Working Alliance Transformer for Psychotherapy Dialogue Classification

Baihan Lin[1], **Guillermo Cecchi**[2], **Djallel Bouneffouf**[2]
[1]Icahn School of Medicine at Mount Sinai, New York, NY
[2]IBM TJ Watson Research Center, Yorktown Heights, NY
baihan.lin@mssm.edu, {gcecchi@us., djallel.bouneffouf@}ibm.com

Abstract

As a predictive measure of the treatment outcome in psychotherapy, the working alliance measures the agreement of the patient and the therapist in terms of their bond, task and goal. Long been a clinical quantity estimated by the patients' and therapists' self-evaluative reports, we believe that the working alliance can be better characterized using natural language processing technique directly in the dialogue transcribed in each therapy session. In this work, we propose the Working Alliance Transformer (WAT), a Transformer-based classification model that has a psychological state encoder which infers the working alliance scores by projecting the embedding of the dialogues turns onto the embedding space of the clinical inventory for working alliance. We evaluate our method in a real-world dataset with over 950 therapy sessions with anxiety, depression, schizophrenia and suicidal patients and demonstrate an empirical advantage of using information about therapeutic states in the sequence classification task of psychotherapy dialogues.

1 Introduction

The working alliance between the therapist and the patient is an important measure of the clinical outcome and a qualitative predictor of therapeutic effectiveness in psychotherapy (Wampold, 2015; Bordin, 1979). The alliance entails a number of cognitive and emotional aspects of the interaction between these two agents, such as their shared understanding of the objectives to be attained and the tasks to be completed, as well as the bond, trust, and respect that will develop during the course of the therapy. While traditional methods to quantify the alliance depend on self-evaluative reports with point-scales valuation by patients and therapists about whole sessions (Horvath, 1981), the digital era of mental health can enable new research fronts utilizing real-time transcripts of the dialogues between the patients and therapists. By analyzing the psychotherapy dialogues, we are interested in studying the usage of natural language processing technique to extract out turn-level features of the working alliance and see if it can help better inform us of the clinical condition of the patient.

Here we present Working Alliance Transformer (WAT), a transformer-based classification model to classify the psychotherapy sessions into different psychiatric conditions. Our methods consists of a psychological state encoder that quantifies the degree of patient-therapist alliance by projecting each turn in a therapeutic session onto the representation of clinically established working alliance inventories, using language modeling to encode both turns and inventories, which was originally proposed in (Lin et al., 2022) as an analytical tool. This allows us not only to quantify the overall degree of alliance but also to identify granular patterns its dynamics over shorter and longer time scales.

We collated and preprocessed the Alex Street Counseling and Psychotherapy Transcripts dataset (Street, 2023), which consists of transcribed recordings of over 950 therapy sessions between multiple anonymized therapists and patients that belong to four types of psychiatric conditions: anxiety, depression, schizophrenia and suicidal. (The data publisher mentions that they have more clinical conditions other than the analyzed 4 classes, but due to the licensing and access limitations, we can only obtain the 4 classes we presented.) This multi-part collection includes speech-translated transcripts of the recordings from real therapy sessions, 40,000 pages of client narratives, and 25,000 pages of reference works. As open science and data sharing initiatives in the psychiatry domains become more prominent, we believe our methodologies can be adapted in a responsible way to a broader spectrum of clinical conditions. On this dataset, we evaluate quantitatively the effectiveness of this inference method in improving the classification / diagnosis capability of deep learning models to linguistically

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 64–69

predict psychiatric conditions from therapy transcripts. Lastly, we discuss how our approach may be used as a companion tool to provide feedback to the therapist and to augment learning opportunities for training therapists.

2 Methods

We describe our pipeline in Figure 1. Given the transcripts of a therapy session and the medical records of the patient. The dialogue are separated into pairs of turns as the timestamps. We can either choose to only use the turns by the patients, or by the therapists, or use both, as a paired input. Empirically, the patients' turns are usually more narrative, as they are describing themselves, while the therapists' turns are usually more declarative, as they are usually confirming the patients, or leading conversations to certain topic.

Each patient response turn S_i^p followed by a therapist response turn S_i^t is treated as a dialogue pair. In total, these materials include over 200,000 turns together for the patient and therapist and provide access to the broadest range of clients for our linguistic analysis of the therapeutic process of psychotherapy. On the other hand, we have access to the Working Alliance Inventory (WAI), the clinical instrument. The modern WAI consists of 36 statements in a self-report questionnaire which measures the therapeutic bond, task agreement, and goal agreement (Horvath, 1981; Tracey and Kokotovic, 1989; Martin et al., 2000), where the Since the original 12-item version (Tracey and Kokotovic, 1989), the inventory has used parallel versions for clients and therapist with good psychometric properties and helped establish the importance of therapeutic alliance in predicting treatment outcomes. The modern version of the inventory consists of 36 questions, where the rater (i.e. the patient or the therapist) is asked to rate each statement on a 7-point scale (1=never, 7=always)(Martin et al., 2000). This inventory is disorder-agnoistic, meaning that it measures the alliance factors across all types of therapies, and provides a record of the mapping from the alliance measurement and the corresponding cognitive constructs underlying the measurement under a unified theory of therapeutic change (Horvath and Greenberg, 1994).

The inference goal is to compute a score that characterizes the working alliance given the clinical inventory, with for instance, a feature vector of 36 dimension that correspond to the 36 alliance measure of interests in the inventory. Operationally, the goal is to derive from these 36 items three alliance scales: the task scale, the bond scale and the goal scale. They measures the three major themes of psychotherapy outcomes: (1) the collaborative nature of the patient-therapist relationship; (2) the affective bond between therapist and patient, and (3) the therapist's and patient's capabilities to agree on treatment-related short-term tasks and long-term goals. The score corresponding to the three scales comes from a key table which specifies the positivity or the sign weight to be applied on the questionnaire answer when summing in the end. The full scale is simply the sum of the scores of the three scales. The key table is like a weighting matrix that specifies the directionalities of the scales. After computing the information regarding the predicted clinical outcome with our inferred working alliance scores, this feature vector highlights a bias towards what the clinicians would care about in the psychotherapy given the metrics provided by the working alliance inventory. We would then able to further use this information to potentially inform us of the psychiatric condition of a given patient. As such, we propose the Working Alliance Transformer (WAT), a classification model that utilizes an inference module that informs the downstream classifier where the current state is with respect to the therapeutic trajectory or landscape in the psychotherapy treatment of this patient. Is this patients approaching a breakthrough? Or is he or she susceptible to a rupture of trust? These therapeutic information about alliance can vary across clinical conditions, and thus, potentially beneficial to the

Algorithm 1 Working Alliance Transformer (WAT)

Input: a session with T turns
Output: a label for psychiatric condition
for i = 1,2,$\cdots$, T **do**
  Transcribe dialogue turn pairs (S_i^p, S_i^t)
  for $(I_j^p, I_j^t) \in$ inventories (I^p, I^t) **do**
    $W_j^{p_i}$ = similarity$(Emb(I_j^p), Emb(S_i^p))$
    $W_j^{t_i}$ = similarity$(Emb(I_j^t), Emb(S_i^t))$
  end for
  (Patient) $x_c = concat(Emb(S_i^t), W^{p_i})$
  (Therapist) $x_t = concat(Emb(S_i^t), W^{t_i})$
  (Dyad) $x = concat(x_t, x_c)$
  Aggregated feature $X.append(x)$
end for
obtain prediction $y = Transformer(X)$

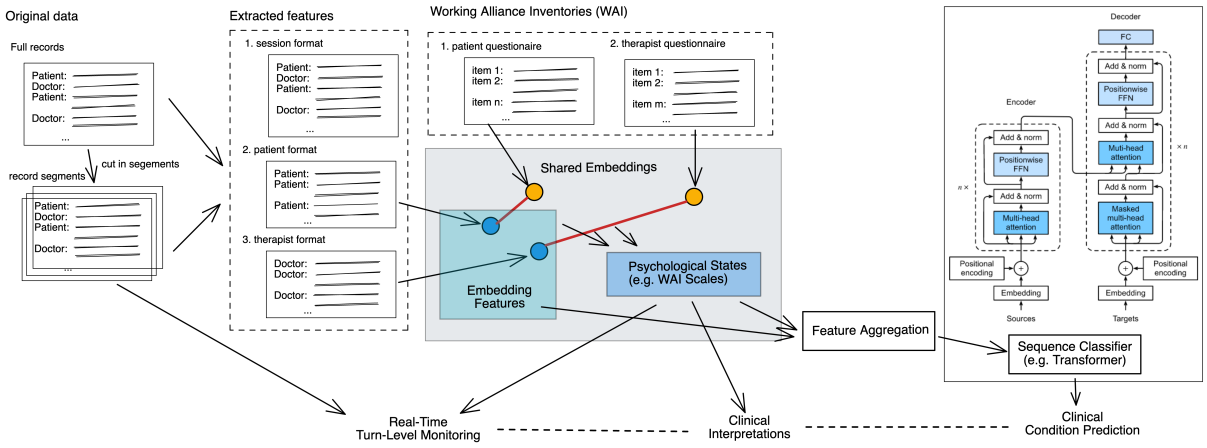

Figure 1: Architecture of working alliance transformer for psychiatric condition classification using the psychological state encoder from working alliance

diagnosis and monitoring of psychiatric disorders.

Algorithm 1 outlines the classification process. During the session, the dialogue between the patient and therapist are transcribed into pairs of turns. We denote the patient turn as S_i^p followed by the therapist turn S_i^t, as a dialogue pair. Similarly, the inventories of working alliance questionnaires come in pairs (I^p for the patient, and I^t for the therapist, each with 36 statements). We compute the distributed representations of both the dialogue turns and the inventories with the sentence embeddings. The working alliance scores can then be computed as the cosine similarity between the embedding vectors of the turn and its corresponding inventory vectors. Following (Lin et al., 2022), we use SentenceBERT (Reimers and Gurevych, 2019) and Doc2Vec embedding (Le and Mikolov, 2014) as our sentence embeddings for the working alliance inference. With that, for each turn (either by patient or by therapist), we obtain a 36-dimension working alliance score. For the classification, we concatenate the 36-dimension working alliance scores computed from the current turn in the dialogue, along with the sentence embedding of the current turn, as our feature vector to fed into our Transformer sequence classifier.

The analytical features enabled by the working alliance inference are not only useful for the classification we investigate in this study but also other downstream tasks, such as predictive modeling and real-time analytics. In our case, the turns in a dialogue or monologue are fed into the sentence embedding sequentially as individual entries. And then, given the sentence embedding, we feed them each into the psychological state encoder that infer the psychological or therapeutic state of the dialogue at this turn. The encoder will generate a vector that characterizes the state, such as the 36-dimension working alliance scores, corresponding to the 36 working alliance inventory items. Then, the model aggregate both the sentence embedding feature vector and the psychological state vector. In this case, we concatenate them together as a first step. Since we feed our input sentence by sentence (or turn by turn), we have a sequence of combined feature vector, which is then fed into a sequence classifier. We use the transformer (Vaswani et al., 2017) as our classifier for its effectiveness in various sequence-based learning tasks, and potential interpretability from its attention weights. The output of this classification model is the predicted clinical condition of this sequence. The sequence of turns we feed to generate a label is a trimmed segment of the session of psychotherapy transcript.

3 Results

Here we present the transcript classification results.

Experimental setting. The psychotherapy dataset we evaluate is highly *imbalanced* across the four clinical conditions (495 anxiety sessions, 373 depression sessions, 71 schizophrenia sessions, and 12 suicidal sessions). If we directly train our models on this dataset, the classifier is likely to be highly biased towards the majority class. To correct for this imbalanceness issue, we are using the sampling technique. Instead of going through the entire training data in epochs, we train the models in sampling iterations. In each iteration we randomly choose a class and then randomly sample one session from the class pool. Before we sample

Table 1: Classification accuracy (%) of psychotherapy sessions

	SentenceBERT			Doc2Vec		
	Patient turns	Therapist turns	Both turns	Patient turns	Therapist turns	Both turns
WAT (working alliance embedding)	**27.6**	**27.0**	**26.0**	**34.1**	25.7	**31.9**
WAT (working alliance score)	26.1	23.4	25.5	28.9	23.7	**31.9**
Embedding Transformer	24.8	24.0	25.5	31.8	**26.2**	29.9
WA-LSTM (working alliance embedding)	**35.0**	**36.9**	**23.3**	**46.0**	27.7	29.6
WA-LSTM (working alliance score)	24.5	34.2	22.6	30.2	24.7	**43.4**
Embedding LSTM	23.0	36.0	22.9	44.3	**31.1**	31.1

the sessions, we split the dataset into 20/80 as our test set and training set. Then during the training or the test phase, we perform the sampling technique for each iteration only in the fully separated training and test sets. Then, for each sampled session, we feed into the classification model the first 50 dialogue turns of our transcript, turn by turn, and the sequence classifier will output a label predicting which psychiatric condition this session belongs to.

Model architecture. We evaluate two classifier backbones. The first one is the classical transformer model. For the multi-head attention module, we set the number of heads to be 4 and the dimension of the hidden layer to be 64. The dropout rates for the positional encoding layer and the transformer blocks are both set to be 0.5. The second backbone is a 64-neuron Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997).

Ablation and baseline models. For each of the two classifiers, we compare three types of features as the input we feed into the sequence classifier component. The first one, the working alliance embedding, is the concatenated feature vector of both the sentence embedding vector and the psychological state vector (which in our case, is the 36-dimension inferred working alliance scores). The second type of feature, the working alliance score, is an ablation model which only uses the state vector (the working alliance score vector). The third type of feature, the embedding, is the baseline which only uses the sentence embedding vector directly. In other words, The working alliance score introduces the bias for WAI. The sentence embedding doesn't. The working alliance embedding is the feature that combines both with concatenation. And since we have two sentence embeddings to choose from (the sentence BERT and Doc2Vec), they each have 9 models in the evaluation pool. Other than the classifier types (Transformer or LSTM), the embedding types (Sentence-BERT or Doc2Vec) and the feature types (working alliance embedding, working alliance scores, or simply sentence embedding), we also compared using only the dialogue turns from the patients, from the therapists, and from both the patients and the therapists. In the case where we use the turns from both the patients and the therapists, we consider them as a pair, and concatenate them together as a combined feature. This is as opposed to treating them as subsequent sequences, because we believe that the therapist's response are loosely semantic labels for the patient's statements, and thus, serve different semantic contexts that should be considered side by side, instead of sequentially, which would assume a homogeneity between time steps.

Training procedure. For all 12 models, we train them for over 50,000 iterations using the stochastic gradient descent with a learning rate of 0.001 and a momentum of 0.9. Since the training set is relatively small for our neural network models, we observe some of the models exhibit overfitting at early stages before we finish the training. As a result, we report the performance of their checkpoints where they converge and have a plateau performance. Then in the testing phase, we randomly sample class-balanced 1,000 samples.

Empirical results. We report the classification accuracy as our evaluation metrics. Since we have four classes, and the evaluation is corrected for imbalanceness with the sampling technique.

Overall, we observe a benefit of using the working alliance embedding as our features in Transformer and LSTM-based model architectures. Among all the models, the WA-LSTM model with working alliance embedding using only the patient turns obtains the best classification result (46%), followed by the WA-LSTM model using only the working alliance score using both turns from the patients and therapists (43.4%). This suggest the advantage of taking into account the predicted clinical outcomes in characterizing these sessions given their clinical conditions. We also notice that the inference of the therapeutic working alliance with Doc2Vec appears to be more beneficial in model-

ing the patient turns than the therapist turns, while the SentenceBERT-based inference appears to be advantageous in both therapist and patient features.

Comparing the two sequential learners, the Transformer, due to the additional attention mechanism, yields a more stable learning phase. When using SentenceBERT as its embedding, we observe a modest benefit when training on only the patient turns, which might suggest an interference of features between the therapists' and patients' working alliance information. The Transformers using the working alliance embedding, i.e. both the sentence embedding and their therapeutic states (i.e. the inferred working alliance score vector) are the best performing ones. When using Doc2Vec as the embedding, the best performing models are both the Transformers using some of the working alliance information from our inference module as features.

4 Discussion

Our analytic approach reveals insightful features of therapeutic relationship and their usefulness in terms of clinical diagnosis merely based on the patient-doctor conversations. In our prior work, we observe systematic differences in the mean inferred alliance scores between patients and therapists, and also across disorders (Lin et al., 2022). However the in-session evolution of the inferred scores provide a much more interesting perspective, as shown in our dialogue sequence classification results. In particular, while all conditions show a systematic misalignment of scores between patients and therapists, this is significantly starker for suicidality, something observable in the mean as well as in the time trace for full and sub-scales, which can be useful for early detection of suicidal thoughts.

As more and more successful applications of AI are deployed in clinical domains, there are many ethical considerations we practitioners of machine learning should be aware of and take into considerations. When dealing with patient data, the privacy and security is a top priority. Following the suggestion of best practices from (Matthews et al., 2017), all examples in this paper as well as the dataset we analyzed are properly anonymized with pre- and post-processing techniques. In addition, the dataset itself was sourced with proper license from ProQuest's Alexander Street platform. We remove all personally identifiable information (meta data, user name, identifiers, doctors' name) from the dataset.

Since the clinical domain of this work is mental health and psychological well-being, there are additional ethical considerations. Emerging techniques in wearable devices, digital health records, brain imaging measurements, smartphone applications and social media are gradually transforming the landscape of the monitoring and treatment of mental health illness. However, most of these attempts are proof of concept as identified by this review (Graham et al., 2019), and requires extensive caution to prevent from the pitfall of over-interpreting preliminary results. The limitations of these prior studies, including our work here, reside in the difficulty of a systematic clinical validation and a uncertain future expectation of the technological readiness for patient care and therapeutic decision making approved by authorities. For instance, it was recently shown that despite the high predictability of statistical learning-based methods in analyzing large datasets in support of clinical decisions in psychiatry, existing machine learning solutions is highly susceptible to overfitting in realistic tasks which has usually a small sample sizes in the data, missing data points for some subjects, and highly correlated variables (Iniesta et al., 2016). These properties in real-world applications limits the out-of-sample generalizability of the results.

5 Conclusions

In this work, we present a Transformer-based classification model that characterizes the sequence of therapeutic states as beneficial feature to improve the classification of psychological dialogues into different psychiatric conditions. It combines the domain expertise from clinically validated psychiatry inventories with the distributed deep representations of language modeling provide a turn-level encoding of working alliance at a turn-level resolution. We demonstrate on a real-world psychotherapy dialogue dataset that using this additional granular representation of the interaction dynamics between patients and therapists is beneficial both for interpretable post-session insights and linguistically diagnosing the patients.

Our results suggest that the inferred scores of therapeutic or psychological states of patient-doctor alignment can be useful in downstream tasks, such as diagnosis. Although not a main focus in this work, future work would include a more systematic investigation of such downstream tasks, and exploiting the attention mechanism of the transformer blocks for interpretations.

References

Edward S Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252.

Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21(11):1–18.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Adam O Horvath. 1981. *An exploratory study of the working alliance: Its measurement and relationship to therapy outcome*. Ph.D. thesis, University of British Columbia.

Adam O Horvath and Leslie S Greenberg. 1994. *The working alliance: Theory, research, and practice*, volume 173. John Wiley & Sons.

Raquel Iniesta, D Stahl, and Peter McGuffin. 2016. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological medicine*, 46(12):2455–2465.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2022. Deep annotation of therapeutic working alliance in psychotherapy. *arXiv preprint arXiv:2204.05522*.

Daniel J Martin, John P Garske, and M Katherine Davis. 2000. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, 68(3):438.

Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. 2017. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2189–2201.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alexander Street. 2023. counseling and psychotherapy transcripts series.

Terence J Tracey and Anna M Kokotovic. 1989. Factor structure of the working alliance inventory. *Psychological Assessment: A journal of consulting and clinical psychology*, 1(3):207.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bruce E Wampold. 2015. How important are the common factors in psychotherapy? an update. *World Psychiatry*, 14(3):270–277.