Decoding emotions: exploring the validity of sentiment analysis in psychotherapy

Steffen T. Eberhardt, Jana Schaffrath, Danilo Moggia, Brian Schwartz, Martin Jaehde, Julian A. Rubel, Tobias Baur, Elisabeth André, Wolfgang Lutz

Angaben zur Veröffentlichung / Publication details:

Eberhardt, Steffen T., Jana Schaffrath, Danilo Moggia, Brian Schwartz, Martin Jaehde, Julian A. Rubel, Tobias Baur, Elisabeth André, and Wolfgang Lutz. 2025. "Decoding emotions: exploring the validity of sentiment analysis in psychotherapy." *Psychotherapy Research* 35 (2): 174–89. https://doi.org/10.1080/10503307.2024.2322522.

Routledge
Taylor & Francis Group

Psychotherapy Research

ISSN: 1050-3307 (Print) 1468-4381 (Online) Journal homepage: www.tandfonline.com/journals/tpsr20

Decoding emotions: Exploring the validity of sentiment analysis in psychotherapy

Steffen T. Eberhardt, Jana Schaffrath, Danilo Moggia, Brian Schwartz, Martin Jaehde, Julian A. Rubel, Tobias Baur, Elisabeth André & Wolfgang Lutz

To cite this article: Steffen T. Eberhardt, Jana Schaffrath, Danilo Moggia, Brian Schwartz, Martin Jaehde, Julian A. Rubel, Tobias Baur, Elisabeth André & Wolfgang Lutz (2025) Decoding emotions: Exploring the validity of sentiment analysis in psychotherapy, Psychotherapy Research, 35:2, 174-189, DOI: 10.1080/10503307.2024.2322522

To link to this article: https://doi.org/10.1080/10503307.2024.2322522

Psychotherapy Research, 2025
Vol. 35, No. 2, 174–189, https://doi.org/10.1080/10503307.2024.2322522

RESEARCH ARTICLE

Decoding emotions: Exploring the validity of sentiment analysis in psychotherapy

STEFFEN T. EBERHARDT[1], JANA SCHAFFRATH[1], DANILO MOGGIA[1], BRIAN SCHWARTZ[1], MARTIN JAEHDE[1], JULIAN A. RUBEL[2], TOBIAS BAUR[3], ELISABETH ANDRÉ[3], & WOLFGANG LUTZ[1]

[1]*Trier University, Trier, Germany;* [2]*Osnabrück University, Osnabrück, Germany &* [3] *Augsburg University, Augsburg, Germany*

(Received 1 March 2023; revised 19 January 2024; accepted 15 February 2024)

Abstract
Objective Given the importance of emotions in psychotherapy, valid measures are essential for research and practice. As emotions are expressed at different levels, multimodal measurements are needed for a nuanced assessment. Natural Language Processing (NLP) could augment the measurement of emotions. The study explores the validity of sentiment analysis in psychotherapy transcripts.
Method We used a transformer-based NLP algorithm to analyze sentiments in 85 transcripts from 35 patients. Construct and criterion validity were evaluated using self- and therapist reports and process and outcome measures via correlational, multitrait-multimethod, and multilevel analyses.
Results The results provide indications in support of the sentiments' validity. For example, sentiments were significantly related to self- and therapist reports of emotions in the same session. Sentiments correlated significantly with in-session processes (e.g., coping experiences), and an increase in positive sentiments throughout therapy predicted better outcomes after treatment termination.
Discussion Sentiment analysis could serve as a valid approach to assessing the emotional tone of psychotherapy sessions and may contribute to the multimodal measurement of emotions. Future research could combine sentiment analysis with automatic emotion recognition in facial expressions and vocal cues via the Nonverbal Behavior Analyzer (NOVA). Limitations (e.g., exploratory study with numerous tests) and opportunities are discussed.

Keywords: emotions; sentiment analysis; natural language processing (NLP); transcripts; multimodal measurement

Clinical or methodological significance of this article: The study provides several indications in support of the validity of sentiment analysis for measuring emotions in psychotherapy session transcripts. It also highlights the importance of multimodal measurement to gain more in-depth knowledge of emotional processes during treatment and the potential sentiment analysis could have for measurement-based and data-informed psychotherapy.

Emotions play a crucial role in mental health and well-being (Houben et al., 2015; Kraiss et al., 2020). Emotional distress is a primary reason for seeking psychological help, and interventions that promote emotion regulation and coping are key components of psychotherapy (Whelton, 2004). In order to effectively treat emotional disorders (Bullis et al., 2019), it is essential to assess and monitor emotions

Correspondence concerning this article should be addressed to STEFFEN T. EBERHARDT, Trier University, Am Wissenschaftspark 25 +27, Trier 54296, Germany. Email: eberhardt@uni-trier.de

during therapy (Greenberg, 2012; Peluso & Freund, 2018). In precision mental health care, continuous measurement and feedback of emotions can inform clinical decision-making and improve outcomes (Lutz et al., 2021). However, accurately and efficiently assessing emotions in psychotherapy remains a challenge.

As emotions are expressed on multiple levels, including facial expression (Ekman & Friesen, 2003), physiology (Kreibig, 2010), behavior (Tamir & Bigman, 2018), and language (Koolagudi & Rao, 2012; Shapira et al., 2021), multimodal measures are necessary for a comprehensive assessment (Eid et al., 2014; Lutz et al., 2021; Scherer, 2005). Despite the widespread application of multimodal approaches to assessing emotions in several domains (Guo et al., 2014; Soleymani et al., 2011), they have been widely overlooked in psychotherapy research (Eid et al., 2014). Multimodal approaches in psychology seek a more complete and nuanced understanding of psychological phenomena by combining various aspects, dimensions, components, or modalities using diverse methods (D'Mello & Kory, 2015; Kazdin, 1996). While in communication and computer science, multimodality primarily refers to multiple human communicative channels, in psychotherapy research, multimodal measurements may also involve combining different methods and sources (Lutz et al., 2021). Traditional emotion measures, such as self-reports (e.g., Profile of Mood States, POMS; McNair et al., 1992), therapist reports (e.g., Atzil-Slonim et al., 2019), observer ratings (e.g., Facial Action Coding System, FACS; Ekman & Friesen, 2003), or physiological data (e.g., Tschacher & Meier, 2020; Bar-Kalifa et al., 2019), have provided valuable insights into emotional experiences in therapy. However, these traditional methods also have drawbacks, such as self-report bias, reliance on trained observers, cost, and time intensity (Kihlstrom et al., 2000).

Integrating new approaches could help surpass the limitations of traditional methods. Machine learning has been useful in mental health research (e.g., mental health detection: Abd Rahman et al., 2020; mental health outcome: Su et al., 2020; mobile mental health: Han et al., 2021; treatment recommendations: Lutz et al., 2022; for a review, see: Shatte et al., 2019) and could facilitate emotion measurement in psychotherapy (Tanana et al., 2021). For example, Terhürne et al. (2022) used the Nonverbal Behavior Analyzer (NOVA) to detect emotions from facial expressions in video-recorded psychotherapy sessions. Natural language processing (NLP) is another area of machine learning focused on analyzing and understanding human language. For example, Atzil-Slonim et al. (2021) used topic models to identify clients' levels of functioning and alliance ruptures, and Smink et al. (2019) employed text mining to relate change processes to therapeutic outcomes.

Sentiment analysis, a subfield of NLP, aims to classify the emotional tone expressed in texts as positive, negative, or neutral (Kumar et al., 2016; Mohammad, 2016; Yue et al., 2019). Since psychotherapy is primarily based on language, sentiment analysis could be a helpful tool to capture the rudimentary emotional tone within a therapy session. It can rapidly process large amounts of text and provide researchers with a vast and diverse dataset to study emotions (Nakayama et al., 2021; Syzdek, 2020). Sentiment analysis has been applied to written (e.g., chat messages) and transcribed (e.g., therapy sessions) texts. Provoost et al. (2019) found that sentiment analysis applied to online cognitive behavioral therapy texts performed nearly as well as human raters. Shapira et al. (2021) discovered associations between emotional words and the reported distress in psychotherapy session transcripts. Tanana et al. (2021) found that modern machine-learning methods trained on psychotherapy utterances performed better at predicting sentiment in session transcripts than dictionary-based methods and models trained in other domains such as novels, news articles, and social media posts. Despite a tradition of various qualitative methods and coding systems for the analysis of psychotherapy transcripts, these novel approaches are still in the early stages. Further research is needed to establish their validity in psychotherapy. The current study addresses this gap by exploring the validity of sentiment analysis within a wide nomological net of theoretically related constructs using a multimodal approach (Cronbach & Meehl, 1955). The goal was to examine the construct and criterion validity of the sentiment analysis in psychotherapy session transcripts.

The construct validity of sentiments in our study is theoretically supported by their potential alignment with patient self-reports of emotions and therapist ratings of patient emotions. If sentiments accurately capture the emotional tone of therapy sessions, we expect to find consistency between sentiments and patients' direct emotional self-reports. Additionally, sentiments should correspond with therapist ratings, which provide an external assessment of patient emotions.

Furthermore, examining the potential association between sentiments and specific in-session processes could offer critical insight into sentiment validity. Our study investigated three key in-session processes: interpersonal experiences, coping experiences, and affective experiences (Rubel et al., 2017). Interpersonal experiences assess the patient's perception of the therapeutic relationship, reflecting the quality of the alliance. It is conceivable that a strong therapeutic relationship, characterized by trust, emotional support, collaboration, and comfort, may foster open dialogue and the expression of positive emotions during therapy sessions (Fitzpatrick & Stalikas, 2008). Coping experiences, on the other hand, measure the patient's perceived ability to master challenges, emphasizing the corrective aspects of therapy. In theory, successful coping could be associated with positive sentiment due to the feelings of mastery, self-efficacy, achievement, and positivity often linked with effective coping (Lazarus & Folkman, 1984). Affective experiences capture moments when patients engage emotionally with their problems. Negative sentiments may potentially accompany these experiences, as engaging with distressing issues frequently results in their expression (Greenberg & Watson, 2006).

Finally, exploring the relationship between sentiments and therapy outcomes, including patients' levels of functioning and symptom severity, allows the validation of sentiments. If sentiments reflect the emotional tone of therapy sessions, they might be linked to patients' psychological well-being and broader psychotherapy outcomes (Leahy, 2008). Confirming sentiments' alignment with therapy outcomes would provide evidence supporting the construct validity of this analytical tool, further emphasizing its ability to capture the emotional tone of psychotherapy sessions and their potential links to patients' therapeutic progress.

The following hypotheses were examined:

> H1: Construct Validity - Sentiments and Emotion Measures: Patient sentiments obtained by sentiment analysis have significant correlations with emotion measures. Specifically, sentiments are associated with patients' self-reports of emotions in the same session: Either positive sentiments are positively associated or negative sentiments are negatively associated with self-reports of positive emotions (H1a). Correspondingly, either positive sentiments are negatively associated or negative sentiments are positively associated with patients' self-reports of negative emotions (H1b). The same pattern is expected for the associations between sentiments and therapist reports of their patients' emotions: Positive sentiments are positively associated or negative sentiments are negatively associated with therapist reports of positive emotions (H1c). Positive sentiments are negatively associated or negative sentiments are positively associated with therapist reports of negative emotions (H1d).

> H2: Criterion Validity - Sentiments and Process Measures: Patient sentiments obtained via sentiment analysis have significant correlations with psychotherapy process measures. Positive sentiments are positively or negative sentiments are negatively associated with interpersonal (H2a) and coping experiences (H2b), while positive sentiments are negatively or negative sentiments positively associated with problem-related affective experiences (H2c).

> H3: Criterion Validity - Sentiments and Outcome Measures: Patient sentiments have significant correlations with psychotherapy outcome measures. Positive sentiments are positively or negative sentiments negatively related to the level of functioning (H3a), while positive sentiments are negatively or negative sentiments positively associated with symptom severity (H3b). Moreover, an increase in positive sentiments predicts more functional emotion regulation or an increase of negative sentiment more dysfunctional emotion regulation, whereas an increase of positive sentiments predicts better or an increase of negative sentiment worse outcomes after treatment termination (H3c).

All associations with session-level measures will be examined on both the within- and between-patient level. The hypotheses will be disconfirmed if both positive and negative sentiments show no significant within- and between-patient associations with the respective validation scale.

Method

Patients and Therapists

We analyzed a sample of $N = 35$ patients ($M = 40$ years, $SD = 12.5$, range: 17–62) with 85 recorded and manually transcribed sessions. Patients were included in the study (a) if they and their therapists gave informed consent to the usage of their data for research purposes, (b) if at least one of their sessions was transcribed, and (c) if they reported their emotional experience for the transcribed sessions. No restrictions were made based on demographic variables or psychopathology. Patients had, on average, 2.5 ($SD = .89$, range: 1–4) transcribed sessions. Table S1 shows patient characteristics and diagnoses based on the Structured Clinical Interview for DSM-IV (SCID; First et al., 1995). The majority (85.7%) of patients were of German origin, and all therapy sessions were conducted in the German language. The patients were treated by 22 therapists ($M = 28$ years, $SD = 4$, range: 25–40, 86.4% female) who had a master's degree in psychology, were in a 3–5-year psychotherapy training program, and had at least 1.5 years of clinical experience. The mean number of patients per therapist was 1.6 ($SD = .83$, range: 1–4).

Treatment

Between 2017 and 2020, personalized integrative cognitive–behavioral therapies (CBT) were conducted at a university outpatient training and research clinic in Southwest Germany. Therapists had supervision and training in manualized treatments and transtheoretical concepts. The supervisors were licensed cognitive behavior therapists with a minimum of five years of clinical experience and advanced training in supervision, maintaining ongoing professional development and continuous education. Therapists could customize the treatment to each patient's progress and needs by incorporating different techniques and change principles (Lutz et al., 2023). They had access to a comprehensive feedback and navigation tool (Trier Treatment Navigator, TTN; Lutz et al., 2022), which provides pretreatment recommendations and progress monitoring. If a patient is not improving as expected, therapists can use clinical support tools suggested by the system. The Inventory of Therapeutic Interventions and Skills (ITIS, Boyle et al., 2019) was employed as an adherence measure to assess the application of CBT techniques and strategies in the sessions. Therapy sessions lasted 50 min and took place weekly. Patients received individual therapy with a mean of 25 sessions (*SD* = 15, range: 3–63). The dropout rate was 23.9%, compared to an overall rate of 22.6% in the clinic (Lutz et al., 2019).

Measures

Patients completed questionnaire batteries at intake and after treatment termination and short scales before and after each session as part of routine process and outcome monitoring at the outpatient clinic. Therapists answered brief questionnaires at the end of the session. The sessions were video-recorded using Telycam TLC-700-S-R cameras and Beyerdynamic BM 32 W microphones. The videos were used for transcription and adherence ratings by trained raters. The transcribers followed Mayring's (1990) transcription guidelines.

Emotion measures. Self- and therapist reports were used to assess patients' emotions with seven items based on the Profile of Mood States (POMS; McNair et al., 1992) at the end of each session. Patients and therapists were asked, "How [sad, ashamed, anxious, angry, content, energetic, relaxed] did [you / your patient] feel in today's session?" on a rating scale from 0 *(not at all)* to 100 *(extremely)*. The four scores for sadness, shame, anxiety, and anger were averaged to create a negative emotion score ($\text{POMS}_{\text{PAT, NEG}}$ & $\text{POMS}_{\text{THE, NEG}}$), while contentment, energy, and relaxation ratings were averaged to compute a positive emotion score for self- and therapist reports ($\text{POMS}_{\text{PAT, POS}}$ & $\text{POMS}_{\text{THE, POS}}$). The reliability in the current study ranged between $\alpha = .84$ and $\alpha = .92$ (see also Table S2).

Sentiment as a measure of the emotional tone of patients' statements in psychotherapy transcripts was assessed using the Multilingual Language Model Toolkit for Sentiment Analysis (XLM-T; Barbieri et al., 2022), a transformer-based NLP model derived from the Cross-lingual Language Model based on RoBERTa (XLM-R, Conneau et al., 2020). It is trained on 2.5 TB of multilingual text data and additionally fine-tuned for sentiment analysis on 198 million tweets in eight languages, including German. XLM-T was chosen due to its proficiency in German and potential adaptability to psychotherapy transcripts, supported by its extensive and diverse training data. It has demonstrated reliability and superiority over other baselines like RoBERTa-base and XLM-R, achieving an F1 score of 77.35% in German (Barbieri et al., 2022).

Sentiment analysis was conducted on patients' sentences, while therapists' statements were not analyzed. This prioritization was due to the assumed importance of patients' emotions in conjunction with therapy processes and outcomes. Annotations of non-verbal cues in the transcript (e.g., laughter and silences) were excluded to ensure the sentiment analysis was solely based on patients' natural language, eliminating any influence from observational or other non-verbal input. XLM-T estimated the probability of each patient sentence reflecting positive or negative sentiment, averaged to assess the overall positive and negative sentiment (SENT_{POS} & SENT_{NEG}) in each session.

Process measures. We chose patients' in-session experiences as representatives of therapeutic processes for validating the sentiments. Patients' in-session experiences were measured using a short version of the Bern Post-Session Report (BPSR; Flückiger et al., 2010) at the end of each session. Rubel et al. (2017) found three factors that were used in this study: Interpersonal Experiences (EXP_{INT}, 4 items), Coping Experiences (EXP_{COP}, 6 items), and Affective Experiences (EXP_{AFF}, 2 items). Interpersonal Experiences represent patients' perceived relationship quality with their therapists. Coping Experiences reflect patients' perceived corrective experiences regarding mastery and clarification. Affective Experience is a process in which patients' problems are addressed in an emotionally engaging way. The items were assessed on a Likert scale from –3 (*not at all*) to 3 (*yes, exactly*). They were averaged to obtain the final scores for each subscale. Reported internal consistencies were α = .89 (Coping Experience), α = .90 (Interpersonal Experience), and α = .85 (Affective Experience; Rubel et al., 2017). In the current study, reliability ranged between α = .85 and α = .94 (see also Table S2).

Outcome measures. The Hopkins Symptom Checklist-11 (HSCL-11; Lutz et al., 2006) is an 11-item self-report for the assessment of symptom severity administered at the start of each session. Patients were asked to what degree they experienced the eleven symptoms in the last seven days. Patients rated the symptoms on a Likert scale from 1 (*not at all*) to 4 (*extremely*). The mean score represents the patients' global level of symptom distress for the preceding week. The score is highly correlated with the Global Severity Index of the Brief Symptom Inventory (BSI; $r = .91$) and has a high internal consistency (α = .92; Lutz et al., 2006). The reliability in the current study was α = .93.

The Global Assessment of Functioning (GAF) is a commonly used rating scale for mental health assessment and treatment planning. Therapists rated their patients' functioning on a scale from 0 to 100 at the end of each session, with higher values reflecting better functioning in various domains (e.g., work, school, and relationships; Aas, 2010, 2011).

The Outcome Questionnaire-30 (OQ-30; Ellsworth et al., 2006) is a 30-item self-report outcome measure designed to assess various aspects of psychological functioning, such as subjective complaints, interpersonal relationships, and fulfillment of social roles. Patients responded on a 5-point Likert scale ranging from 0 (*never*) to 4 (*almost always*). The overall mean score is indicative of the level of psychological distress, with higher values indicating more constraints. The reliability in the current study was α = .96 at pre- and post-measurement.

The Patient Health Questionnaire 9 (PHQ-9; Löwe et al., 2002) is a widely used, 9-item self-report designed to assess the severity of depressive symptoms in the last two weeks (Kroenke et al., 2001). The items are answered on a Likert scale ranging from 0 (*not at all*) to 3 (*almost every day*). The total score depicts the overall distress caused by depressive symptoms. The PHQ-9 has demonstrated high reliability and validity in both clinical and research settings (Kroenke et al., 2001) and is commonly used as part of a comprehensive assessment in primary care and mental health settings (Spitzer et al., 2006). The reliability in the current study was α = .93 at pre- and α = .90 at post-measurement.

The Generalized Anxiety Disorder-7 (GAD-7; Löwe et al., 2007) is a 7-item self-report rating scale for screening generalized anxiety disorder and assessing the severity of symptoms. Patients rate how often the symptoms were present in the last two weeks on a Likert scale ranging from 0 (*not at all*) to 3 (*almost every day*). The GAD-7 has demonstrated high reliability and validity (Spitzer et al., 2006). The reliability in the current study was α = .93 at pre- and α = .87 at post-measurement.

The Affective Style Questionnaire (ASQ; Graser et al., 2012) assesses emotion regulation styles via 20 items related to suppression, reappraisal, and acceptance of emotions, each answered on a 1 (*not at all*) to 5 (*extremely*) Likert scale. A mean score was used, with the suppression subscale inversely coded, signifying that higher scores denote more functional emotion regulation styles. The reliability in the current study was α = .90 at pre- and post-measurement.

Data Analysis

Correlational, multitrait-multimethod, and multilevel analyses were used to analyze the sentiments' construct and criterion validity. The R code is available at the open science framework (OSF): https://doi.org/mc4h. Missing data were addressed using multiple imputations, employing the R package MICE (v3.16.0; Van Buuren & Groothuis-Oudshoorn, 2011). The three-level data structure was accounted for during the imputation process using the R package miceadds (v3.16-18; Robitzsch & Grund, 2023). Ten imputed datasets were generated, each with 30 iterations. Imputation was conducted independently of the sentiment variables. All subsequent analyses involving missing data were performed independently in each imputed dataset. For result aggregation, we used the mice::pool function. In cases where the mice::pool function was not applicable, we employed the pool_scalar_RR function from the R package miceafter (v0.5.0; Heymans & Twisk, 2022). This function pools estimates and standard errors and calculates test statistics and *p*-values for the pooled estimates following Rubin's Rules (Rubin, 1987).

Construct validity

Multitrait-multimethod (MTMM) analysis. The intent of the MTMM analysis (Campbell & Fiske, 1959) was to assess the convergent and discriminant validity of the sentiment analysis. Two traits (positive and negative emotions) were assessed with three methods (POMS$_{\text{PAT}}$, POMS$_{\text{THE}}$, and SENT). In the MTMM matrix, positive correlations in the monotrait-heteromethod blocks represent convergent validity, as they indicate the agreement between measures of the same trait using different assessment methods. In the heterotrait-heteromethod blocks, which contain correlations between measures of different traits using different assessment methods, negative correlations would support discriminant validity. To account for the nesting of sessions within patients, multilevel correlations were employed using the multilevel.cor function from the R package misty (v0.5.3; Yanagida, 2023) to compute within- and between-patient correlation matrices.

Hierarchical linear models (HLMs). To further account for the multilevel data structure, HLMs were fitted using the R package lme4 (v1.1-30; Bates et al., 2015). HLMs are appropriate to handle hierarchical data as they do not require the independence assumption and facilitate modeling relationships at within- and between-patient levels. Random-intercept random-slope HLMs were fitted to analyze the construct validity of the sentiments (2 sentiments: positive & negative × 2 self-reports: positive & negative × 2 therapist reports: positive & negative). Sentiments were used as predictors of self-reports and therapist reports in the same session. More specifically, patients' mean sentiment scores across sessions were calculated and employed as a level-2 predictor. Mean-centered sentiment scores were calculated by subtracting patients' average sentiment scores from their session-level scores and added as a level-1 predictor (see approach L3a; Hamaker & Muthén, 2020). We have visualized the within-patient effects based on these mean-centered sentiment scores using the R package ggplot2 (v3.3.6; Wickham, 2016). We used the model-implied intercepts and slopes specific to each patient to visualize the within-patient effects. The graphs are based on unstandardized scores, whereas the statistics reported in the text are based on scaled scores.

While patients were additionally nested within therapists, our study primarily focused on patients, and thus, nesting within therapists was not modeled in the HLMs. An advantage of HLMs is their compatibility with cluster-robust standard errors. This correction permits accurate estimation of the variability in the coefficient estimates given clustering within therapists, avoiding the necessity to explicitly model the nesting within the HLMs (McNeish et al., 2017). The approach aids in focusing the analysis on the patient while still accounting for potential variability introduced by different therapists. The cluster-robust standard errors were computed using the model_parameters function from the R package parameters (v0.21.1; Lüdecke et al., 2020). The results of the HLMs have priority over the results of the multilevel correlations in the testing of hypotheses.

Criterion validity

Correlational analysis. The sentiments were correlated with process (i.e., in-session experiences) and outcome variables (GAF & HSCL-11) in the same session to evaluate criterion validity. To facilitate comparison between the various emotion measures, a correlation matrix was created with the emotion measures in the columns and the process and outcome variables in the rows. The multilevel.cor function from the R package misty (v0.5.3; Yanagida, 2023) was used to compute within- and between-patient correlations.

Hierarchical linear models (HLMs). Random-intercept random-slope HLM models were fitted to predict processes and outcomes (2 sentiments: positive & negative × 3 processes: interpersonal, coping, & affective experiences × 2 outcomes: GAF & HSCL-11). The analytic approach was similar to the HLMs for the examination of the construct validity. Sentiments were used as predictors of processes and outcomes in the same session. Patients' mean sentiment scores were used as a level-2 predictor and patient mean-centered sentiment scores were added as a level-1 predictor. The within-patient effects were visualized based on these mean-centered sentiment scores and their model-implied intercepts and slopes for each patient using the R package ggplot2 (v3.3.6; Wickham, 2016). The graphs are based on unstandardized scores, whereas the statistics reported in the text are based on scaled scores. The cluster-robust standard errors were computed using the model_parameters function from the R package parameters (v0.21.1; Lüdecke et al., 2020) to account for nesting in therapists.

Multiple regressions. A two-step procedure was employed to assess the predictive criterion validity of the sentiments. First, random-intercept random-slope HLMs were used to fit session numbers as predictors of both positive and negative sentiments. This step allowed the estimation of individual sentiment intercepts, representing patients' initial sentiment levels (sentiment baseline) and sentiment slopes, which signify the direction and strength of each patient's sentiment change throughout treatment (sentiment change). In the second step, the generated point estimates of these intercepts and slopes from the initial HLMs were incorporated into separate multiple regression models. These regression models aimed to predict post-treatment outcomes, including measures of emotional disorders (PHQ-9_{POST} & GAD-7_{POST}), emotion regulation (ASQ_{POST}), and overall outcome (OQ-30_{POST}) while controlling for the initial scores of these dependent variables (e.g., OQ-30_{PRE}). All variables in the multiple linear regression were z-standardized. Cluster-robust standard errors were computed using the coeftest function from the R package lmtest (v0.9-40; Zeileis & Hothorn, 2002) to account for nesting in therapists.

Results

The descriptive statistics of the emotion, process, and outcome measures can be found in Table S2. The distributions of the variables are visualized in Figures S1–S2.

Construct Validity

MTMM analysis. An MTMM matrix (Table I) was built to explore the sentiments' construct validity (H1). SENT_{POS} was significantly positively associated with $\text{POMS}_{\text{PAT, POS}}$ ($r_w = .351$, $p_w = .003$; $r_b = .651$, $p_b = .001$) and SENT_{NEG} was significantly negatively associated with $\text{POMS}_{\text{PAT, POS}}$ within patients ($r_w = -.404$, $p_w < .001$), but not between patients ($r_b = -.323$, $p_b = .118$; H1a). Correspondingly, SENT_{POS} was significantly negatively associated with $\text{POMS}_{\text{PAT, NEG}}$ within patients ($r_w = -.394$, $p_w < .001$), but not between patients ($r_b = -.342$, $p_b = .240$), while SENT_{NEG} was significantly positively associated with $\text{POMS}_{\text{PAT, NEG}}$ within and between patients ($r_w = .351$, $p_w = .003$, $r_b = .446$, $p_b = .023$; H1b).

Similarly, SENT_{POS} was significantly positively associated with $\text{POMS}_{\text{THE, POS}}$ ($r_w = .493$, $p_w < .001$; $r_b = .629$, $p_b = .001$), while SENT_{NEG} was significantly negatively associated with $\text{POMS}_{\text{THE, POS}}$ within patients ($r_w = -.392$, $p_w = .002$), but not between patients ($r_b = -.192$, $p_b = .118$; H1c). Furthermore, SENT_{POS} was significantly negatively associated with $\text{POMS}_{\text{THE, NEG}}$ within patients ($r_w = -.438$, $p_w < .001$), but not between patients ($r_b = -.130$, $p_b = .726$), while SENT_{NEG} was positively associated with $\text{POMS}_{\text{THE, NEG}}$ within patients ($r_w = .323$, $p_w = .028$), but not between patients ($r_b = .373$, $p_b = .106$; H1d).

Hierarchical linear models (HlMs). Patients with at least two transcribed sessions were included in the HLM analyses. Therefore, 79 of 85 (92.9%) sessions nested in 29 of 35 (82.9%) patients were used. Table II shows the fixed within- and between-patient effects using sentiments to predict the other emotion measures (POMS_{PAT} & POMS_{THE}).

Sentiments had significant fixed effects on $\text{POMS}_{\text{PAT, POS}}$ within and between patients in the expected direction ranging from $b_b = -0.23$, $p_b = .033$, to $b_b = 0.47$, $p_b < .001$ (H1a). Except for one between-patient effect (SENT_{NEG} →

Table I. Construct Validity: Multitrait-Multimethod Matrix for Patients' Emotions Measured by Patients (POMS_{PAT}), Therapists (POMS_{THE}), and Sentiment Analysis (SENT) Using Multilevel Correlations.

Methods		POMS_{PAT}		POMS_{THE}		Sentiment Analysis	
	Traits	Positive	Negative	Positive	Negative	Positive	Negative
POMS_{PAT}	Positive	.90	$-.64^{***}$ [−0.96, −0.32]	$.86^{***}$ [.64, 1.07]	$-.58^{**}$ [−1.00, −.15]	$.65^{**}$ [.20, 1.00]	−.32 [−.75, .10]
	Negative	$-.43^{***}$ [−.66, −.19]	.92	−.38 [−0.82, .06]	.91 [−1.00, 1.00]	−.34 [−.95, .27]	$.45^{*}$ [.04, .86]
POMS_{THE}	Positive	$.36^{**}$ [.12, .61]	−.19 [−.46, .09]	.90	−.22 [−.78, .34]	$.63^{*}$ [.21, 1.00]	−.19 [−.65, .26]
	Negative	$-.32^{*}$ [−0.59, −0.05]	$.47^{***}$ [.24, .70]	$-.57^{***}$ [−.82, −.33]	.84	−.13 [−.88, .62]	.37 [−.11, .85]
Sentiment Analysis	Positive	$.35^{**}$ [.11, .59]	$-.39^{***}$ [−.63, −.16]	$.49^{***}$ [.27, .71]	$-.44^{***}$ [−.70, −.18]	—	−.37 [−0.89, .15]
	Negative	$-.40^{***}$ [−0.64, −0.17]	$.35^{**}$ [.11, .59]	$-.39^{**}$ [−.65, −.13]	$.32^{*}$ [.03, .62]	$-.77^{***}$ [−.88, −.65]	—

Note. $N_w = 85$ sessions. $N_b = 35$ patients. POMS = Profile of Mood States. PAT = Patients. THE = Therapists. Dark gray cells represent monotrait-heteromethod correlations indicating convergent validity. Medium gray cells represent heterotrait-heteromethod correlations indicating discriminant validity. Light gray cells represent heterotrait-monomethod correlations indicating specificity. The diagonal represents the reliability (Cronbach's α) in the current study. Lower triangular: Within-patient correlations. Upper triangular: Between-patient correlations.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table II. Construct Validity: Fixed Within- and Between-Patients Effects Using Sentiments to Predict Emotions Measured by Patients (POMS$_{\text{PAT}}$) and Therapists (POMS$_{\text{THE}}$) in the Same Session in HLMs.

Measure	Effect	SENT$_{\text{POS}}$			SENT$_{\text{NEG}}$		
		b	*t*	*p*	*b*	*t*	*p*
POMS$_{\text{PAT, POS}}$	Within	0.15	2.26*	.027	–0.23	–2.98**	.004
	Between	0.47	5.77***	< .001	–0.23	–2.17*	.033
POMS$_{\text{PAT, NEG}}$	Within	–0.22	–5.02**	.001	0.24	3.10**	.003
	Between	–0.27	–2.66*	.010	0.20	1.78	.080
POMS$_{\text{THE, POS}}$	Within	0.23	2.56*	.011	–0.18	–1.99*	.047
	Between	0.50	5.11***	< .001	–0.24	–2.09*	.037
POMS$_{\text{THE, NEG}}$	Within	–0.26	–1.93	.054	0.19	1.41	.158
	Between	–0.18	–1.44	.151	0.20	1.82	.069

Note. $n = 79$ sessions nested in 29 patients with at least two sessions each. POMS = Profile of Mood States. PAT = Patients. THE = Therapists. SENT = Sentiment. POS = Positive. NEG = Negative. HLMs = Hierarchical Linear Models.

POMS$_{\text{PAT, NEG}}$, $b_b = 0.20$, $p_b = .080$), sentiments also had significant fixed effects on POMS$_{\text{PAT, NEG}}$ in the expected direction between $b_b = –0.27$, $p_b = .033$, to $b_w = 0.24$, $p_w < .001$ (H1b). Sentiments also had significant effects on POMS$_{\text{THE, POS}}$ ($b_b = –0.24$, $p_b = .037$, to $b_b = 0.50$, $p_b < .001$, H1c), but none of the effects were significant for POMS$_{\text{THE, NEG}}$ ($b_w = –0.26$, $p_w = .054$, to $b_b = 0.20$, $p_b = .069$, H1d). Convergence issues were noted in models involving POMS$_{\text{THE}}$. The warning messages are available in the Supplemental Material. The within-patient effects are also visualized in Figure 1 (for POMS$_{\text{PAT}}$) and Figure S3 (for POMS$_{\text{THE}}$).

Criterion Validity

Correlational analyses

Processes. A correlation matrix (Table III) was created to evaluate the sentiments' criterion validity. SENT$_{\text{POS}}$ had a significantly positive between-patient correlation with EXP$_{\text{INT}}$, $r_b = .621$, $p_b = .013$ (H2a), and EXP$_{\text{COP}}$, $r_b = .524$, $p_b = .031$ (H2b), and a significantly negative within-patient correlation with EXP$_{\text{AFF}}$, $r_w = -.413$, $p_w < .001$ (H2c). Moreover, SENT$_{\text{NEG}}$ was also significantly positively correlated with EXP$_{\text{AFF}}$ within patients, $r_w = .322$, $p_w = .010$ (H2c). The direction of the within-patient correlation between SENT$_{\text{POS}}$ and EXP$_{\text{INT}}$ was not as expected, $r_w = -.099$, $p_w = .488$ (H2a).

Outcome. The sentiments showed correlations in the expected direction with the outcome measures (H3): GAF showed significant correlations with the sentiments between $r_b = -.410$, $p_b = .037$, and $r_w = .480$, $p_w < .001$, except for the between-patient correlation between SENT$_{\text{POS}}$ and GAF, $r = .097$, $p = .768$ (H3a). The HSCL-11 had a significantly negative within-patient correlation with SENT$_{\text{POS}}$, $r = -.290$, $p = .020$, and a significant between-patient correlation with SENT$_{\text{NEG}}$, $r = .655$, $p < .001$, while the between-patient correlation with SENT$_{\text{POS}}$, $r = -.258$, $p = .323$, and the within-patient correlation with SENT$_{\text{NEG}}$, $r = .184$, $p = .175$, were not significant (H3b).

Hierarchical linear models (HLMs)

Processes. Table IV shows the fixed effects of the sentiments predicting the process measures (in-session experiences) in the HLMs using cluster-robust standard errors to additionally correct for nesting in therapists. In contrast to the correlational analysis, sentiments had no significant effect on EXP$_{\text{INT}}$, $b = –0.13$, $p = .213$, to $b = 0.22$, $p = .238$ (H2a). The within-patient effects are visualized in Figure S4.

Outcomes. Table V also presents the fixed effects of the sentiments predicting the outcome measures (GAF & HSCL-11). SENT$_{\text{NEG}}$ had a significant between-patient effect on the level of functioning (GAF), $b = -0.35$, $p = .002$ (H3a), and symptom severity (HSCL-11), $b = 0.50$, $p = .006$ (H3b), in the expected direction, while within-patient effects were not significant. SENT$_{\text{POS}}$ had no significant effect on the outcome measures. The within-patient effects are visualized in Figure S5.

Multiple regressions. The results of the regression models exploring the sentiments' predictive criterion validity (H3c) are summarized in Table V. An increase in SENT$_{\text{POS}}$ was negatively associated with overall distress (OQ-30$_{\text{POST}}$), $b = -0.38$, $p = .024$, and generalized anxiety (GAD-7$_{\text{POST}}$), $b = -0.48$, $p = .019$. An increase in SENT$_{\text{NEG}}$ significantly predicted more depressive

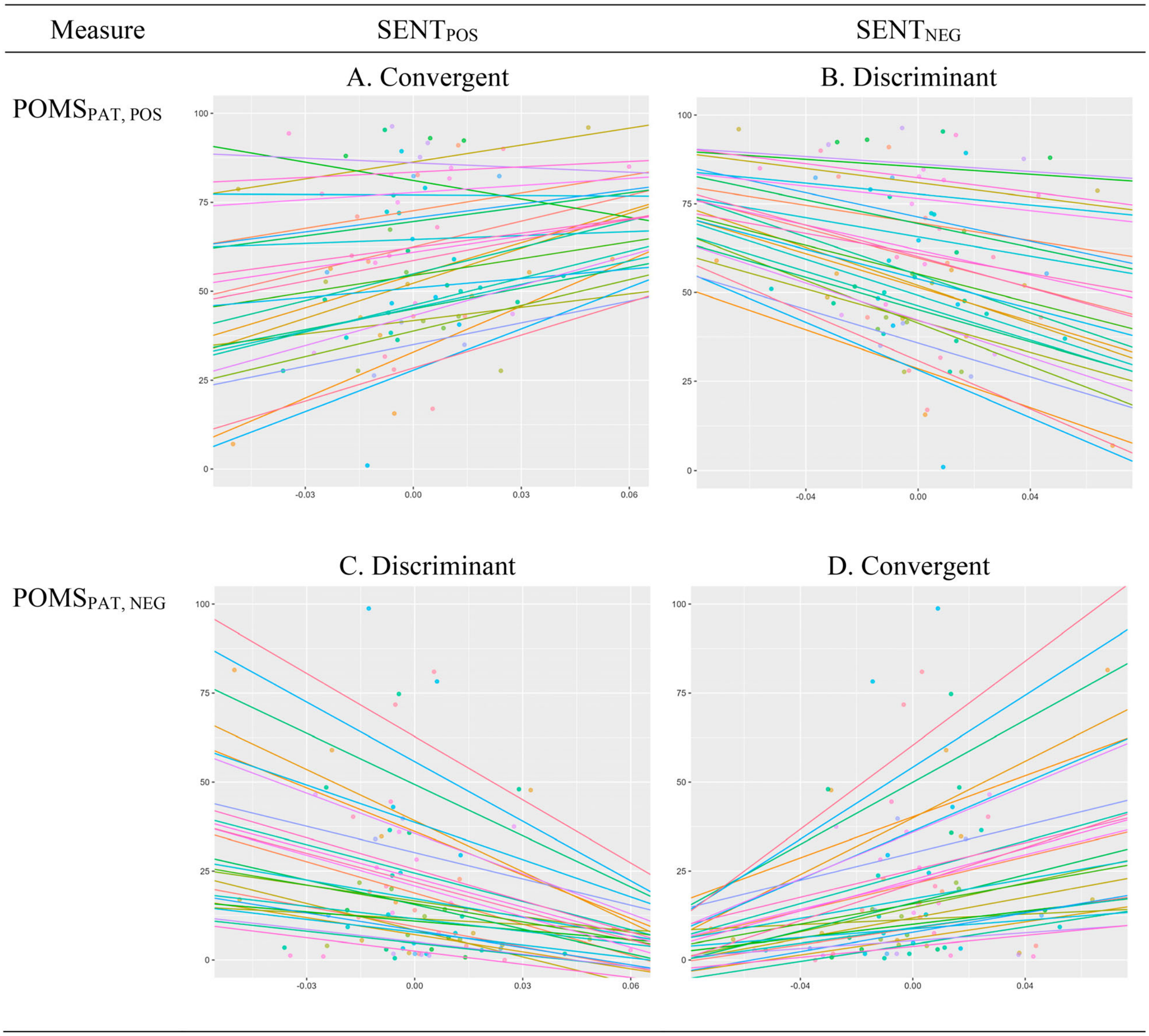

Figure 1. Construct Validity: Within-Patient Effects Using Sentiments to Predict Patients' Emotion Self-Reports in the Profile of Mood States (POMS$_{PAT}$) in the Same Session.
Note: $N = 79$ sessions nested in 29 patients with at least two sessions each. Each line represents a patient. POMS = Profile of Mood States. SENT = Sentiment. PAT = Patient. POS = Positive. NEG = Negative. Convergent = Relationship between measures of the same construct. Discriminant = Relationship between measures of different constructs.

symptoms (PHQ-9 $_{POST}$), $b = 0.68$, $p = .041$, and stronger generalized anxiety (GAD-7$_{POST}$), $b = 1.10$, $p = .020$. In contrast, an increase in SENT$_{POS}$ was positively associated with more functional emotion regulation (ASQ $_{POST}$), $b = 0.52$, $p < .001$. All other associations between sentiment change and outcome after treatment termination were non-significant.

Discussion

The aim of the present study was to explore the validity of sentiment analysis as a measure of the basic emotional tone in patients' psychotherapy transcripts. Regarding construct validity, we assumed that the sentiments correlate with self- and therapist reports of patients' emotions in the same session (H1). The MTMM analysis largely supported the sentiments' construct validity. Seven of eight tested multilevel correlations between sentiments and self- and therapist reports were significant and in the expected direction. When correcting for nesting in therapists in HLMs using cluster-robust standard errors, five of eight tested associations were significant and in the expected direction. However, no significant within- or between-patient effects on

Table III. Criterion Validity: Multilevel Correlations Between Sentiments, Process, and Outcome Measures Compared to Patient Self-Reports (POMS_{PAT}) and Therapist Reports (POMS_{THE}).

Valence		Positive			Negative		
Measure		POMS_{PAT}	POMS_{THE}	Sentiment	POMS_{PAT}	POMS_{THE}	Sentiment
				Process Measures			
Interpersonal	r_w	.31*	.26	−.10	−.06	−.34*	.10
Experience (EXP_{INT})	CI_w	[.05, .57]	[−.03, .54]	[−.37, .18]	[−.36, .23]	[−.61, -.07]	[−.18, .38]
	r_b	.45*	.42*	.62*	−.46*	−.56**	−.26
	CI_b	[.06, .83]	[.02, .81]	[.06, 1.00]	[−.91, −.01]	[−1.00, −.12]	[−.72, .20]
Coping	r_w	.48***	.34**	.21	−.41***	−.45***	−.19
Experience (EXP_{COP})	CI_w	[.27, .70]	[.09, .60]	[−.06, .47]	[−.66, −.16]	[−.68, −.21]	[−.46, .08]
	r_b	.73***	.69***	.52*	−.20	−.32	−.35
	CI_b	[.49, .97]	[.38, 1.00]	[−.00, 1.00]	[−.70, .30]	[−.84, .21]	[−.78, .08]
Affective	r_w	−.16	−.20	−.41***	.68***	.41**	.32*
Experience (EXP_{AFF})	CI_w	[−.44, .11]	[−.47, .08]	[−.64, −.18]	[.53, .84]	[.16, .66]	[.07, .57]
	r_b	−.20	.19	−.11	.53*	.55*	.26
	CI_b	[−.76, .36]	[−.41, .79]	[−.88, .66]	[.09, .98]	[−.03, 1.00]	[−.29, .80]
				Outcome Measures			
Global Assessment of	r_w	.48***	.61***	.48***	−.29*	−.66***	−.41**
Functioning (GAF)	CI_w	[.24, .71]	[.40, .83]	[.25, .71]	[−.56, −0.03]	[−.86, −0.47]	[−.71, −.12]
	r_b	.37	.20	.10	−.61***	−.19	−.41*
	CI_b	[−.05, .79]	[−.27, .67]	[−.57, .77]	[−.97, −0.24]	[−.76, .38]	[−.82, −.00]
Hopkins Symptom	r_w	−.20	−.25	−.29*	.25	.46***	.18
Checklist-11 (HSCL-11)	CI_w	[−.48, .08]	[−.52, .02]	[−.54, −.04]	[−.01, .52]	[.23, .70]	[−.09, .45]
	r_b	−.67***	−.42*	−.26	.53**	.53**	.66***
	CI_b	[−.95, −.39]	[−.77, −.06]	[−.80, .28]	[.20, .86]	[.17, .90]	[.37, .94]

Note. N_w = 85 sessions. N_b = 35 patients. POMS = Profile of Mood States. r_w = Within-patient correlation coefficient. r_b = Between-patient correlation coefficient. The correlation coefficients were pooled across the ten imputed datasets. CI_w: 95% confidence interval for the within-patient correlation. CI_b: 95% confidence interval for the between-patient correlation.
* $p < .05$. ** $p < .01$. *** $p < .001$.

therapists' reports of negative emotions were found for either positive or negative sentiments. Thus, hypothesis H1d was not supported, while the remaining hypotheses H1a to H1c could be confirmed.

Exploring criterion validity, sentiments also showed significant associations with process measures (H2). Either positive or negative sentiments were significantly related to coping experiences (H2b) and problem-related affective experiences (H2c) in

Table IV. Criterion Validity: Fixed Within- and Between-Patients Effects Using Sentiments to Predict Process and Outcome Measures in the Same Session in HLMs.

Measure		SENT_{POS}			SENT_{POS}		
	Effect	*b*	*t*	*p*	*b*	*t*	*p*
			Process Measures				
EXP_{INT}	Within	−.13	−1.26	.213	.17	1.60	.114
	Between	.22	1.19	.238	.01	0.08	.938
EXP_{COP}	Within	.11	2.36*	.021	−.11	−1.67	.100
	Between	.28	3.38**	.001	−.25	−1.38	.172
EXP_{AFF}	Within	−.25	−3.59***	.001	.18	2.28*	.026
	Between	−.22	−1.64	.106	.21	1.78	.080
			Outcome Measures				
GAF	Within	.15	1.33	.183	−.12	−1.03	.301
	Between	.20	1.29	.197	−.35	−3.14**	.002
HSCL-11	Within	−.11	−1.98	.052	.05	0.70	.489
	Between	−.20	−1.26	.212	.50	2.84**	.006

Note. n = 79 sessions nested in 29 patients with at least two sessions each. EXP_{INT} = Interpersonal Experiences. EXP_{COP} = Coping Experiences. EXP_{AFF} = Affective Experiences. GAF = Global Assessment of Functioning. HSCL-11 = Hopkins Symptom Checklist-11.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table V. Predictive Criterion Validity: Sentiments Predicting Outcome Measures (OQ-30, PHQ, GAD, & ASQ) After Termination of Treatment Controlling for Intake Severity in Multiple Regressions.

	Positive Sentiment					Negative Sentiment				
Variables	β	*SE*	*t*	*p*	95% CI	β	*SE*	*t*	*p*	95% CI
				Outcome Questionnaire-30 (OQ-30 $_{\text{POST}}$)						
Sentiment Baseline	0.34*	.13	2.56	.013	[0.07, 0.60]	0.59	.38	1.54	.124	[−0.16, 1.35]
Sentiment Change	−0.38*	.17	−2.26	.024	[−0.72, −0.05]	0.76	.40	1.87	.062	[−0.04, 1.55]
OQ $_{\text{PRE}}$	0.72***	.12	5.82	<.001	[0.47, 0.96]	0.63***	.14	4.51	<.001	[0.35, 0.90]
				Patient Health Questionnaire-9 (PHQ-9 $_{\text{POST}}$)						
Sentiment Baseline	0.22	.13	1.74	.084	[−0.03, 0.48]	0.73*	.33	2.18	.029	[0.07, 1.38]
Sentiment Change	−0.27	.16	−1.69	.091	[−0.58, 0.04]	0.68*	.33	2.04	.041	[0.03, 1.34]
PHQ $_{\text{PRE}}$	0.70***	.14	5.00	<.001	[0.43, 0.98]	0.55***	.15	3.71	<.001	[0.26, 0.84]
				Generalized Anxiety Disorder-7 (GAD-7 $_{\text{POST}}$)						
Sentiment Baseline	0.37	.19	1.92	.058	[−0.01, 0.76]	1.09**	.41	2.64	.009	[0.28, 1.90]
Sentiment Change	−0.48*	.20	−2.36	.019	[−0.88, −0.08]	1.10*	.46	2.36	.020	[0.18, 2.02]
GAD $_{\text{PRE}}$	0.55**	.19	2.86	.005	[0.17, 0.94]	0.33	.21	1.60	.113	[−0.08, 0.74]
				Affective Style Questionnaire (ASQ $_{\text{POST}}$)						
Sentiment Baseline	−0.07	.19	−0.37	.712	[−0.46, 0.32]	−0.94	.52	−1.82	.069	[−1.95, 0.07]
Sentiment Change	0.52*	.23	2.24	.027	[0.06, 0.98]	−0.86	.48	−1.78	.076	[−1.81, 0.09]
ASQ $_{\text{PRE}}$	0.48***	.14	3.45	<.001	[0.20, 0.75]	0.43*	.18	2.38	.018	[0.08, 0.79]

Note. $n = 29$ patients with at least two sessions each. Sentiment baseline and change scores were based on the intercepts and slopes from hierarchical linear models (HLM) with at least two sessions nested in the patients. PRE = Measured before the beginning of treatment. POST = Measured after treatment termination. All variables in the multiple linear regressions were z-standardized. β-weights and standard errors (SE) were pooled, and the test statistics and *p*-values are based on Rubin's Rule.

* $p < .05$. ** $p < .01$. *** $p < .001$.

the expected direction. However, no significant effects were found for interpersonal experiences in the HLM. Thus, H2a was not supported.

Findings were also in support of criterion validity when exploring the associations between sentiments and outcome measures of psychotherapy (H3). Either positive or negative sentiments were significantly related to the level of functioning (H3a) and symptom distress (H3b) in the expected direction. The between-patient correlation among negative sentiments and symptom severity in the HSCL-11 was particularly high ($r = .66$). Moreover, the sentiments also showed predictive criterion validity. An increase in positive sentiment throughout therapy significantly predicted less distress on the OQ-30, less anxiety on the GAD-7, and more functional emotion regulation on the ASQ, while an increase in negative sentiment predicted more depressive symptoms on the PHQ-9 and more anxiety on the GAD-7 after treatment termination (H3c). Overall, the results supported eight of ten hypotheses in favor of the sentiments' validity, while two hypotheses were disconfirmed (H1d & H2a).

Implications for Research and Practice

Our study showed that sentiments were significantly correlated with various process and outcome measures, demonstrating their potential for the evaluation of patient progress. Sentiment analysis may provide therapists with an additional measure to track progress and assess the achievement of therapeutic goals related to emotional well-being. Continuous monitoring of sentiment could signal sessions with sudden increases in negative sentiment and help therapists identify clients in crisis. Research could focus on change sensitivity, capturing dynamic trajectories of sentiments throughout therapy. Examining established patterns of change, such as sudden gains or losses, could shed more light on the validity and utility of sentiment analysis.

The integration of sentiment analysis could supplement and enhance traditional measures of emotion by providing automated and objective assessments of emotional expressions. It could help compensate for the limitations of traditional self-report measures of emotions, which can be biased by social desirability or memory recall. The multimodal measurement approach in this study revealed a few discrepancies between therapist ratings of patient emotions and those achieved by sentiment analysis. On average, positive sentiments were negatively correlated with therapist ratings of negative emotions (Figure S3$_{\text{C}}$). However, the correlation was positive for some patients, indicating that therapists may have been unable to identify their patients' emotions correctly or that patients' non- and paraverbal emotional expressions differed from what they said. Therapists may profit from feedback on such discrepancies in emotional expression, especially since patient-focused research has demonstrated the general benefits of feedback and data-informed psychological therapies (de Jong et al., 2021; Lutz et al., 2022). Therefore, it is crucial to develop systems that integrate and provide easy access to emotional process feedback in clinical practice, training, and supervision (e.g., Trier Treatment Navigator, TTN; Lutz et al., 2019; Lutz et al., 2022). One potential barrier to the widespread implementation of sentiment analysis in clinical practice is the need for transcripts. To address this issue, ways to streamline the transcription process could be explored, such as using large language models for automated transcription.

Future research could benefit from incorporating additional sources, including psychophysiological measures like heart rate variability (Hehlmann et al., 2021), video analyses (Schwartz et al., 2022), and paraverbal language features (Nof et al., 2021). Sentiment analysis could be integrated with automatic emotion recognition in facial expressions and vocal cues using the Nonverbal Behavior Analyzer (NOVA; Baur et al., 2013; Baur et al., 2020; Baur et al., 2020). With its integrated video player, NOVA could help therapists locate and analyze crucial moments in therapy sessions with high emotional intensity. This feature could enable therapists to engage in meaningful discussions with their patients or supervisors about these critical incidents. In this way, integrating sentiment analysis into therapist training and supervision can provide feedback on the emotional dynamics of therapy sessions and contribute to the ongoing development of therapeutic skills.

It has the potential to expand the study of emotional processes within a therapy session by providing moment-to-moment analyses. These fine-grained assessments have the advantage of capturing fluctuations in emotions over time. In addition, sentiment analysis can be conducted on large amounts of data, making it a promising tool for investigating the dynamics of emotional experiences in therapy. In practical applications, sentiment analysis could also be helpful when combined with other large language models, topic models, and NLP algorithms to aid in the automatic measurement or detection of more complex constructs relevant to psychotherapy, such as alliance ruptures (Atzil-Slonim et al., 2021). These benefits suggest that sentiment analysis is a valuable addition to the psychotherapy research methodology and has the potential to deepen our understanding of emotional processes in psychotherapy.

Limitations & Strengths

Methodological limitations include the use of a non-specific NLP method (Denecke & Deng, 2015; Tanana et al., 2016, 2021), no evidence for inter-transcriber reliability, and no investigation into the reliability of sentiments or sentiment interaction between therapist and patient (Aafjes-van Doorn & Müller-Frommeyer, 2010; Syzdek, 2020). A further limitation is the exploratory nature of the study with a multitude of hypothesis-relevant tests (i.e., 16 for construct validity, 20 for criterion validity, and 8 for predictive validity). Therefore, the findings need to be replicated in studies with a larger number of sessions and correction for multiple testing. Potential bias may have been introduced by using point estimates from hierarchical linear models in separate multiple regression models. Convergence warnings were encountered in our analyses using HLMs, particularly those examining associations between sentiments and therapist reports of patients' emotions (POMS_{THE}), urging caution in interpreting these results. The small sample size, mainly due to the complexities of collecting complete multimodal measures and the time-consuming transcription process, and the non-normal distribution of interpersonal experiences, potentially stemming from consistently high patient ratings of the therapeutic alliance, also represent significant constraints.

Despite these limitations, the study has several strengths. It assessed the validity of sentiment analysis in psychotherapy, exploring its associations within an extensive nomological network and applied advanced statistical methods, such as multiple imputation and cluster-robust standard errors. The study was conducted in a naturalistic, face-to-face setting and demonstrated the predictive validity of sentiments at various time points using longitudinal data. Furthermore, a multimodal approach was adopted, utilizing a diverse array of validation scales, and a multilingual algorithm was implemented, showing promise for research across different languages.

While the insights obtained from this study are comprehensive, they remain initial and should be interpreted with caution. Future research should include more patients, apply corrected testing, and integrate more session transcripts. We are exploring more advanced transcription methods to reduce previous limitations and build upon these initial findings for subsequent research on sentiment analysis in psychotherapy.

Conclusion

Sentiment analysis has demonstrated potential, revealing several meaningful associations and providing initial evidence supporting its validity as a measure of the fundamental emotional tone in psychotherapy sessions. Sentiments were significantly related to other measures of emotion as well as processes and outcomes of psychotherapy. Thus, sentiments could augment the multimodal measurement of emotions in psychotherapy research. Future research should explore combining sentiments with other modalities to enhance our understanding of emotional dynamics in therapy and evaluating their utility in measurement-based and feedback-informed psychotherapy.

Author Contributions and Conflict of Interest

STE, DM, BS, JAR, TB, EA, and WL contributed to the conception and design of the study. STE, DM, BS, and MJ performed the statistical and sentiment analysis. STE and JS wrote the first draft of the manuscript. DM, BS, JAR, TB, EA, and WL revised the manuscript. All authors participated in proofreading the manuscript and approved the submitted version. All authors declare that they have no conflicts of interest. We want to thank Kaitlyn Poster, Ph.D., for proofreading the manuscript and Patrick Bungart for his assistance.

Funding

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG; LU 660/16-1, grant number 493169211 & LU 660/19-1, grant number 504507043).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Ethics

The present study analyzed routine data from a university outpatient clinic in Trier, Germany. All procedures in this study complied with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975 and its later amendments. Written, informed consent allowing anonymized data to be used for research purposes was obtained from all participants. The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Supplemental Data

Supplemental data for this article can be accessed at https://doi.org/10.1080/10503307.2024.2322522.

ORCID

Steffen T. Eberhardt http://orcid.org/0000-0002-9900-4671
Jana Schaffrath http://orcid.org/0000-0003-0106-0243
Danilo Moggia http://orcid.org/0000-0001-6321-4450
Brian Schwartz http://orcid.org/0000-0003-4695-4953
Martin Jaehde http://orcid.org/0009-0001-6441-9543
Julian A. Rubel http://orcid.org/0000-0002-9625-6611
Tobias Baur http://orcid.org/0000-0002-2797-605X
Elisabeth André http://orcid.org/0000-0002-2367-162X
Wolfgang Lutz http://orcid.org/0000-0002-5141-3847

References

Aafjes-van Doorn, K., & Müller-Frommeyer, L. (2020). Reciprocal language style matching in psychotherapy research. *Counselling and Psychotherapy Research*, *20*(3), 449–455. https://doi.org/10.1002/capr.12298

Aas, I. M. (2010). Global Assessment of Functioning (GAF): Properties and frontier of current knowledge. *Annals of General Psychiatry*, *9*(1), 1–11. https://doi.org/b75rsp

Aas, I. M. (2011). Guidelines for rating global assessment of functioning (GAF). *Annals of General Psychiatry*, *10*(1), 1–11. https://doi.org/10.1186/1744-859X-10-2

Abd Rahman, R., Omar, R., Noah, K., Danuri, S. A. M., Al-Garadi, M. S. N. M., & A, M. (2020). Application of machine learning methods in mental health detection: A systematic review. *IEEE Access*, *8*, 183952–183964. https://doi.org/10.1109/ACCESS.2020.3029154

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., Juravski, D., Bar-Kalifa, E., Gilboa-Schechtman, E., Tuval-Mashiach, R., Shapira, N., & Goldberg, Y. (2021). Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy*, *58*(2), 324–339. https://doi.org/10.1037/pst0000362

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 258–266. https://doi.org/mc4b

Bar-Kalifa, E., Prinz, J. N., Atzil-Slonim, D., Rubel, J. A., Lutz, W., & Rafaeli, E. (2019). Physiological synchrony and therapeutic alliance in an imagery-based treatment. *Journal of*

Counseling Psychology, *66*(4), 508–517. https://doi.org/10.1037/cou0000358

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baur, T., Clausen, S., Heimerl, A., Lingenfelser, F., Lutz, W., & André, E. (2020). *NOVA: A tool for explanatory multimodal behavior analysis and its application to psychotherapy*. MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, 577–588.

Baur, T., Damian, I., Lingenfelser, F., Wagner, J., & André, E. (2013). *NOVA: Automated analysis of nonverbal signals in social interactions*. Human Behavior Understanding: 4th International Workshop, HBU 2013, 160–171.

Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M. F., Schuller, B., & André, E. (2020). eXplainable cooperative machine learning with NOVA. *KI - Künstliche Intelligenz*, *34*(2), 143–164. https://doi.org/10.1007/s13218-020-00632-3

Boyle, K., Deisenhofer, A. K., Rubel, J. A., Bennemann, B., Weinmann-Lutz, B., & Lutz, W. (2019). Assessing treatment integrity in personalized CBT: The inventory of therapeutic interventions and skills. *Cognitive Behaviour Therapy*, *49*(3), 210–227. https://doi.org/10.1080/16506073.2019.1625945

Bullis, J. R., Boettcher, H. T., Sauer-Zavala, S., Farchione, T. J., & Barlow, D. H. (2019). What is an emotional disorder? A transdiagnostic mechanistic definition with implications for assessment, treatment, and prevention. *Clinical Psychology: Science and Practice*, *26*(2). https://doi.org/10.1111/cpsp.12278

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M.. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv:2006.13979v2*. https://doi.org/10.48550/arXiv.2006.13979

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Jong, K., Conijn, J. M., Gallagher, R. A., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, https://doi.org/10.1016/j.cpr.2021.102002

Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, *64*(1), 17–27. https://doi.org/10.1016/j.artmed.2015.03.006

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*(3), 1–36. https://doi.org/10.1145/2682899

Eid, M., Geiser, C., & Nussbeck, F. W. (2014). Multitrait-multimethod analysis in psychotherapy research: New methodological approaches. In W. Lutz, & S. Knox (Eds.), *Quantitative and qualitative methods in psychotherapy research* (pp. 44–52). Routledge.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues* (Vol. 10). Malor Books.

Ellsworth, J. R., Lambert, M. J., & Johnson, J. (2006). A comparison of the Outcome Questionnaire-45 and Outcome Questionnaire-30 in classification and prediction of treatment outcome. *Clinical Psychology & Psychotherapy*, *13*(6), 380–391. https://doi.org/10.1002/cpp.503

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1995). *Structured clinical interview for DSM-IV axis I disorders – patient edition (SCID-I/P, version 2.0)*. Biometrics Research Department, New York State Psychiatric Institute.

Fitzpatrick, M., & Stalikas, A. (2008). Integrating positive emotions into theory, research, and practice: A new challenge for psychotherapy. *Journal of Psychotherapy Integration*, *18*(2), 248–258. https://doi.org/10.1037/1053-0479.18.2.248

Flückiger, C., Regli, D., Zwahlen, D., Hostettler, S., & Caspar, F. (2010). Der Berner Patienten-und Therapeutenstundenbogen 2000 [The Bern Post Session Report 2000, Patient and therapist versions: Measuring psychotherapeutic processes]. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *39*(2), 71–79. https://doi.org/10.1026/1616-3443/a000015

Graser, J., Bohn, C., Kelava, A., Schreiber, F., Hofmann, S. G., & Stangier, U. (2012). Der „Affective Style Questionnaire (ASQ)": Deutsche Adaption und Validitäten [The „Affective Style Questionnaire (ASQ)": German adaption and validity]. *Diagnostica*, *58*(2), 100–111. https://doi.org/10.1026/0012-1924/a000056

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

Greenberg, L. S., & Watson, J. C. (2006). *Emotion-focused therapy for depression*. American Psychological Association.

Guo, F., Cao, Y., Ding, Y., Liu, W., & Zhang, X. (2014). A multimodal measurement method of users' emotional experiences shopping online. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *25*(5), 585–598. https://doi.org/10.1002/hfm.20577

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Han, J., Zhang, Z., Mascolo, C., André, E., Tao, J., Zhao, Z., & Schuller, B. W. (2021). Deep learning for mobile mental health: Challenges and recent advances. *IEEE Signal Processing Magazine*, *38*(6), 96–105. https://doi.org/10.1109/MSP.2021.3099293

Hehlmann, M. I., Schwartz, B., Lutz, T., Gómez Penedo, J. M., Rubel, J. A., & Lutz, W. (2021). The use of digitally assessed stress levels to model change processes in CBT – a feasibility study on seven case examples. *Frontiers in Psychiatry*, *12*, 613085. https://doi.org/10.3389/fpsyt.2021.613085

Heymans, M. W., & Twisk, J. W. R. (2022). Handling missing data in clinical research. *Journal of Clinical Epidemiology*, *151*, 185–188. https://doi.org/10.1016/j.jclinepi.2022.08.016

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Kazdin, A. E. (1996). Combined and multimodal treatments in child and adolescent psychotherapy: Issues, challenges, and research directions. *Clinical Psychology: Science and Practice*, *3*(1), 69–100. https://doi.org/10.1111/j.1468-2850.1996.tb00059.x

Kihlstrom, J. F., Eich, E., Sandbrand, D., & Tobias, B. A. (2000). Emotion and memory: Implications for self-report. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 81–99). Lawrence Erlbaum Associates Publishers.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*(2), 99–117. https://doi.org/10.1007/s10772-011-9125-1

Kraiss, J. T., Peter, M., Moskowitz, J. T., & Bohlmeijer, E. T. (2020). The relationship between emotion regulation and well-being in patients with mental disorders: A meta-analysis. *Comprehensive Psychiatry*, *102*, 152189. https://doi.org/10.1016/j.comppsych.2020.152189

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kumar, K. S., Desai, J., & Majumdar, J. (2016). *Opinion mining and sentiment analysis on online customer review*. 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 1–4.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer.

Leahy, R. L. (2008). The therapeutic relationship in cognitive-behavioral therapy. *Behavioural and Cognitive Psychotherapy*, *36* (6), 769–777. https://doi.org/10.1017/S1352465808004852

Löwe, B., Müller, S., Brähler, E., Kroenke, K., Albani, C., & Decker, O. (2007). Validierung und Normierung eines kurzen Selbstratinginstrumentes zur Generalisierten Angst (GAD-7) in einer repräsentativen Stichprobe der deutschen Allgemeinbevölkerung. *PPmP – Psychotherapie · Psychosomatik · Medizinische Psychologie*, *57*(2), A050. https://doi.org/10.1055/s-2007-970669

Löwe, B., Spitzer, R. L., Zipfel, S., & Herzog, W. (2002). *PHQ-D: Gesundheitsfragebogen für Patienten; Manual, Komplettversion und Kurzform*. Pfizer.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445

Lutz, W., Deisenhofer, A. K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B. (2022). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology*, *90*(1), 90–106. https://doi.org/10.1037/ccp0000642

Lutz, W., Deisenhofer, A.-K., Weinmann-Lutz, B., & Barkham, M. (2023). Data-informed clinical training and practice. In L. G. Castonguay, & C. E. Hill (Eds.), *Becoming better psychotherapists: Advancing training and supervision* (pp. 191–213). American Psychological Association.

Lutz, W., de Jong, K., Rubel, J. A., & Delgadillo, J. (2021). Measuring, predicting, and tracking change in psychotherapy. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 89–133). Wiley.

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A. K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438. https://doi.org/10.1016/j.brat.2019.103438

Lutz, W., Schwartz, B., & Delgadillo, J. (2022). Measurement-based and data-informed psychological therapy. *Annual Review of Clinical Psychology*, *18*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Lutz, W., Tholen, S., Schürch, E., & Berking, M. (2006). Reliabilität von Kurzformen gängiger psychometrischer Instrumente zur Evaluation des therapeutischen Fortschritts in Psychotherapie und Psychiatrie. *Diagnostica*, *52*(1), 11–25. https://doi.org/10.1026/0012-1924.52.1.11

Mayring, P. (1990). *Einführung in die qualitative Sozialforschung [Introduction to qualitative social research]*. Psychologie Verlags Union.

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

McNeish, D. M., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. https://doi.org/10.1037/met0000078

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 201–237). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100508-8.00009-6.

Nakayama, M., Hatanakaka, C., Konakawa, H., Suzuki, Y., Koh, A., Sugihara, Y., & Kawai, T. (2021). *Japanese dictionary for sentiment analysis of counselling text*. Proceedings of the 9th International Conference on Human-Agent Interaction, 311–315.

Nof, A., Amir, O., Goldstein, P., & Zilcha-Mano, S. (2021). What do these sounds tell us about the therapeutic alliance: Acoustic markers as predictors of alliance. *Clinical Psychology & Psychotherapy*, *28*(4), 807–817. https://doi.org/10.1002/cpp.2534

Peluso, P. R., & Freund, R. R. (2018). Therapist and client emotional expression and psychotherapy outcomes: A meta-analysis. *Psychotherapy*, *55*(4), 461–472. https://doi.org/10.1037/pst0000165

Provoost, S., Ruwaard, J., Van Breda, W., Riper, H., & Bosse, T. (2019). Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: An exploratory study. *Frontiers in Psychology*, *10*, 1065. https://doi.org/10.3389/fpsyg.2019.01065

Robitzsch, A., & Grund, S.. (2023). miceadds: Some additional multiple imputation functions, especially for "mice." https://cran.r-project.org/package=miceadds

Rubel, J. A., Rosenbaum, D., & Lutz, W. (2017). Patients' in-session experiences and symptom change: Session-to-session effects on a within-and between-patient level. *Behaviour Research and Therapy*, *90*, 58–66. https://doi.org/10.1016/j.brat.2016.12.007

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Schwartz, B., Rubel, J. A., Deisenhofer, A.-K., & Lutz, W. (2022). Movement-based patient-therapist attunement in psychological therapy and its association with early change. *Digital Health*, *8*, https://doi.org/10.1177/20552076221129098

Shapira, N., Lazarus, G., Goldberg, Y., Gilboa-Schechtman, E., Tuval-Mashiach, R., Juravski, D., & Atzil-Slonim, D. (2021). Using computerized text analysis to examine associations between linguistic features and clients' distress during psychotherapy. *Journal of Counseling Psychology*, *68*(1), 77. https://doi.org/10.1037/cou0000440

Shatte, A., Hutchinson, D., & Teague, S. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Smink, W. A., Fox, J., Sang, E. T. K., Sools, A. M., Westerhof, G. J., & Veldkamp, B. P. (2019). Understanding therapeutic change process research through multilevel modeling and text mining. *Frontiers in Psychology*, *10*, https://doi.org/10.3389/fpsyg.2019.01186

Soleymani, M., Pantic, M., & Pun, T. (2011). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, *3*(2), 211–223. https://doi.org/10.1109/T-AFFC.2011.37

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, *10*(1), 116–132. https://doi.org/10.1038/s41398-020-0780-3

Syzdek, B. M. (2020). Client and therapist psychotherapy sentiment interaction throughout therapy. *Psychological Studies*, *65* (4), 520–530. https://doi.org/10.1007/s12646-020-00567-7

Tamir, M., & Bigman, Y. E. (2018). Expectations influence how emotions shape behavior. *Emotion*, *18*(1), 15–25. https://doi.org/10.1037/emo0000351

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment*, *65*, 43–50. https://doi.org/10.1016/j.jsat.2016.01.006

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

Terhürne, P., Schwartz, B., Baur, T., Schiller, D., Eberhardt, S. T., André, E., & Lutz, W. (2022). Validation and application of the Non-Verbal Behavior Analyzer: An automated tool to assess non-verbal emotional expressions in psychotherapy. *Frontiers in Psychiatry*, *13*, 1026015. https://doi.org/10.3389/fpsyt.2022.1026015

Tschacher, W., & Meier, D. (2020). Physiological synchrony in psychotherapy sessions. *Psychotherapy Research*, *30*(5), 558–573. http://doi.org/10.1080/10503307.2019.1612114

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Whelton, W. J. (2004). Emotional processes in psychotherapy: Evidence across therapeutic modalities. *Clinical Psychology and Psychotherapy*, *11*(1), 58–71. https://doi.org/10.1002/cpp.392

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Yanagida, T. (2023). *misty: Miscellaneous Functions*. https://cran.r-project.org/package=misty.

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, *60*(2), 617–663. https://doi.org/10.1007/s10115-018-1236-4

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10.