HHS Public Access

Author manuscript

Clin Psychol Sci. Author manuscript; available in PMC 2024 August 05.

Published in final edited form as:

Clin Psychol Sci. 2024 May ; 12(3): 435–446. doi:10.1177/21677026231172694.

Machine-Learning-Based Prediction of Client Distress From Session Recordings

Patty B. Kuo[1], **Michael J. Tanana**[1], **Simon B. Goldberg**[2], **Derek D. Caperton**[3,1], **Shrikanth Narayanan**[4], **David C. Atkins**[5], **Zac E. Imel**[1]

[1]University of Utah

[2]University of Wisconsin, Madison

[3]Calgary Counselling Centre

[4]University of Southern California

[5]University of Washington

Abstract

Natural language processing (NLP) is a subfield of machine learning that may facilitate the evaluation of therapist-client interactions and provide feedback to therapists on client outcomes on a large scale. However, there have been limited studies applying NLP models to client outcome prediction that have (a) used transcripts of therapist-client interactions as direct predictors of client symptom improvement, (b) accounted for contextual linguistic complexities, and (c) used best practices in classical training and test splits in model development. Using 2,630 session recordings from 795 clients and 56 therapists, we developed NLP models that directly predicted client symptoms of a given session based on session recordings of the previous session (Spearman's rho =0.32, p<.001). Our results highlight the potential for NLP models to be implemented in outcome monitoring systems to improve quality of care. We discuss implications for future research and applications.

Keywords

Machine learning; natural language processing; outcome prediction; psychotherapy

Psychotherapy represents a class of mostly conversational treatments focused on behavior change or reduction of client distress. The primary goal of research on psychotherapy is to determine what sorts of conversations are most associated with the desired change. However, it has proven quite difficult to make consistent predictions about the outcome of psychotherapy from the content of the conversation (e.g., Webb et al., 2010). Client ratings of the therapeutic relationship consistently predict response to treatment but are distal to the

Correspondence regarding this article should be addressed to Patty Kuo (patty.kuo@pennmedicine.upenn.edu).
Patty B. Kuo, Department of Educational Psychology, University of Utah. Michael J. Tanana, College of Social Work, University of Utah. Simon Goldberg, Department of Counseling Psychology and Center for Healthy Minds, University of Wisconsin, Madison. Derek D. Caperton, Calgary Counselling Centre; Department of Educational Psychology, University of Utah. Shrikanth Narayanan, Viterbi School of Engineering, University of Southern California. David C. Atkins, Department of Psychiatry, University of Washington. Zac Imel, Department of Educational Psychology, University of Utah.

session itself, and place a burden on clients to repeatedly rate their experience. Our work focuses on evaluating machine learning based technology to make predictions about client symptom change directly from the recording of a session.

Measurement based psychotherapy relies on obtaining clinically meaningful information from clients, observers, or therapists based on process and outcome ratings. In measurement based care (MBC), therapists use client ratings to guide discussion regarding treatment goals and approach (Fortney et al., 2017; Goldberg et al., 2018; Lewis et al., 2019; Scott & Lewis, 2015; Shimokawa et al., 2010). Similarly, psychotherapy research has relied on client, therapist, and observer ratings of clinical processes to predict client outcomes. For example, researchers have used therapist and client self-report measures to assess factors such as therapeutic alliance (Flückiger et al., 2018) and multicultural competency (Tao et al., 2015). Similarly, researchers have used behavioral coding to examine therapist adherence to specific interventions (Webb et al., 2010) and emotional expression (Elliott et al., 2011). While traditional MBC generally has increased feedback on therapeutic processes and quality of care, methodological issues continue to limit therapists receiving feedback on client outcomes based on clinical interactions. Clients are often asked to retrospectively rate their experiences in sessions; consequently, client ratings may be more indicative of their general experiences in therapy rather than specific interactions with therapists. Therapist self-report measures similarly do not provide information on how specific therapist-client interactions drive client outcomes, and are often less correlated with outcomes than client self-report ratings (e.g. Elliott et al., 2011; Tao et al., 2015). Observer ratings of sessions can provide valuable insight into how therapist-client interactions inform client symptom improvement, but behavioral coding is often costly and time intensive (e.g., Moyers et al., 2003). Implementing routine outcome monitoring in large healthcare systems also results in increased time burdens for clients to complete measures, and therapists to collect data. Given current limitations in assessment of therapeutic processes, alternative methods of predicting client symptom improvement directly from therapist-client interactions are needed in order to scale up research on how therapist behaviors decrease client distress, and facilitate provision of feedback to therapists in a manner that minimizes client and therapist burden.

Natural language processing (NLP) is a powerful alternative to traditional methodologies that rely on self-report measures and behavioral coding to assess client outcomes and therapeutic processes. NLP is a subfield of machine learning (ML) that integrates linguistics and computer science to automatically make statistical predictions based on patterns in large bodies of text (Hirschberg & Manning, 2015; Jurafsky & Martin, 2014). NLP techniques derive these predictions by converting text into smaller units (e.g., phrases - sometimes called n-grams), assigning quantitative metrics to units of text, and examining how patterns in these quantitative metrics are associated with specific outcomes of interest (e.g., topics, interventions, emotion, changes in symptoms). NLP classification models are particularly useful because they analyze large bodies of text in relatively short time periods; efforts to manually analyze the same bodies of text using traditional coding methods would be slower and require extensive human resources. NLP technologies have been used in wide-ranging applications, from detecting urgency in spoken utterances (Landesberger et al., 2020) to examining themes in poetry (Kao & Jurafsky, 2012).

NLP classification has been leveraged to describe and predict therapeutic processes directly from the words spoken by therapists and clients, rather than rely on client, therapist, or observer ratings of clinical processes. For example, topic models have been used to automatically assess motivational interviewing fidelity (Atkins et al., 2014) and classify types of psychotherapy treatment used by providers (Imel, Steyvers, et al., 2015). In addition, NLP models have been used to predict the sentiment of therapist and client statements (Syzdek, 2020; Tanana et al., 2016; Tanana et al., 2021). Furthermore, NLP has shown promise in predicting therapeutic processes such as therapist empathy (Xiao et al., 2012, 2015) and therapeutic alliance (Goldberg et al., 2020).

Past application of NLP-based tools to identify therapeutic processes suggest that NLP models can be similarly used to automatically predict client symptom improvement from in-session content. NLP based tools that predict client outcomes have the potential to be implemented in large healthcare systems to automatically monitor symptoms directly from session data, rather than relying on self-report measures that increase burden on clients and therapists. Three recent studies have used computationally derived ratings of important therapeutic processes to predict treatment outcomes. Ewbank et al., 2020 applied a deep learning model to classify therapist messages during an Internet-enabled text-based cognitive behavioral therapy delivered via synchronous messaging (n = 17,572). They grouped therapist messages into categories (e.g., use of change methods, arranging next session) and found several features had small associations with treatment improvement, including use of change methods and therapeutic praise (odds ratios = 1.11 and 1.21, respectively). In a sample of 729 psychotherapy sessions, Shapira et al., 2021 found that lower frequency of negative emotion words (using a pre-defined dictionary) in a given session was associated with lower distress at the subsequent session. In addition, a decline in first-person singular words was associated with pre- to post-treatment improvement. Finally, Atzil-slonim et al. (2021) applied topic models to transcripts of 873 sessions, and used model derived topics to examine how changes in topics of conversation between therapists and clients predicted symptom change; they found that increases in topics associated with higher client functioning were associated with greater symptom improvement.

Ewbank et al. (2020), Shapira et al. (2020), and Atzil-Slonim et al. (2021) demonstrate the promise of using session content to not only scale up the evaluation of session content, but also derive important metrics to predict outcomes. Application of NLP models to identify therapist interventions and emotional valence allows for more detailed understanding on a large scale of how these components inform client change. At once, these studies could be further methodologically strengthened. First, past studies used NLP models to first label therapist-client interactions, then used separate analyses to correlate client outcomes with the labels rather than raw text. Directly predicting client outcomes from therapist-client interactions may capture more nuanced linguistic patterns that are important to client symptom change, but uncaptured through the particular human conceived labels. Topic modeling approaches, such as those used in Atzil-Slonim (2021), are effective in identifying interpretable text patterns without relying on prior human conceived schemes; however, topic models often do not take into account nuanced meaning and context in text that can affect the topics that are generated (Barde et al., 2017).

Second, previous studies developed NLP models that used dictionary based and word2vec methods to identify labels of interest; recent models incorporating more contextual information that could improve measurement of client symptoms. NLP researchers have recently developed more advanced models such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) that account for (a) more information from text corpora of interest, and (b) language specific complexities (e.g. shifts in word meaning based on context). Models such as BERT have been found to exceed performance of older models that use dictionary based methods and word2vec (Dai et al., 2019; Konstantinov et al., 2021). Third, Ewbank et al.'s models were applied to text-based therapy, and similar models may benefit from being trained on spoken language interactions. Although text-based therapy can be effective in reducing client distress (Aguilera et al., 2017; Marcelle et al., 2019), interventions provided via text may unfold differently compared to traditional spoken interventions. Furthermore, the vast majority of mental health care remains delivered via spoken language, which presents a number of additional analytic complexities related to speech processing. Fourth, previous studies did not apply classical ML test-train splits to evaluate model accuracy in predicting outcome. ML researchers typically train and test models on distinct subsets of their data in order to ensure that final model evaluation results are not due to models identifying outcomes based on unique features in the data (Jurafsky & Martin, 2014). Using the same sample for testing and training could result in models that are overfitted to the primary text corpora, and perform poorly in other text corporas. Finally, Shapira et al. and Atzil-Slonim et al. used text generated by human transcribers; the application of models derived from human transcriptions (which take longer than a client simply filling out a symptom measure) could be strengthened with the use of automated transcriptions. Recent advances have improved automated transcription of recordings through automated speech recognition, where NLP tools quickly segment and convert therapist and client speech into text. While there may be some error in how NLP tools convert speech to text, application of NLP tools to transcribe session recordings allows for examination of therapist-client interactions on much larger scales. Reliance on human-generated transcripts is time consuming, and significantly limits the scalability of researching how session content informs client improvement. Thus, evaluations of NLP to predict treatment outcomes directly from recordings are warranted.

Purpose of the Study

The use of NLP methods to predict treatment outcomes directly from clinical encounters has numerous practical applications, including supporting quality assurance efforts, providing feedback to providers, and enriching research efforts seeking to link elements of clinical interactions to treatment effects. Indeed, NLP-based predictions of treatment outcome directly from session recordings could help augment the otherwise limited objective feedback that is routinely available to clinicians providing psychotherapy (Tracey et al., 2014) while minimizing time burden on clients. However, to date there are limited studies that have used transcripts of therapist-client interactions as direct predictors of client symptom improvement. Furthermore, previous applications of NLP models to prediction of client outcomes were limited by (a) reliance on models that do not take into account contextual linguistic features, (b) applications to text-based psychotherapy, and (c) lack of

differentiation between training and test data samples. In order to address limitations in past NLP research in predicting client symptoms, we developed and evaluated NLP models that automatically predicted client symptom ratings of a given session based on transcripts of their previous therapy session using 2,630 transcripts of therapy sessions from 795 clients and 56 therapists. Based on best practices in NLP research, we used separate training, validation, and test samples of transcripts of therapy sessions (Jurafsky & Martin, 2014).

Transparency and Openness

Preregistration

The current study was not preregistered.

Data, materials, code, and online resources

We are unable to provide access to our data and analysis script as our dataset is comprised of transcripts of therapy sessions. Clients in our sample did not consent to having their data made available to the public. Our analysis script was tailored to our dataset.

Reporting

This study involved an analysis of existing data rather than new data collection.

Ethical Approval

This study was approved by the University of Utah's IRB (protocol number 00083132) and was carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

Methods

Overview of Text Corpora

Our data consisted of 2,630 recordings from individual counseling sessions at a large university counseling center. Recordings were collected as part of an NIH funded grant (2 R01 AA018673). Transcripts were generated from recordings using a combination of Kaldi (Povey et al., 2011), a speech recognition toolkit, and another set of models specifically tuned to the language used in psychotherapy, which we will refer to as the 'speech pipeline' (see Flemotomos et al., 2022) for details on the speech recognition, diarization, role assignment and voice activity detection systems). Sessions where both therapists and clients consented to have their sessions recorded were processed with the speech pipeline. We only analyzed sessions where client symptom ratings were available for the next session. Processing steps involved voice activity detection (e.g. identifying whether someone is talking), diarization (e.g. identifying whether the speaker was the client or therapist), and transcription of audio to text (e.g. automated speech recognition).

To create our sample of training and test transcripts, we randomly assigned clients into training, validation, and test sets so that roughly 80% of the transcripts were in the training set, 10% were in the validation set, and 10% were in the test set. Percentage allocation of transcripts to training, validation, and test sets were based on standard NLP practices to

balance (a) maximizing the amount of data that models are trained upon in order to learn various patterns in text, and (b) allowing for varied validation and test sets to adequately assess model performance (for more information, see Jurafsky & Martin, 2014). Clients could only be part of one set. For example, a client who was assigned to the training set could not be in the validation or test set. We were unable to randomize transcripts at the therapist level due to the relatively smaller number of therapists in our sample. Consequently, therapists could have concurrently been included in the training, validation, and test sets. Overall, 2,044 transcripts were assigned to the training set, 314 transcripts were assigned to the validation set, and 272 transcripts were assigned to the test set.

Participants

Client demographics.—Our sample consisted of 1,002 clients. Regarding racial-ethnic identity, 74.2% identified as white, 8.4% as Latinx, 8.3% as Asian American, 5.6% as multi-racial, 1.6% as African American/Black, and 2.0% self-identified their racial-ethnic identity. Regarding gender, 55.2% identified as cisgender women, 40.8% as cisgender men, and 3.4% self-identified their gender identity (e.g. transgender man, transgender woman, gender queer). Regarding sexual orientation, 70.9% identified as heterosexual/straight, 12.2% as bisexual, 5.7% indicated they were questioning their sexual orientation, 5.2% as gay, 3.3% self-identified their sexual orientation (e.g. pansexual, demisexual), and 2.6% as lesbian. Regarding religious and spiritual identity, 27.4% as members of the Church of Jesus Christ of Latter Day Saints, 14.8% self-identified their religious and spiritual identity, 13.8% as atheist, 12.1% as Catholic, 11.5% identified as agnostic, 5.8% as spiritual but not religious, 3.3% preferred not to disclose their religious and spiritual identity, 3.2% as Protestant, 2.3% as Buddhist, 1.9% as Jewish, 1.9% as Lutheran, 1.7% as Hindu, and 1.3% as secular. The average age of clients was 23.3 years (SD=4.7), and client ages ranged from 18 to 54.

Table 1 presents the percentage breakdown of all presenting concerns endorsed by clients in our sample. The top ten client presenting concerns were anxiety (69.7%), depression (63.3%), academic performance (43.1%), self-esteem (41.0%), loneliness (36.4%), social anxiety (33.2%), relationship concerns with partner (26.9%), family of origin (21.8%), relationship concerns with friends (19.9%), and body image (17.1%).

Therapist demographics.—Our sample consisted of 56 therapists. Regarding racial-ethnic identity, 62.8% identified as white, 11.9% identified as Asian, 4.8% identified as Black, 4.8% self-identified their racial-ethnic identity, and 2.4% identified as multi-racial. Regarding gender identity, 63.4% identified as female, 31.7% as male, and 4.9% as gender-queer. Regarding sexual orientation, 64.2% identified as heterosexual/straight, 14.2% self-identified their sexual orientation, 9.5% identified as bisexual, 7.1% identified as lesbian or gay, 4.6% indicated they were questioning their sexual orientation, and 2.4% did not disclose their sexual orientation. The average age of therapists was 32.2 years. Therapists consisted of psychologists, social workers, psychology and social work interns, and psychology and social work practicum students. Therapists in our sample used a variety of psychological interventions stemming from cognitive-behavioral therapy, interpersonal, psychodynamic, and feminist-multicultural frameworks.

Measures

Client demographics questionnaire (Standardized Demographic Set).—The client demographics questionnaire (SDS) is administered to all clients prior to their intake session, and includes information pertaining to client cultural identities (e.g. race/ethnicity, gender, sexuality, religion, nationality, and ability status). The SDS also gathers information regarding the client's family, mental health treatment, and substance use histories. The SDS is used by university counseling centers that are affiliated with the Center for Collegiate Mental Health (CCMH), and was developed from feedback from over 100 counseling centers. The SDS is revised each year to be as comprehensive and inclusive as possible.

College Center Assessment of Psychological Symptoms-34 (CCAPS-34; Locke et al., 2012).—The CCAPS-34 is the primary outcome measure in our study, and is a 34-item self-report measure assessing symptoms across a broad range of symptoms among university students. The CCAPS-34 is a subset of the larger CCAPS-62 scale, and is measured on a five point likert scale ranging from zero to four. The CCAPS-34 consists of seven subscales- depression, generalized anxiety, social anxiety, academic distress, eating concerns, hostility, and alcohol use; the CCAPS-62 subscale contains an extra subscale assessing family stress. Cut scores indicating elevated distress are 1.70 for depression, 1.70 for generalized anxiety, 2.50 for social anxiety, 2.40 for academic distress, 1.80 for eating concerns, 1.43 for hostility, and 1.40 for substance use. In addition to the subdomains, the CCAPS-34 also generates a distress index that incorporates items across subscales to assess general distress of clients; we used the distress index as our primary outcome, and the cut score for the distress index is 2.15. Higher scores on the distress index indicate greater overall distress. Sample items in the CCAPS-34 include "*My heart races for no good reason*" and "*I don't enjoy being around people as much as I used to*." The CCAPS was normed on a sample of 448,904 university students across the United States, and has strong evidence of convergent validity with other measures assessing specific distress domains, and high test-retest reliability (0.78 to 0.91; Locke et al., 2011). Internal consistency of CCAPS subscales range from 0.78 to 0.91. In our sample, the CCAPS-34 is administered to clients prior to the start of every session.

Statistical Analyses

Preprocessing and Feature Extraction

We prepared transcripts for analyses (e.g. preprocessing) by exporting each of the psychotherapy sessions into text documents that could be read by NLP models, with speaker labels for therapists indicated by "T:", and clients indicated by "P:" (e.g. "T: How were you feeling today. C: I am feeling a little depressed. T: I am sorry to hear that."). After preprocessing transcripts, we extracted text for analyses (e.g. feature extraction) using the default 'Robustly Optimized BERT Pretraining Approach' (RoBERTa) byte-pair encoding, which breaks down a sentence into words and word pairs. This algorithm strikes a balance between word-based language models and character-based language models (Sennrich et al., 2015). The vocabulary for the byte-pair encoder was 50,000. All inputs are wrapped in the standard starting and ending tags for RoBERTa ("<s>" and "</s>"). We also standardized all raw symptom scores to help stabilize error propagation in the models. Given our interest

in developing tools that could independently predict client symptoms from transcripts of therapy sessions in a context where client ratings are not available, rather than examining the causal importance of therapist-client interactions in predicting client outcomes, we did not use client symptoms as predictors in our models. Furthermore, content of therapy sessions are intertwined with client symptoms (e.g. shifts in distress could be due to content that is discussed during sessions, and session content may also be informed by current symptoms). Consequently, including client symptoms as predictors in our models could unintentionally result in models that are statistically overcontrolled, removing true statistical signal that is related to symptom ratings.

Overview of Models

We used RoBERTa (Liu et al., 2019) as our primary representation of the text. RoBERTa is a transformer neural network that can take sentences or paragraphs and output numeric vectors that can then be used for other prediction tasks. Neural networks are machine learning algorithms that are broadly modeled after human neuron connections, and are composed of nodes (e.g. computational units) separated into an input, hidden, and output layer (Jurafsky & Martin, 2014). Nodes in the input layer transmit predictor information to nodes in the hidden layers, which detects and uses patterns in the text to make predictions that are ultimately transmitted to the output layer. RoBERTa is pretrained on a corpus of 160 GB of English text to predict words hidden from the model. The pretrained model was downloaded from the Huggingface repository (https://huggingface.co/). We used the base-english pre-trained model which has 355 million parameters (hidden nodes of 1024, 16 attention heads that change weights assigned to features based on observed patterns, and 24 layers). The base model was used rather than the large model to limit the amount of computer memory used for each training example.

One of the major challenges with using transformer-based models for analyzing entire psychotherapy sessions is that computation complexity (and memory requirements) increases exponentially as the length of the inputs increase. Generally, there is a limit of 1,024 input words or word parts, while 50-minute psychotherapy sessions typically range from 10,000 to 20,000 word parts. In order to solve this issue, we broke sessions down into chunks with max lengths of 1024. We extracted the output above the classifier token (“<s>”) from each chunk. This leaves us with a variable sized matrix of (hidden size × number of chunks). We initially experimented with two different methods for simplifying this variable sized matrix into a standard-size vector: average-pooling and max-pooling. Average-pooling simplifies the matrix by calculating the average values of numeric outputs across layers, while max-pooling calculates the maximum value numeric outputs across layers. Initial piloting suggested that the max-pooling strategy provided the best performance. After max-pooling we were left with a 1024 length vector. We then added a continuous layer that reduced the vector into a single, continuous prediction.

All models were trained using the Adam optimizer (Kingma & Ba, 2014) and we used Mean Squared Error loss for the final outcome prediction. We used the validation set to tune the learning rate (e.g. parameter that controls model adjustments based on estimated errors) and tested the best model on the test set. We used pytorch as well as the Huggingface framework

to train all models. Models were trained on an Nvidia Quadro 8000 with 48GB of computer memory. We tuned the number of epochs (e.g. number of times the model reviews the training set and updates model parameters) using the validation set. We used Spearman's Rho as our main measure of model performance, as the outcome variable was not normally distributed and we were interested in how well our model ordered clients compared to their own symptom reporting. We also report R-squared as a measure of absolute performance on the test set (this version of R-squared can be negative, since we did not fit any parameters to the test set).

In addition, we used an n-gram term frequency, inverse document frequency (TF-IDF) model with a regularized linear regression model (for more information on this model see Goldberg et al., 2020). This model is a simple statistical model that predicts a continuous outcome from counts of words and phrases (called n-grams). TF-IDF examines how the presence of specific words in the text corpora contributes to relevant outcomes of interests; words in TF-IDF models are given less weight if they frequently occur across documents in the text corpora. For example, words such as "the", "it", "and" would be given less weight as they frequently occur across therapy sessions, and consequently would not help models differentiate between different sessions. For these analyses we used the open source sklearn toolkit in python.

Results

Client outcome scores at the start of treatment on the CCAPS-34 ranged from 0 to 3.45 (M=1.75; SD=0.70), and client outcome scores at termination ranged from 0 to 3.9 (M=1.29; SD=0.75). Overall client outcome scores ranged from 0 to 3.9 (M=1.69; SD=0.72). Regarding baseline severity of clients across CCAPS-34 subdomains, mean baseline scores were 1.60 for depression, 1.89 for generalized anxiety, 2.05 for social anxiety, 1.99 for academic distress, 0.79 for eating concerns, 0.66 for hostility, and 0.35 for substance use. Clients in our sample attended on average 8.06 sessions (SD=3.68), with treatment length ranging from 2 to 24 sessions. Approximately 87.1% of clients terminated treatment below the clinically significant distress index cut-off score, and 49.3% of clients terminated treatment below the low distress index cut-off score. About 24% of clients achieved reliable change at termination. Cohen's D values between the first and last observed CCAPS ratings were −0.53 for overall distress, −0.34 for depression, −0.33 for generalized anxiety, −0.22 for social anxiety, −0.23 for academic distress, −0.11 for eating concerns, −0.24 for hostility, and −0.12 for substance use.

Symptom Prediction

The average number of words in each transcript was 6,997.30 (SD =1819.12), and ranged from 19 to 15,526. Across all transcripts there were 43,234 unique word tokens. We examined learning rates ranging from 1e-5 to 1e-10, using the validation set with a linear decay of the learning rate over the course of training. Additionally, we tested between 15k and 40k learning steps. The performance on the validation set varied widely depending on the learning rate and the number of learning steps (between Rho=.38 and −.04). The best model used a learning rate of 1e-8 and trained for 9 epochs (both were tuned using the

validation set, not the test set). On the test set, the final model had a spearman's Rho of .32 ($p<.001$) and an R^2 value of 7%, meaning 7% of a client's symptom rating in the subsequent session could be predicted by the raw linguistic content of the prior session.

As a comparison, we ran an n-gram model to assess the degree to which the pre-trained RoBERTa model improved performance. We tested unigrams, bigrams and trigrams as well as regularizers (c parameters; variables that adjust model estimates based on error) ranging from .01 to 100 on the validation set. The best model was a trigram model with a regularizer coefficient of .01. This model did not predict the outcome better than chance (Spearman's Rho=.08, p=.18, $R^2 < .001$). This indicates that the transfer learning used by the RoBERTa model had a meaningful impact on the prediction of outcomes in these sessions beyond simple word frequency counts.

Discussion

Recent advances in technology have resulted in innovative methods for scaling up the evaluation of therapeutic processes (e.g., Goldberg et al., 2020) and client symptom improvement. Previous studies used NLP derived ratings of therapeutic processes (e.g. emotion words from a pre-defined dictionary in Shapira et al., 2020; interventions in Ewbank et al., 2020; topics of conversation in Atzil-Slonim et al. 2021) to predict client outcomes, but relied on older dictionary based models, did not use classical test-train splits, and did not directly predict client symptoms from session recordings. Consequently, the purpose of the current study was to build upon past NLP psychotherapy research by developing models that predict symptom improvement of a given session based on the recording of the preceding session. Using 2,630 session recordings, we were able to develop and train a model that was able to significantly predict client symptoms at an effect size similar to other psychotherapy process variables (e.g. alliance; Flückiger et al., 2018), yet do not rely on asking clients to rate sessions or therapists to guess about how a client is doing. Our results are a promising start to developing NLP tools that could quickly predict client outcome trajectories, and provide therapists with important information to tailor and improve quality of care. Furthermore, our findings highlight how session recordings contain meaningful linguistic signals that have the potential to inform, and provide contextual information for, client and therapist session process ratings. Future studies would benefit from continuing to leverage session recordings to better understand which aspects of therapist-client interactions contribute to client change. For example, researchers could analyze how shifts in client, therapist, and both client and therapist language across sessions is associated with treatment trajectories. Similarly, researchers could also examine how therapist and client language in specific content areas (e.g. exploration of identities, suicidality, behavior change) may account for symptom change.

It is important to note that our final model outperformed a comparison to an n-gram TF-IDF model, as well as past NLP models that predict specific continuous variables from session transcripts. Our comparison n-gram model did not predict session outcomes better than chance, and effect sizes from previous studies applying NLP models to session outcomes were quite small. For example, Goldberg et al. (2020) was only able to modestly predict alliance from session recordings (ρ=.15). Similarly, effect sizes for Shapira et al. (2020)

from multilevel models using NLP extracted linguistic features from transcripts to predict outcomes were small ($0.011<f^2< 0.022$; $\eta^2 = 0.08$). Our model may have outperformed past models due to the application of RoBERTa. Instead of relying on the presence of specific words (as is done in traditional n-gram models; Jurafsky & Martin, 2014), RoBERTa examines how the *relationship* between phrases is associated with changes in outcomes of interest. Given the complexities associated with therapist-client interactions, RoBERTa may be better-suited for predicting client symptoms than the previously employed simpler models. For example, a client expressing a desire to terminate therapy (e.g., "I think I should stop coming to therapy") could be positive (e.g. treatment goals achieved) or negative (e.g. premature termination). The presence of these words alone is insufficient to indicate symptom improvement. Contextual information from the rest of the session is necessary to understand the outcome trajectory of the client. NLP models that account for more context have similarly outperformed older dictionary based models in predicting sentiment in therapy contexts (Tanana et al., 2020). Future studies predicting therapeutic processes and outcomes directly from session recordings may greatly benefit from including RoBERTa models in model comparisons. Another reason our models may have performed better than Goldberg et al. (2020), is because symptom measures may assess client factors that may be more linguistically observable in session (e.g. statements regarding mood, physical symptoms) compared to a client's feelings about the relationship. Raw text may not capture self-report client measures of therapeutic processes such as alliance as fully because these measures assess internal experiences of interactions; other cues such as vocal tone and body posture may be more strongly correlated with self-report measures of therapeutic processes (Imel et al., 2014; Ramseyer & Tschacher, 2011).

Limitations and Future Directions

There were several limitations in the current study. First, our models were based on predominantly White, cisgender, heterosexual clients and therapists in a university setting; furthermore, we did not collect information on socioeconomic status. Although the sample was relatively large for a psychotherapy process evaluation, it was restricted to one treatment milieu and thus our NLP models may have been based on clinical interactions that were unique to our sample, which limits generalizability. Future research should develop models based on session recordings from different clinical settings (e.g. community mental health, hospital) across intersecting client and therapist identities. Given documented bias in machine learning and NLP models (Mehrabi et al., 2021; Yapo & Weiss, 2018), researchers should be aware of how clinical settings and identities of clients and therapists in their sample impact language associated with symptom improvement. Researchers would benefit from testing models in different samples (i.e., external validation), and comparing model performance across samples. Second, our models did not include other client information, such as client demographics and past mental health history, in prediction of symptom change. Consequently, information relevant to symptom improvement was excluded from our models. Future studies could incorporate more client information into models predicting treatment trajectories. Third, while our models were able to predict symptom scores from session recordings, our models did not identify what specific content in sessions impacted predictions the most. Future studies can continue to examine and implement methods of better interpreting how language shifts in therapist-client interactions drive NLP model

predictions of client outcomes. Qualitative and behavioral coding methodologies (e.g. Gonçalves et al., 2011) could also be used to examine sessions that were associated with increased, or decreased, symptom improvement. Themes, interventions, and other labels from these studies could be used to further refine NLP models, and facilitate automated identification of specific session content that predict client improvement. Fourth, contextual NLP models like RoBERTa are powerful in predicting outcomes of interest based on the word patterns in the context of specific documents, but the word patterns-outcome relationship is not necessarily generalizable outside of the document context. In other words, it is not a specific phrase in any session that influences the prediction of outcomes, but the use of specific phrases in the context of the specific session that impacts outcomes. Conversely, it is difficult to interpret how specific text features extracted from these models predict overall outcomes of interest (e.g. “why does language predict symptoms?”), as the impact of these text features varies based on specific documents (Belinkov et al., 2020). The interpretability of contextual based NLP models is an active area of research in NLP, and future studies could focus on innovative ways of applying these methodologies in psychotherapy contexts to better derive methods of interpreting model findings on client symptom improvement.

Implications

Our study has several important implications for research and clinical practice. First, our initial model results indicate that context based NLP models have great potential in identifying patterns of therapist-client interactions that contribute to client outcomes. Moreover, as NLP modeling approaches improve (e.g., shifting from unigram models to RoBERTa), their practical utility is only likely to increase. Psychotherapy researchers could implement NLP-based model predictions in studies seeking to clarify therapeutic processes at a massive scale (e.g., within the National Health Service), providing unprecedented opportunities for investigating linkages between psychotherapy processes and outcome. Large scale examinations of mechanisms of change in therapy could also allow researchers to identify patterns in language associated with more effective therapists, and in turn facilitate efforts to augment training of future therapists. Second, at clinic or health systems levels aggregate outcome predictions could be used to support quality improvement efforts (e.g., evaluating the impact of a clinic- or system-wide training initiative). Agencies could use outcome predictions to identify groups of clients at risk for treatment failure, or clinicians who consistently underperform, in order to provide additional support or training (Imel, Sheng, et al., 2015). Automated feedback could be particularly helpful for clinicians who no longer receive supervision, and whose outcomes may actually decline as they gain experience (Goldberg, Rousmaniere, et al., 2016). For example, (Goldberg, Babins-Wagner, et al., 2016) found that client outcomes gradually improved at a mental health agency over the course of seven years after an outcome monitoring system was implemented; furthermore, the effect size of improvement in client outcomes was almost three times as large as gradual decreases in outcomes over time observed in Goldberg, Rousmaniere, et al. (2016). Implementation of NLP based outcome monitoring systems could facilitate targeted feedback by quickly flagging session information that contextualizes client symptomatology, and encourage greater therapist reflection (Hirsch et al., 2018; Imel et al., 2019). This would

be particularly valuable in community settings where providers often have large caseloads, and are unable to review all of their session information due to time constraints.

Given the potential of systemic quality improvement in large clinical health settings (Boswell et al., 2015), future research could focus on developing symptom prediction based feedback systems, and evaluating feasibility and acceptability of these systems. While current feedback systems provide estimated predictions of symptoms based on past symptom scores (Goldberg, Babins-Wagner, et al., 2016; Lambert, 2015), or summaries of specific interventions (Imel et al., 2019), to date there are no feedback systems predicting symptom change directly from session content. Feasibility and acceptability studies could focus on examining ways to facilitate clinician openness to using feedback systems, particularly given the context of clinician resistance to receiving feedback, and concerns of trusting model predictions (Creed et al., 2021; Hirsch et al., 2018; Imel et al., 2019). Similarly, evaluations of feedback systems could focus on understanding how clinicians integrate knowledge of their client's predicted scores into interventions in subsequent sessions. Although much work remains to be done developing, testing, and implementing these systems in clinical settings, NLP models may hold great promise in improving the quality and efficiency of clinical care and ultimately the lives of clients.

Acknowledgments

Research reported in this publication was supported by National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award number 2 R01 AA018673. PBK was supported by the National Institute on Minority Health and Health Disparities under Award Number F31 MD014941. SBG was supported by the National Center for Complementary & Integrative Health of the National Institutes of Health under Award Number K23AT010879. DCA was supported by the National Institute on Alcohol Abuse and Alcoholism of the National Institutes of Health under award number K02 AA023814. MJT, SN, DCA, and ZEI are each co-founders and minority shareholders in Lyssn.io a technology company focused on developing machine learning tools to evaluate the quality of behavioral health conversations.

References

Aguilera A, Bruehlman-Senecal E, Demasi O, & Avila P (2017). Automated Text Messaging as an Adjunct to Cognitive Behavioral Therapy for Depression: A Clinical Trial. Journal of Medical Internet Research, 19(5), e148. [PubMed: 28483742]

Atkins DC, Steyvers M, Imel ZE, & Smyth P (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implementation Science: IS, 9, 49. [PubMed: 24758152]

Barde, & Bainwad AM (2017). An overview of topic modeling methods and tools. 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 745–750. 10.1109/ICCONS.2017.8250563

Belinkov Y, Gehrmann S, & Pavlick E (2020, July). Interpretability and analysis in neural NLP. In Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts (pp. 1–5).

Boswell JF, Kraus DR, Miller SD, & Lambert MJ (2015). Implementing routine outcome monitoring in clinical practice: benefits, challenges, and solutions. Psychotherapy Research: Journal of the Society for Psychotherapy Research, 25(1), 6–19. [PubMed: 23885809]

Creed TA, Kuo PB, Oziel R, Reich D, Thomas M, O'Connor S, Imel ZE, Hirsch T, Narayanan S, & Atkins DC (2021). Knowledge and Attitudes Toward an Artificial Intelligence-Based Fidelity Measurement in Community Cognitive Behavioral Therapy Supervision. Administration and Policy in Mental Health. 10.1007/s10488-021-01167-x

Dai Z, Wang X, Ni P, Li Y, Li G, & Bai X (2019, October). Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Suzhou, China. 10.1109/cisp-bmei48845.2019.8965823

Devlin J, Chang M-W, Lee K, & Toutanova K (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North. Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota. 10.18653/v1/n19-1423

Elliott R, Bohart AC, Watson JC, & Greenberg LS (2011). Empathy. Psychotherapy , 48(1), 43–49. [PubMed: 21401273]

Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, & Blackwell AD (2020). Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. JAMA Psychiatry , 77(1), 35–43. [PubMed: 31436785]

Flemotomos N, Martinez VR, Chen Z, Singla K, Ardulov V, Peri R, … & Narayanan S (2022). Automated evaluation of psychotherapy skills using speech and language technologies. Behavior Research Methods, 54(2), 690–711. [PubMed: 34346043]

Flückiger C, Del Re AC, Wampold BE, & Horvath AO (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. Psychotherapy , 55(4), 316–340. [PubMed: 29792475]

Fortney JC, Unützer J, Wrenn G, Pyne JM, Smith GR, Schoenbaum M, & Harbin HT (2017). A Tipping Point for Measurement-Based Care. Psychiatric Services , 68(2), 179–188. [PubMed: 27582237]

Goldberg SB, Babins-Wagner R, Rousmaniere T, Berzins S, Hoyt WT, Whipple JL, Miller SD, & Wampold BE (2016). Creating a climate for therapist improvement: A case study of an agency focused on outcomes and deliberate practice. Psychotherapy , 53(3), 367–375. [PubMed: 27631868]

Goldberg SB, Buck B, Raphaely S, & Fortney JC (2018). Measuring Psychiatric Symptoms Remotely: a Systematic Review of Remote Measurement-Based Care. Current Psychiatry Reports, 20(10), 81. [PubMed: 30155749]

Goldberg SB, Flemotomos N, Martinez VR, Tanana MJ, Kuo PB, Pace BT, Villatte JL, Georgiou PG, Van Epps J, Imel ZE, Narayanan SS, & Atkins DC (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. Journal of Counseling Psychology, 67(4), 438–448. [PubMed: 32614225]

Goldberg SB, Rousmaniere T, Miller SD, Whipple J, Nielsen SL, Hoyt WT, & Wampold BE (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. Journal of Counseling Psychology, 63(1), 1–11. [PubMed: 26751152]

Gonçalves MM, Ribeiro AP, Mendes I, Matos M, & Santos A (2011). Tracking novelties in psychotherapy process research: The innovative moments coding system. Psychotherapy Research: Journal of the Society for Psychotherapy Research, 21(5), 497–509. [PubMed: 21480054]

Hirschberg J, & Manning CD (2015). Advances in natural language processing. Science, 349(6245), 261–266. [PubMed: 26185244]

Hirsch T, Soma C, Merced K, Kuo P, Dembe A, Caperton DD, Atkins DC, & Imel ZE (2018). “It’s hard to argue with a computer:” Investigating Psychotherapists’ Attitudes towards Automated Evaluation. DIS. Designing Interactive Systems (Conference), 2018, 559–571. [PubMed: 30027158]

Imel ZE, Barco JS, Brown HJ, Baucom BR, Baer JS, Kircher JC, & Atkins DC (2014). The association of therapist empathy and synchrony in vocally encoded arousal. Journal of Counseling Psychology, 61(1), 146–153. [PubMed: 24274679]

Imel ZE, Pace BT, Soma CS, Tanana M, Hirsch T, Gibson J, Georgiou P, Narayanan S, & Atkins DC (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. Psychotherapy , 56(2), 318–328. [PubMed: 30958018]

Imel ZE, Sheng E, Baldwin SA, & Atkins DC (2015). Removing very low-performing therapists: A simulation of performance-based retention in psychotherapy. Psychotherapy , 52(3), 329–336. [PubMed: 26301424]

Imel ZE, Steyvers M, & Atkins DC (2015). Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. Psychotherapy , 52(1), 19–30. [PubMed: 24866972]

Jurafsky D, & Martin JH (2014). Speech and Language Processing. Pearson.

Kao J, & Jurafsky D (2012). A computational analysis of style, affect, and imagery in contemporary poetry. Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature, 8–17.

Kingma DP, & Ba J (2014). Adam: A Method for Stochastic Optimization. https://www.semanticscholar.org/paper/a6cb366736791bcccc5c8639de5a8f9636bf87e8

Konstantinov A, Moshkin V, & Yarushkina N (2021). Approach to the use of language models BERT and Word2vec in sentiment analysis of social network texts. In Recent Research in Control Engineering and Decision Making (pp. 462–473). Springer International Publishing.

Lambert MJ (2015). Progress feedback and the OQ-system: The past and the future. Psychotherapy , 52(4), 381–390. [PubMed: 26641368]

Landesberger J, Ehrlich U, & Minker W (2020). "What is it?" How to Collect Urgent Utterances using a Gamification Approach. 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 19–22.

Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, Kassab H, Hoffman M, Scott K, Lyon A, Douglas S, Simon G, & Kroenke K (2019). Implementing Measurement-Based Care in Behavioral Health: A Review. JAMA Psychiatry , 76(3), 324–335. [PubMed: 30566197]

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, & Stoyanov V (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. In https://openreview.net › forum https://openreview.net › forum. https://openreview.net/pdf?id=SyxS0T4tvS

Locke BD, McAleavey AA, Zhao Y, Lei P-W, Hayes JA, Castonguay LG, Li H, Tate R, & Lin Y-C (2012). Development and Initial Validation of the Counseling Center Assessment of Psychological Symptoms–34. Measurement and Evaluation in Counseling and Development, 45(3), 151–169.

Marcelle ET, Nolting L, Hinshaw SP, & Aguilera A (2019). Effectiveness of a Multimodal Digital Psychotherapy Platform for Adult Depression: A Naturalistic Feasibility Study. JMIR mHealth and uHealth, 7(1), e10948. [PubMed: 30674448]

Mehrabi N, Morstatter F, Saxena N, Lerman K, & Galstyan A (2021). A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv, 54(6), 1–35.

Moyers T, Martin T, Catley D, Harris KJ, & Ahluwalia JS (2003). Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. Behavioural and Cognitive Psychotherapy, 31(2), 177–184.

Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlícek P, Qian Y, Schwarz P, Silovský J, Stemmer G, & Veselý K (2011). The Kaldi Speech Recognition Toolkit. https://www.semanticscholar.org/paper/3a1a2cff2b70fb84a7ca7d97f8adcc5855851795

Ramseyer F, & Tschacher W (2011). Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. Journal of Consulting and Clinical Psychology, 79(3), 284–295. [PubMed: 21639608]

Scott K, & Lewis CC (2015). Using Measurement-Based Care to Enhance Any Treatment. Cognitive and Behavioral Practice, 22(1), 49–59. [PubMed: 27330267]

Sennrich R, Haddow B, & Birch A (2015). Neural Machine Translation of Rare Words with Subword Units. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1508.07909

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), 77–87. [PubMed: 32352823]

Shimokawa K, Lambert MJ, & Smart DW (2010). Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. Journal of Consulting and Clinical Psychology, 78(3), 298–311. [PubMed: 20515206]

Syzdek BM (2020). Client and therapist psychotherapy sentiment interaction throughout therapy. Psychological Studies. 10.1007/s12646-020-00567-7

Tanana M, Dembe A, Soma CS, Imel Z, Atkins D, & Srikumar V (2016). Is Sentiment in Movies the Same as Sentiment in Psychotherapy? Comparisons Using a New Psychotherapy

Sentiment Database. Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology, 33–41.

Tanana MJ, Soma CS, Kuo PB, Bertagnolli NM, Dembe A, Pace BT, Srikumar V, Atkins DC, & Imel ZE (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. Behavior Research Methods, 53(5), 2069–2082. [PubMed: 33754322]

Tao KW, Owen J, Pace BT, & Imel ZE (2015). A meta-analysis of multicultural competencies and psychotherapy process and outcome. Journal of Counseling Psychology, 62(3), 337–350. [PubMed: 26167650]

Tracey TJG, Wampold BE, Lichtenberg JW, & Goodyear RK (2014). Expertise in psychotherapy: an elusive goal? The American Psychologist, 69(3), 218–229. [PubMed: 24393136]

Webb CA, Derubeis RJ, & Barber JP (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. Journal of Consulting and Clinical Psychology, 78(2), 200–211. [PubMed: 20350031]

Xiao B, Can D, Georgiou PG, Atkins D, & Narayanan SS (2012). Analyzing the Language of Therapist Empathy in Motivational Interview based Psychotherapy. Signal and Information Processing Association Annual Summit and Conference (APSIPA), … Asia-Pacific. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2012. https://www.ncbi.nlm.nih.gov/pubmed/27602411

Xiao B, Imel ZE, Georgiou PG, Atkins DC, & Narayanan SS (2015). "Rate My Therapist": Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. PloS One, 10(12), e0143055. [PubMed: 26630392]

Yapo A, & Weiss J (2018). Ethical implications of bias in machine learning. Proceedings of the 51st Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences. 10.24251/hicss.2018.668

Table 1

Presenting concerns of clients

Presenting concern	% of clients endorsed
Anxiety	69.7
Depression	63.3
Academic performance	43.1
Self esteem	41.0
Loneliness	36.4
Social anxiety	33.2
Relationship concerns with partner	26.9
Family of origin	21.8
Relationship concerns with friends	19.9
Body image	17.1
Existential identity concerns	17.0
Shyness	15.1
Thoughts of suicide	14.5
Test performance anxiety	14.5
Career issues	13.1
Grief	11.6
Trauma	9.6
Financial concerns	9.4
Religious identity exploration	6.8
Work	6.8
Anger	6.3
Obsessive compulsive disorder	5.1
Self injury	5.1
Pornography	4.8
Bipolar disorder	4.8
Sexual health concerns	4.6
Sexual orientation identity exploration	4.5
Sexual assault	3.7
Substance use	3.2
Being a parent	3.1
Learning disability	2.8
Eating disorder	2.2
Hallucinations	1.7
Suicide attempt	1.1
Discrimination	0.8
Divorce separation	0.8
Racism	0.6
Legal concerns	0.6
Homelessness	0.3