .001$) compared to clients marked by lower coherence levels (−1 *SD*; *Est.* = 0.10, $p = .002$).

Discussion

In the present study, we used state-of-the-art emotion recognition language models to automatically label clients' emotions speech-turn by speech-turn, session-by-session throughout psychotherapy. Our goal was to leverage these models to test for coherence between clients' verbal expression of emotions and their subjective experience of emotions and whether this coherence corresponded to an improvement in clients' functioning.

The results of our prerequisite analysis showed that AlephBERT performed the best and achieved moderate accuracy in automatically labeling clients' emotions at the utterance level. AlephBERT's superior performance is consistent with the findings of previous NLP studies in which this model achieved state-of-the-art results

[2] To calculate the Pearson's correlation between verbal expression and self-reported emotions, we included treatments with at least 10 sessions.

[3] Level-2 random intercept model was utilized since the addition of random slopes to the model did not significantly improve the model fit.

Table 2
Means, Standard Deviations, and Intercorrelations for the Study Variables

Study variable	[M_{start}, M_{end}]	*SD*	Zero-order correlations				
			1	2	3	4	5
1. SRE.NEG	[5.83, 5.09]	2.33	—	−0.49 ($p < .001$)	0.18 ($p < .001$)	−0.26 ($p < .001$)	−0.28 ($p < .001$)
2. SRE.POS	[9.67, 9.45]	2.67		—	−0.25 ($p < .001$)	0.18 ($p < .001$)	0.36 ($p < .001$)
3. VEE.NEG	[0.39, 0.37]	0.14			—	−0.38 ($p < .001$)	−0.13 ($p < .001$)
4. VEE.POS	[0.11, 0.11]	0.06				—	0.15 ($p < .001$)
5. ORS	[21.80, 25.59]	8.18					—

Note. Means are presented for first sessions (M_{start}) and last sessions (M_{end}). Zero-order correlations applied the variable means computed across all treatment sessions. SRE.NEG/POS = session-level negative/positive self-reported emotions; VEE.NEG/POS = session-level negative/positive verbally expressed emotions; ORS = session-level Outcome Rating Scale; *SD* = standard deviation of the variables across all treatment sessions.

on benchmark data sets for different Hebrew NLP tasks such as categorical classification (Seker et al., 2022). While previous studies on emotional processes in psychotherapy tend to be based on relatively small sample sizes owing to the labor-intensive nature of human coding, or on self-reports that cannot capture the moment-to-moment expression of emotions within psychotherapy sessions, automatic labeling provides opportunities for examining emotional processes on a larger scale and higher specificity (~139,061 utterances of clients' labeled emotions from 872 sessions).

It is interesting to qualitatively explore examples in which the emotion recognition model misclassified the emotion category. One of the challenges for the model was the verbal ambiguity of Hebrew. For example, in the sentence "I'm dying on my dad, he had good intentions," the human annotators correctly labeled this as a positive emotion utterance since they identified the common Hebrew slang expression to express affection (in Hebrew slang, "I am dying on him" implies "I like him very much"). However, its intended emotion was missed by the automatic emotion recognition model, mostly likely because of the negative connotation of "dying." The model also tended to misclassify emotions in complex sentences that included the negation of emotion. For example, in the sentence "I did not have a good trip, but in the lab everyone thought that I enjoyed it and had fun," the human annotator correctly identified the negative emotion, but the model misclassified it as positive. NLP studies have noted that perfect performance should not be expected from emotion recognition models, given that even humans do not completely agree on these types of ratings (Tanana et al., 2021). However, as indicated by the results, the model approached human performance.

The automatic labeling of emotions allowed us to assess in the next step the level of coherence between verbally expressed and self-reported emotions. In line with Hypothesis 2, the findings indicated emotional coherence between verbally expressed emotions and self-reported emotions for both positive and negative emotions. This is consistent with previous reports indicating coherence between self-reported emotions and other channels of emotional response such as physiology or facial expression (e.g., Lohani et al., 2018). It, however, extends these results by showing coherence between emotional experience and emotional expression. In addition, whereas previous research has assessed coherence at only a few time points (e.g., Hastings et al., 2009), the present study investigated emotional coherence that occurs session-by-session across therapy. The significant association between self-reported emotions and verbally expressed emotions further supports the capacity of the automatic emotion recognition model to accurately recognize individuals' emotional states.

Table 3
Fixed and Random Effects of Clients' Self-Reported Emotions and Verbally Expressed Emotions

Study parameter	Negative emotions				Positive emotions			
	Est. (*SE*)	[CI 95%]	*p*	Std. Est.	Est. (*SE*)	[CI 95%]	*p*	Std. Est.
Fixed effects								
Intercept (γ_{00})	4.53 (0.26)	[4.03, 5.04]	<.001		8.79 (0.29)	[8.23, 9.35]	<.001	
VEE (γ_{10})	0.01 (0.001)	[0.01, 0.02]	<.001	0.15	0.06 (0.01)	[0.05, 0.08]	<.001	0.22
Random effects								
Level 1 (sessions)								
Residual	1.50	[1.42, 1.59]			1.61	[1.52, 1.69]		
Level 2 (clients)								
Intercepts	2.05	[1.61, 2.61]			2.11	[1.70, 2.62]		
VEE	0.074	[0.05, 0.10]			0.04	[0.02, 0.08]		

Note. Est. = estimate; SE = standard error; CI = confidence interval; Std. Est. = standard estimate; VEE = verbally expressed emotions.

Table 4

Fixed Effect of Clients' Coherence and Session Number as Predictors for Clients' Self-Reported Functioning Level (ORS Score)

Study parameter	Negative emotions				Positive emotions			
	Est. (*SE*)	[CI 95%]	*p*	Std. Est.	Est. (*SE*)	[CI 95%]	*p*	Std. Est.
Intercept (γ_{00})	22.32 (0.96)	[20.44, 24.20]	<.001		22.39 (0.97)	[20.49, 24.29]	<.001	
Session number (γ_{10})	0.16 (0.02)	[0.12, 0.20]	<.001	0.021	0.16 (0.02)	[0.11, 0.20]	<.001	0.020
Coherence (γ_{20})	0.83 (3.32)	[−5.84, 7.50]	.804	0.032	1.61 (3.16)	[−4.75, 7.97]	.614	0.061
Coherence × Session Number (γ_{30})	0.19 (0.08)	[0.04, 0.35]	.014	0.007	0.01 (0.08)	[−0.14, 0.15]	.940	0

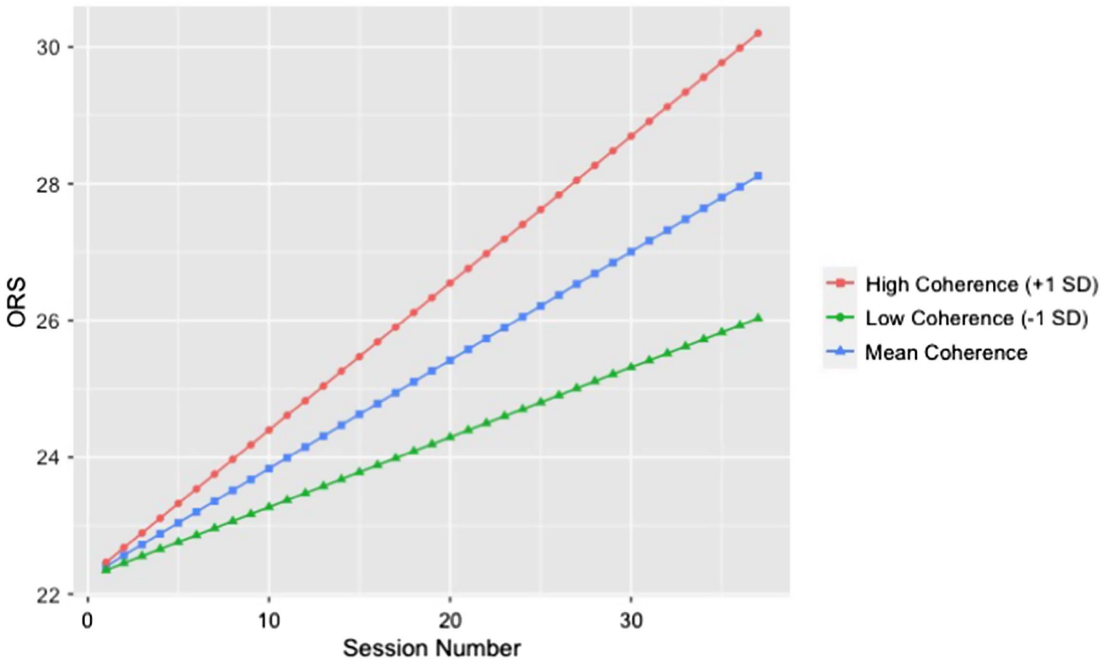
Note. ORS = Outcome Rating Scale; Est. = estimate; *SE* = standard error; CI = confidence interval; Std. Est. = standard estimate.

The finding partially supported the third hypothesis associating higher coherence between negative verbally expressed emotions and self-reported emotions with greater improvement in functioning. This finding strengthens various psychotherapy theories that emphasize the importance of coherence between subjective experience and verbal expression of emotions (e.g., Greenberg, 2012; Lane et al., 2022). The finding is also in line with previous studies that demonstrated association between emotional coherence and functioning levels (e.g., Leonhardt et al., 2018) but goes one step further by showing that the level of emotional coherence is associated with improvement in functioning levels over the course of psychotherapy. This finding highlights the importance of coherence between verbal expression of emotions and subjective experience of emotions to treatment outcome. However, this conclusion is drawn from correlational data, which does not allow for a definitive inference of a causal relationship between emotional coherence and treatment outcome. An alternative interpretation could be that individuals whose functioning improved over time may have managed to better align their verbal expressions and their emotional experiences.

It is interesting to speculate why emotional coherence was only linked to improvement in functioning for negative emotions, but not for positive emotions. This contrasts with a lab study that reported that higher emotional coherence in positive emotions was associated with higher functioning levels (Mauss et al., 2011). One potential explanation is that during psychotherapy, both therapists and clients often focus on negative emotions and tend to overlook positive emotions (e.g., Atzil-Slonim et al., 2019). Because clients often enter therapy with painful emotions, these emotions may take center stage. This may facilitate coherence between expressing and experiencing negative emotions, which may lead to enhanced functioning. In contrast, the tendency to neglect positive emotions

Figure 1

ORS Trajectory Over the Course of Treatment Moderated by Clients' Coherence Levels



Note. Trajectories of functioning (ORS score) improvement throughout treatment (session number). Three trajectories are shown, differentiated by clients' coherence level. This demonstrated the amplified functioning improvement for higher coherence levels for negative emotions. ORS = Outcome Rating Scale. See the online article for the color version of this figure.

might limit opportunities to align their expression and experience, hindering potential therapeutic benefits.

Support for this explanation comes from an interesting observation that was not part of our set of hypotheses but emerged from the data; namely, that self-reported emotions had higher levels of positive emotions than negative emotions, whereas verbal expressions evidenced higher levels of negative emotions than positive emotions. This may be linked to the fact that people tend to express negative emotions more frequently than positive emotions (Baumeister et al., 2001) and to give socially desirable responses in self-reports that present themselves positively (Kazdin, 2008). This implies that therapists might benefit from being more receptive to subtle expressions of positive emotions. By helping clients express both negative and positive emotions in alignment with their inner experience, clients may become more adept at communicating and managing their emotions effectively.

Limitations, Future Directions, and Clinical Implications

One limitation of this study is related to the fact that the positive and negative valence scores were aggregated for verbally expressed emotions and self-reported emotions. Findings have indicated the importance of individuals' ability to differentiate between specific negative emotions (e.g., sad vs. anxious; Schoebi & Randall, 2015), such that emotional coherence for specific emotions is likely to be meaningful. Given the finding of medium-to-large associations among the specific emotions within each valence in the POMS, we opted to use aggregated scores to adjust the scales between self-reported emotions and verbally expressed emotions. Future studies would benefit from examining whether coherence in specific emotions is differently associated with treatment outcomes.

Another limitation is that the analysis assessed coherence at the session level and focused on the association between verbally expressed emotions and self-reported emotions. Recent technological advances in audio analysis and facial recognition now permit a more fine-grained analysis of different emotional responses. Future studies would benefit from using multimodal measures with a higher time resolution to examine emotional coherence between multiple emotional channels.

The performance of the automatic coding model in this study was relatively modest, although it was comparable to previous studies that have used computational learning methods for similar purposes (e.g Tanana et al., 2021). It is possible that new developments in NLP and larger language models will enable better model performance in the future. Yet another limitation lies in the fact that while the data contained a vast number of utterances, the sample size at the client level was relatively small, thereby limiting its statistical power. Nevertheless, the use of automatic methods facilitates swift coding, thus opening up the possibility for future studies to harness these capabilities effectively for a larger data set.

Another limitation is that our psychotherapy data were in Hebrew and that the study sample was relatively homogeneous, since it mostly consisted of native Hebrew speakers. In addition, treatments were conducted in a clinic that emphasizes a psychodynamic model of treatment. These factors may limit the generalizability of the results to other languages, cultures, and treatment models. Although we consider emotional coherence to be a pan-theoretical component emphasized by most psychotherapy orientations, future studies are required to explore whether the association between emotional coherence and outcome can be replicated with therapists implementing other therapeutic orientations, in other languages and with more diverse samples.

Finally, in the present study, we focused on the clients' emotional processes. However, recent studies have highlighted the importance of studying emotions as a dyadic system (Atzil-Slonim et al., 2018). Accordingly, future studies could explore whether therapists' emotional coherence in different emotional channels is associated with clients' likelihood of emotional coherence over the course of therapy.

These results have several clinical implications. They highlight the importance of helping clients to align their emotional experiences with the language they use to express them. Therapists would benefit from identifying clients or sessions characterized by a high dissociation between emotional experience and the verbal expression of emotions and tailor their interventions to help clients express their emotions or increase their emotional awareness. While our results indicated that emotional coherence primarily correlated with favorable outcomes for negative emotions, the observation that clients rarely express positive emotions during sessions suggests that therapists should place greater emphasis on exploring positive emotions. Automatic emotion recognition models can be integrated into existing feedback systems to provide an indication of the levels of emotional coherence in psychotherapy sessions and allow therapists to modify their interventions accordingly.

The present study employed transformer-based emotion recognition models to automatically annotate clients' emotions, which served to investigate the role of emotional coherence in psychotherapy. This automated labeling approach can also examine many other emotional processes in psychotherapy. Given the rapid pace of technological progress in NLP emotion recognition models, even more advanced methods, such as generative large language models (LLMs), may be utilized in the near future to enhance the accuracy of emotion detection and explore subtle emotional processes in psychotherapy on a larger scale.

References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, *31*(1), 92–116. https://doi.org/10.1080/10503307.2020.1808729

Albert, A., MacWhinney, B., Nir, B., & Wintner, S. (2013). The Hebrew CHILDES corpus: Transcription and morphological analysis. *Language Resources and Evaluation*, *47*(4), 973–1005. https://doi.org/10.1007/s10579-012-9214-z

Atzil-Slonim, D., Bar-Kalifa, E., Fisher, H., Lazarus, G., Hasson-Ohayon, I., Lutz, W., Rubel, J., & Rafaeli, E. (2019). Therapists' empathic accuracy toward their clients' emotions. *Journal of Consulting and Clinical Psychology*, *87*(1), 33–45. https://doi.org/10.1037/ccp0000354

Atzil-Slonim, D., Bar-Kalifa, E., Fisher, H., Peri, T., Lutz, W., Rubel, J., & Rafaeli, E. (2018). Emotional congruence between clients and therapists and its effect on treatment outcome. *Journal of Counseling Psychology*, *65*(1), Article 51. https://doi.org/10.1037/cou0000250

Bar-Kalifa, E., & Sened, H. (2020). Using network analysis for examining interpersonal emotion dynamics. *Multivariate Behavioral Research*, *55*(2), 211–230. https://doi.org/10.1080/00273171.2019.1624147

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Beck, A. T. (Ed.). (1979). *Cognitive therapy of depression*. Guilford press.

Brown, C. L., Van Doren, N., Ford, B. Q., Mauss, I. B., Sze, J. W., & Levenson, R. W. (2020). Coherence between subjective experience and physiology in emotion: Individual differences and implications for well-being. *Emotion*, *20*(5), 818–829. https://doi.org/10.1037/emo0000579

Cao, J., Tanana, M., Imel, Z. E., Poitras, E., Atkins, D. C., & Srikumar, V. (2019). *Observing dialogue in therapy: Categorizing and forecasting behavioral codes*. arXiv preprint arXiv:1907.00326. https://doi.org/10.18653/v1/P19-1563

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chriqui, A., & Yahav, I. (2021). *HeBERT & HebEMO: A Hebrew BERT model and a tool for polarity analysis and emotion recognition*. arXiv preprint arXiv:2102.01909.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Unsupervised cross-lingual representation learning at scale*. arXiv preprint arXiv:1911.02116.

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (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. https://doi.org/10.1177/0146167206287721

Delgadillo, J., & Atzil-Slonim, D. (2023). Artificial intelligence, machine learning and mental health. In H. S. Friedman & C. H. Markey (Eds.), *Encyclopedia of mental health* (3rd ed., pp. 132–142). Elsevier. https://doi.org/10.1016/B978-0-323-91497-0.00177-6

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S., & Blackwell, A. D. (2021). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, *31*(3), 300–312. https://doi.org/10.1080/10503307.2020.1788740

Fisher, H., Atzil-Slonim, D., Bar-Kalifa, E., Rafaeli, E., & Peri, T. (2016). Emotional experience and alliance contribute to therapeutic change in psychodynamic therapy. *Psychotherapy*, *53*(1), 105–116. https://doi.org/10.1037/pst0000041

Fosha, D. (2001). The dyadic regulation of affect. *Journal of Clinical Psychology*, *57*(2), 227–242. https://doi.org/10.1002/1097-4679(200102)57:2<227::AID-JCLP8>3.0.CO;2-1

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Greenberg, L. S. (2012). Emotions, the great captains of our lives: Their role in the process of change in psychotherapy. *American Psychologist*, *67*(8), 697–707. https://doi.org/10.1037/a0029858

Gross, J. J., John, O. P., & Richards, J. M. (2000). The dissociation of emotion expression from emotion experience: A personality perspective. *Personality and Social Psychology Bulletin*, *26*(6), 712–726. https://doi.org/10.1177/0146167200268006

Guadagnoli, E., & Mor, V. (1989). Measuring cancer patients' affect: Revision and psychometric properties of the profile of mood states (POMS). *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*(2), 150–154. https://doi.org/10.1037/1040-3590.1.2.150

Hastings, P. D., Nuselovici, J. N., Klimes-Dougan, B., Kendziora, K. T., Usher, B. A., Ho, M. H. R., & Zahn-Waxler, C. (2009). Dysregulated coherence of subjective and cardiac emotional activation in adolescents with internalizing and externalizing problems. *Journal of Child Psychology and Psychiatry*, *50*(11), 1348–1356. https://doi.org/10.1111/j.1469-7610.2009.02159.x

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., & Neubig, G. (2021). *Towards a unified view of parameter-efficient transfer learning*. arXiv preprint arXiv:2110.04366.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). *Parameter-efficient transfer learning for NLP* [Conference session]. International Conference on Machine Learning.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist*, *63*(3), 146–159. https://doi.org/10.1037/0003-066X.63.3.146

Kramer, U., Pascual-Leone, A., Despland, J. N., & de Roten, Y. (2015). One minute of grief: Emotional processing in short-term dynamic psychotherapy for adjustment disorder. *Journal of Consulting and Clinical Psychology*, *83*(1), 187–198. https://doi.org/10.1037/a0037979

Lane, R. D., Ryan, L., Nadel, L., & Greenberg, L. (2015). Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, *38*, Article e1. https://doi.org/10.1017/S0140525X14000041

Lane, R. D., Subic-Wrana, C., Greenberg, L., & Yovel, I. (2022). The role of enhanced emotional awareness in promoting change across psychotherapy modalities. *Journal of Psychotherapy Integration*, *32*(2), 131–150. https://doi.org/10.1037/int0000244

Leonhardt, B. L., Lysaker, P. H., Vohs, J. L., James, A. V., & Davis, L. W. (2018). The experience and expression of anger in posttraumatic stress disorder: The relationship with metacognition. *Journal of Mental Health*, *27*(5), 432–437. https://doi.org/10.1080/09638237.2018.1466036

Levenson, R. W. (2003). Blood, sweat, and fears: The autonomic architecture of emotion. In P. Ekman, J. J. Campos, R. J. Davidson, & F. B. M. de Waal (Eds.), *Emotions inside out* (pp. 348–366). The New York Academy of Sciences. https://doi.org/10.1196/annals.1280.016

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692.

Lohani, M., Payne, B. R., & Isaacowitz, D. M. (2018). Emotional coherence in early and later adulthood during sadness reactivity and regulation. *Emotion*, *18*(6), 789–804. https://doi.org/10.1037/emo0000345

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175–190. https://doi.org/10.1037/1528-3542.5.2.175

Mauss, I. B., Shallcross, A. J., Troy, A. S., John, O. P., Ferrer, E., Wilhelm, F. H., & Gross, J. J. (2011). Don't hide your happiness! Positive emotion dissociation, social connectedness, and psychological functioning. *Journal of Personality and Social Psychology*, *100*(4), 738–748. https://doi.org/10.1037/a0022410

McNair, D. M., Lorr, M., & Droppleman, L. F. (1992). *POMS manual—Profile of mood questionnaire*. Educational and Industrial Testing Services.

Mergenthaler, E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic

processes. *Psychotherapy Research*, *18*(2), 109–126. https://doi.org/10.1080/10503300701883741

Mergenthaler, E., & Stinson, C. H. (1992). Psychotherapy transcription standards. *Psychotherapy Research*, *2*(2), 125–142. https://doi.org/10.1080/10503309212331332904

Miller, S. D., Duncan, B. L., Brown, J., Sparks, J. A., & Claud, D. A. (2003). The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy*, *2*, 91–100.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), Article 81. https://doi.org/10.1007/s13278-021-00776-6

Negrao, C., II, Bonanno, G. A., Noll, J. G., Putnam, F. W., & Trickett, P. K. (2005). Shame, humiliation, and childhood sexual abuse: Distinct contributions and emotional coherence. *Child Maltreatment*, *10*(4), 350–363. https://doi.org/10.1177/1077559505279366

Pascual-Leone, A. (2018). How clients "change emotion with emotion": A programme of research on emotional processing. *Psychotherapy Research*, *28*(2), 165–182. https://doi.org/10.1080/10503307.2017.1349350

Pos, A. E., Greenberg, L. S., & Warwar, S. H. (2009). Testing a model of change in the experiential treatment of depression. *Journal of Consulting and Clinical Psychology*, *77*(6), 1055–1066. https://doi.org/10.1037/a0017059

Pos, A. E., Paolone, D. A., Smith, C. E., & Warwar, S. H. (2017). How does client expressed emotional arousal relate to outcome in experiential therapy for depression? *Person-Centered and Experiential Psychotherapies*, *16*(2), 173–190. https://doi.org/10.1080/14779757.2017.1323666

Powers, D. M. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*. arXiv preprint arXiv:2010.16061.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE Publications.

Schoebi, D., & Randall, A. K. (2015). Emotional dynamics in intimate relationships. *Emotion Review*, *7*(4), 342–348. https://doi.org/10.1177/1754073915590620

Seker, A., Bandel, E., Bareket, D., Brusilovsky, I., Greenfeld, R., & Tsarfaty, R. (2022). *AlephBERT: Language model pre-training and evaluation from sub-word to sentence level* [Conference session]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers), Dublin, Ireland. https://doi.org/10.18653/v1/2022.acl-long.4

Sened, H., Yovel, I., Bar-Kalifa, E., Gadassi, R., & Rafaeli, E. (2017). Now you have my attention: Empathic accuracy pathways in couples and the role of conflict. *Emotion*, *17*(1), 155–168. https://doi.org/10.1037/emo0000220

Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. In R. A. Levy, J. S. Ablon, & H. Kachele (Eds.), *Psychodynamic psychotherapy research: Evidence-based practice and practice-based evidence* (pp. 9–25). Humana Press.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59*(Suppl. 20), 22–33.

Summers, R. F., & Barber, J. P. (2010). *Psychodynamic therapy: A guide to evidence-based practice*. Guilford Press.

Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior Research Methods*, *53*(5), 2069–2082. https://doi.org/10.3758/s13428-020-01531-z

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Wagner, A. W., Roemer, L., Orsillo, S. M., & Litz, B. T. (2003). Emotional experiencing in women with posttraumatic stress disorder: Congruence between facial expressivity and self-report. *Journal of Traumatic Stress*, *16*(1), 67–75. https://doi.org/10.1023/A:1022015528894

Waldinger, R. J., Schulz, M. S., Hauser, S. T., Allen, J. P., & Crowell, J. A. (2004). Reading others emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. *Journal of Family Psychology*, *18*(1), 58–71. https://doi.org/10.1037/0893-3200.18.1.58

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2005). *Schema therapy: A practitioner's guide*. Guilford Press. https://doi.org/10.1007/978-90-313-7121-1

Received March 13, 2023
Revision received October 30, 2023
Accepted October 30, 2023 ■