NLP meets psychotherapy: Using predicted client emotions and self-reported client emotions to measure emotional coherence

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

[1]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
`www.ukp.tu-darmstadt.de`

[2] Psychotherapy Research Lab (PR Lab)
Department of Psychology
Bar-Ilan University
`www.prlab.co.il/en/`

Abstract

Emotions are experienced and expressed through various response systems. Coherence between emotional experience and emotional expression is considered highly important to clients' well being. To date, emotional coherence has been studied at a single time point using lab-based tasks with relatively small datasets. No study has examined emotional coherence between the subjective experience of emotions and utterance-level emotions over therapy sessions or whether this coherence is associated with clients' well being. Natural language Processing (NLP) approaches have been applied to identify emotions during psychotherapy dialogue, which can be implemented to study emotional processes on a larger scale and with specificity. However, these methods have yet to be used to study coherence between emotional experience and emotional expression over the course of therapy and whether it relates to clients' well-being.

This work presents an end-to-end approach where we use emotion predictions from our transformer based emotion recognition model to study emotional coherence and its diagnostic potential in psychotherapy research. We first employ our transformer based approach on a Hebrew psychotherapy dataset to automatically label clients' emotions at the utterance level in psychotherapy dialogues. We subsequently investigate the emotional coherence between clients' self-reported emotional states and our model-based emotion predictions. We also examine the association between emotional coherence and clients' well being.

The findings indicate a significant correlation between clients' self-reported emotions and positive and negative emotions expressed verbally during psychotherapy sessions. Coherence in positive emotions was also highly correlated with clients well-being. These results illustrate how NLP can be applied to identify important emotional processes in psychotherapy to improve diagnosis and treatment for clients who suffer from mental-health problems.

36 Conference on Neural Information Processing Systems (@PAI4MHNeurIPS 2022), New Orleans.

1 Introduction

One of the main goals of psychotherapy is to help clients develop greater coherence between their emotional experiences and the words that describe them, since this can help them respond more effectively to emotional events [Fosha, 2001, Greenberg, 2012]. Emotional coherence, which is defined as the coordination between different emotional responses as the emotion unfolds over time, is considered crucial to individuals' well-being [Levenson, 2014, Mauss et al., 2005].

Most previous studies assessing emotional coherence have used a lab-based emotional stimuli to evaluate emotional responses (e.g. [Hastings et al., 2009, Negrao et al., 2005, Wagner et al., 2003]) , which cannot capture a genuine real-life emotional response to an interpersonal interaction (such as between clients and therapists). In addition, most of these studies have compared clinical to non-clinical population at a single time point, which limits their ability to determine how different emotional responses change together over time across psychotherapy sessions. Furthermore, these studies focus on several emotional responses, such as between individuals' subjective reports and physiological response [Brown et al., 2020, Lohani et al., 2018] or facial expressions [Lohani et al., 2018].

Psychotherapy researchers have highlighted the importance of coherence between subjective emotional experiences and the words that describe them, since it allows people to give meaning to their experiences which may lead to a better ability to communicate and regulate emotions in a more adaptive way [Greenberg, 2012, Lane et al., 2015]. However, to the best of our knowledge, no empirical study has tested the coherence between these two emotional responses.

Lack of research in this direction may be due to the fact that until recently psychotherapy studies were dependent on human annotators, which limited their ability to capture within-session emotional processes on a large scale (e.g., [Imel et al., 2015]). Studies in the last few year, however, have demonstrated the value of implementing Natural Language Processing (NLP) to understand psychotherapy dialogue [Tanana et al., 2021, Gibson et al., 2017]. Given the vast potential of NLP in data-driven analysis it has also become a key resource for psychotherapy diagnosis as reviewed by [Shatte et al., 2019].

NLP methods that can automatically capture utterance-level emotions at a higher scale and specificity could successfully address the current gaps in the emotional coherence literature and examine emotional coherence in channels that have yet to be investigated. Exploring whether emotional coherence is associated with clients' well-being over the course of therapy may also lead to important conclusions about the diagnostic value of emotional coherence. This work describes an end-to-end approach where a transformer-based Hebrew language model was used to develop utterance-level predictions of client emotions. These automated emotion predictions were then used to study emotional coherence and its associations with client well-being measures. The main contributions of this work are:

- This is the first study to assess the emotional coherence between the subjective experience of client emotions and the verbal expression of their emotions.
- The automated predictions from the transformer model can be applied to scale-up psychotherapy research and investigate therapy dialogue at the utterance-level.
- The association between the clients' subjective experience of emotions and the model's predictions of clients' emotions validates the ability of this approach to effectively capture clients' emotional processes in therapy dialogues.
- Compared to previous studies that have assessed emotional coherence at one time point and in lab-based tasks, the current study used longitudinal data to assess how emotional coherence unfolds from session to session in genuine interactions between clients and therapists.

2 Methods

The following hypotheses were examined:

1. There should be a positive correlation between verbal expression of emotions and clients' subjective experience of emotions.
2. Higher emotional coherence should be associated with clients' better well-being.

2.1 Dataset

We study a data-set comprised of 872 sessions collected as part of a university clinic's regular practice of monitoring clients' progress [Atzil-Slonim et al., 2021]. The language of therapy was Modern Hebrew. A total of 196 sessions from the original dataset were annotated for emotion labels at the utterance level by clinical experts i.e. **BIU-872_Gold**. These emotion labels were coded as: *Negative*, *Positive*, *Neutral:* which was defined as neither positive nor negative emotions and *Mixed:* which was defined as both positive and negative emotions [Greenberg, 2012]. The remainder of the un-annotated sessions were called **BIU-872_Silver**.

After each session, clients self-reported their emotional experience using the Profile of Mood States (POMS; [Cranford et al., 2006]). The POMS consists of 12 words aggregated to describe current negative (e.g, sad) or positive (e.g., happy) emotional states. At the beginning of each session, the clients also completed the Outcome Rating Scale (ORS; [Miller et al., 2003]) to assess their well-being. It consists of four visual analog scale ranging from 0 to 10.

2.2 Labelling Client Emotions for BIU-872_Silver

2.2.1 Task Definition

BIU-872 is a conversation dataset with two speakers i.e. *the client* (C) and *the therapist* (T) and the task is to label client emotions at the utterance-level. Formally, given an input sequence of *N* utterances $[\mathrm{u}_1{}^{\mathrm{p}}, \mathrm{u}_2{}^{\mathrm{p}}.......\mathrm{u}_{\mathrm{N}}{}^{\mathrm{p}}]$, where p=[C, T] and each utterance $\mathrm{u}_{\mathrm{i}}{}^{\mathrm{p}}=[\mathrm{u}_{\mathrm{i},1}, \mathrm{u}_{\mathrm{i},2},........\mathrm{u}_{\mathrm{i},\mathrm{T}}]$ has *T* words $u_{i,j}$ spoken by party *p*, the task is to label emotions only for client utterances $[\mathrm{u}_1{}^{\mathrm{C}}, \mathrm{u}_2{}^{\mathrm{C}}.......\mathrm{u}_{\mathrm{N}}{}^{\mathrm{C}}]$.

2.2.2 Model

To study client emotions, we adapted AlephBERT on a downstream classification task of emotion recognition (ER). We chose AlephBERT for domain training because a) it achieves state of the art (SOTA) results on benchmark datasets for HebrewNLP tasks b) it is pre-trained on a large Hebrew corpus of 52K [Seker et al., 2021] and we are also working with a dialogue dataset in Modern Hebrew. We trained AlephBERT on BIU-872_Gold over 10-fold cross validation to account for the variability in results. Hyper-parameter tuning was done on development dataset (10% of BIU-872_Gold 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 due to skewed emotion label distribution [Chawla et al., 2002]. In line with previous ER studies in psychotherapy, we evaluated the results from the trained model with the F1 micro [Tanana et al., 2021]. We then used our BIU-872_Gold trained AlephBERT model to label client emotions for BIU-872_Silver.

2.3 Emotional Coherence as a diagnostic tool

2.3.1 POMS vs Utterance-level emotions

In the next step, we tested *hypothesis 1* which evaluated coherence between the self-reported client emotions i.e. POMS $\mathbf{P_e}^{\mathbf{m}}$ (cumulative) and utterance-level (i.e. verbal expression) client emotions $\mathbf{U_e}^{\mathbf{m}}$ (normalized), where e = emotion labels (pos, neg) and m = session number. As mentioned in Section 2.1, there are four utterance level coding in the dataset, but we only focus on *positive* and *negative* emotions to study any meaningful change in patient behavior.

Total utterance size and emotion labels varied in count both within and across sessions. To reduce the disparity between long and short sessions and to understand the real significance of each emotion throughout session *m*, we performed emotion label normalization for each session as follows:

$$U_{\mathrm{e}}{}^{\mathrm{m}} = \frac{\#u_{\mathrm{e}}{}^{\mathrm{m}}}{\sum_{i\subset[pos,neg,mix,neu]} \#u_{\mathrm{i}}{}^{\mathrm{m}}} \tag{1}$$

where $\#\mathrm{u_e}{}^{\mathrm{m}}$ is the number of emotion *e* across all utterances for session *m*.

The client POMS was evaluated on six different emotion sub-scales. We collated these sub-scales for *pos*=[calmness,contentment,vigor] and *neg*=[anger,sad,anxiety] emotions to study POMS in a binary emotion setting:

$$P_{\mathrm{e}}{}^{\mathrm{m}} = \sum_{e\subset[pos,neg]} p_{\mathrm{e}}{}^{\mathrm{m}} \tag{2}$$

	BIU-872_Gold	BIU-872_Silver
(P_{pos}, U_{pos})	(0.29,7.8e-05)	(**0.27, 4.3e-12**)
(P_{neg}, U_{neg})	(0.24, 0.001)	(**0.21, 4.1e-08**)

(a) Session-wide Correlation between POMS and Utterance emotion labels

	ORS
Cohr(U_{pos}, P_{pos})	(**0.67, 0.048**)
Cohr(U_{neg}, P_{neg}))	(-0.37, 0.321)

(b) Client-level correlation between Emotional coherence and ORS

Table 1: Correlation analysis on BIU-872

where p_e^m is the total score for emotion sub-scale *e* for the entire session *m*.

Coherence between the POMS and utterance-level emotions was measured for both positive and negative emotions. We used the Pearson implementation from the Scipy library to calculate the correlations i.e. Cohr(U_e, P_e) where *e*=[pos, neg] and the corresponding significance values [Benesty et al., 2009, Kowalski, 1972, Virtanen et al., 2020]. Coherence experiments with BIU-872_Gold aimed at testing of *hypothesis 1* by using expert coded emotion labels as the verbal emotion expression. Then we examined emotional coherence on BIU-872_Silver for application results, where the emotion labels are developed from the transformer-based model. Significance cutoff for correlation results was set at α=0.05.

2.3.2 Association between Emotional Coherence and ORS

In order to test *hypothesis 2*, we calculated correlation between emotional coherence and ORS at the client-level across treatment. The session-wide results for each client *'l'* were first summarized using mean (μ) and then the Pearson correlation was used to calculate the association i.e. $Corr(\mu(Cohr(U_e^l, P_e^l)), \mu(ORS_e^l))$, where e=[pos, neg] and l=client_id.

3 Results and Discussion

3.0.1 Emotion labels for BIU-872_Silver

The test data for BIU-872_Gold prediction results from the transformer model achieved a moderate F1 of 0.66 for ER in psychotherapy. We adapted the pre-trained model to develop utterance-level emotion labels for BIU-872_Silver. The pre-trained ER model allows psychotherapy dialogues to be studied with more granularity at the dialogue-level and on a larger scale than previously examined. While in the current study this model was used to explore emotional coherence, the automatic labeling of emotions also opens up the possibilities to study a wide range of other emotional processes in psychotherapy.

3.0.2 Emotional Coherence between self-reports and verbal expression

The results on BIU-872_Gold in Table 1a show there is a significant and positive correlation between P_{pos} and U_{pos} (0.29) and P_{neg} and U_{neg} (0.24) across all the sessions. This result was based on expert annotated emotion labels, and therefore it validates our *hypothesis 1* using traditional psychotherapy analysis. These results are consistent with previous studies that have reported coherence across various emotional response systems (e.g., [Brown et al., 2020]), but extend beyond them by showing that coherence also occurs between emotional experience and verbal emotion expression.

Table 1a also depicts a significant positive correlation between P_{pos} and U_{pos} (0.27) and P_{neg} and U_{neg} (0.21) for BIU-872_Silver. These results validate the performance of the transformer-based ER approach to automatically detect genuine emotions from text with specificity. This result further underscores the usefulness of this transformer-based model in detecting emotional coherence on a larger scale.

3.0.3 Association between emotional coherence and well being

The results in Table 1b indicate significant results (0.048) with a high correlation (0.67) between Cohr(U_{pos}, P_{pos}) and the ORS measure. This finding supports a largely untested theoretical claim that the emotional coherence between subjective experience and the verbal expression of emotions is important for clients' well-being [Greenberg, 2012, Lane et al., 2015]. The findings of association between emotional coherence and well-being only for positive emotions is in line with previous

studies reporting similar results with other emotional channels [Mauss et al., 2005].
One possible explanation for this finding is that negative emotions tend to be more salient than positive emotions [Baumeister et al., 2001], and hence may be more easily recognized. However, better well-being appears to be achieved when clients express and at the same time recognize their positive emotions.
Our results highlight that by using NLP-based ER models, clinicians may be better able to identify clients or sessions characterized by a high dissociation between emotional experience and verbal expression of emotions and direct their interventions to help clients accordingly.

Ethics Statement

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

Acknowledgments

From UKP Lab, this work has been funded by the LOEWE Distinguished Chair "Ubiquitous Knowledge Processing" (LOEWE initiative, Hesse, Germany). From PR Lab, this work was supported by the Israel Science Foundation, (ISF #2466/21 to Dana Atzil-Slonim).

References

D. Atzil-Slonim, D. Juravski, E. Bar-Kalifa, E. Gilboa-Schechtman, R. Tuval-Mashiach, N. Shapira, and Y. Goldberg. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy*, 2021.

R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs. Bad is stronger than good. *Review of general psychology*, 5(4):323–370, 2001.

J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

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

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

J. A. Cranford, P. E. Shrout, M. Iida, E. Rafaeli, T. Yip, and N. Bolger. 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, 2006.

D. Fosha. The dyadic regulation of affect. *Journal of clinical psychology*, 57(2):227–242, 2001.

J. Gibson, D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan. Attention networks for modeling behaviors in addiction counseling. pages 3251–3255, 2017.

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

P. D. Hastings, J. N. Nuselovici, B. Klimes-Dougan, K. T. Kendziora, B. A. Usher, M.-h. R. Ho, and C. Zahn-Waxler. 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, 2009.

Z. E. Imel, M. Steyvers, and D. C. Atkins. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, 52(1):19, 2015.

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

R. D. Lane, L. Ryan, L. Nadel, and L. Greenberg. Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science. *Behavioral and brain sciences*, 38, 2015.

R. W. Levenson. The autonomic nervous system and emotion. *Emotion review*, 6(2):100–112, 2014.

M. Lohani, B. R. Payne, and D. M. Isaacowitz. Emotional coherence in early and later adulthood during sadness reactivity and regulation. *Emotion*, 18(6):789, 2018.

I. B. Mauss, R. W. Levenson, L. McCarter, F. H. Wilhelm, and J. J. Gross. The tie that binds? coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2):175, 2005.

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

C. Negrao, G. A. Bonanno, J. G. Noll, F. W. Putnam, and P. K. Trickett. Shame, humiliation, and childhood sexual abuse: Distinct contributions and emotional coherence. *Child maltreatment*, 10 (4):350–363, 2005.

A. Seker, E. Bandel, D. Bareket, I. Brusilovsky, R. S. Greenfeld, and R. Tsarfaty. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*, 2021.

A. B. Shatte, D. M. Hutchinson, and S. J. Teague. Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, 49(9):1426–1448, 2019.

M. J. Tanana, C. S. Soma, P. B. Kuo, N. M. Bertagnolli, A. Dembe, B. T. Pace, V. Srikumar, D. C. Atkins, and Z. E. Imel. How do you feel? using natural language processing to automatically rate emotion in psychotherapy. *Behavior research methods*, 53:2069–2082, 2021.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

A. W. Wagner, L. Roemer, S. M. Orsillo, and B. T. Litz. Emotional experiencing in women with posttraumatic stress disorder: Congruence between facial expressivity and self-report. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, 16(1):67–75, 2003.