Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy

Article *in* Psychotherapy Theory Research & Practice · March 2021

DOI: 10.1037/pst0000362

7 authors, including:

Dana Atzil Slonim
Bar Ilan University
52 PUBLICATIONS **391** CITATIONS
SEE PROFILE

Daniel Juravski
Bar Ilan University
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Eran Bar-Kalifa
Ben-Gurion University of the Negev
66 PUBLICATIONS **507** CITATIONS
SEE PROFILE

Eva Gilboa-Schechtman
Bar Ilan University
133 PUBLICATIONS **3,833** CITATIONS
SEE PROFILE

Using Topic Models to Identify Clients' Functioning Levels and Alliance Ruptures in Psychotherapy

Accepted for publication in Psychotherapy

Atzil-Slonim, D[1]., Juravski, D[2]., Bar-Kalifa, E[3]., Gilboa-Schechtman, E[1]., Tuval-Mashiach, R[1]., Shapira, N[2]., & Goldberg, Y[2].

[1] Department of Psychology, Bar-Ilan University, Ramat-Gan, Israel

[2] Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel

[3] Department of psychology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Note: the first two authors equally contributed to this article.

Please address correspondence to: Dana Atzil-Slonim, Bar-Ilan University, Psychology Department, Ramat-Gan, Israel; Phone number: 972-54-6865888; dana.slonim@gmail.com

Abstract

Computerized Natural Language Processing techniques can analyze psychotherapy sessions as texts; thus generating information about the therapy process and outcome and supporting the scaling-up of psychotherapy research. We used topic modeling to identify topics discussed in psychotherapy sessions and explored (1) which topics best identified clients' functioning and alliance ruptures and (2) whether changes in these topics were associated with changes in outcome. Transcripts of 873 sessions from 58 clients treated by 52 therapists were analyzed. Prior to each session, clients self-reported functioning and symptom level. After each session, therapists reported the extent of alliance rupture. Latent Dirichlet Allocation was used to extract latent topics from psychotherapy textual data. Then a Sparse Multinomial Logistic Regression model was used to predict which topics best identified clients' functioning levels and the occurrence of alliance ruptures in psychotherapy sessions. Finally, we used multi-level growth models to explore the associations between changes in topics and changes in outcome. Session-based processing yielded a list of semantic topics. The model identified the labels above chance (65%-75% accuracy). Change trajectories in topics were associated with change trajectories in outcome. The results suggest that topic models can exploit rich linguistic data within sessions to identify psychotherapy process and outcomes.

Keywords: machine learning, natural language processing, text, topic models, psychotherapy process and outcome

Clinical Impact Statement

Question: Can machine learning techniques identify the topics discussed in psychotherapy sessions and examine the associations between these topics and treatment process and outcome?

Findings: Topic modeling yielded semantically meaningful topics that were then used to identify which topics were most closely associated with the level of clients' functioning and those that were most closely associated with rupture occurrence. Changes in these topics were associated with changes in outcome.

Meaning: Topic modeling can enable therapists to be better attuned to specific topics that may signal important events in therapy. Using topic model output, therapists can access a summary of topics discussed in a session, locate specific themes associated with rupture or with clients' deterioration, and orient interventions to improve the situation.

Next steps: Future studies could use topic model output alongside existing monitoring tools to inform therapists of meaningful linguistic processes that occur within psychotherapy sessions.

Using Topic Models to Identify Clients' Functioning Levels and Alliance Ruptures in Psychotherapy

Psychotherapy is based to a great extent on the content of exchanges between clients and therapists, which conveys important information about the participants' modes of communication, mental states, thoughts, and feelings. Until recently, most psychotherapy research has relied on self-report measures or on human coders to quantify the information in psychotherapy sessions. These standardized subjective measures are the building blocks of psychotherapy research, and the process and outcome of treatment cannot be studied without them. However, these methods also have critical shortcomings, including the extent of participants' self-insights, their willingness to complete questionnaires, and their restricted choice of responses (for a review of the limitations of current research methods, see Kazdin, 2016). Furthermore, observational human coding is very labor-intensive, which limits the amount of data that can be analyzed and thus curtails the generalizability of results (Hill & Lambert, 2004). Psychotherapy research could however be enriched by adding more objective and flexible methods that can process copious data and tap the rich information in the human dialogue that occurs in psychotherapy sessions. The ability of computerized Natural Language Processing (NLP) techniques to analyze text from psychotherapy sessions to generate useful information about the therapy process and outcomes may be a way to scale up psychotherapy research (Imel, Steyvers, & Atkins, 2015; Salvatore et al., 2017).

NLP is a subfield of computer science and machine learning that aims to analyze, understand, and produce human language content (Hirschberg & Manning, 2015). There is growing interest in NLP in healthcare research (Peek, Combi, Marin, & Bellazzi, 2015), but few

studies have used NLP to analyze text within psychotherapy sessions (see however exceptions including Atkins, Steyvers, Imel, & Smyth, 2014, and Mergenthaler, 1996; 2008, among others).

There are many NLP methods with possible applications to mental health and psychotherapy research (for a review see Calvo, Milne, Hussain, & Christensen, 2017). When working with textual data, an inherent question is how to use traditional statistical tools, since treating each individual word as a feature (predictor) quickly leads to very sparse and high-dimensional data which typical statistical methods are ill-equipped to handle. To reduce data dimensionality, researchers use different methods to group words together. Most previous studies using computerized methods of text analysis in psychotherapy have employed "dictionaries" of words classified into categories (e.g., Bucci, Kabasakalian, & The RA Research Group, 1992; Mergenthaler, 1996, 2008). These methods take a top-down approach that involves grouping text units based on predefined categories. Other methods take a data-driven/bottom-up approach without any other input. The main strength of these data-driven approaches is that they can identify semantic patterns that may emerge from the text and discover processes that underlie psychotherapy gains (for further discussion of text analysis methods in psychotherapy see Gelo, Salcuni & Colli, 2012).

One promising class of methods is known as topic modeling (Blei, Ng, & Jordan, 2003; Steyvers & Griffiths, 2006), that includes Latent Dirichlet Allocation (LDA; for a tutorial on how to use LDA with psychotherapy data, see Atkins et al., 2012). LDA discovers latent structure in a text and forms meaningful word groups from the text inductively, requiring no external input other than the segmentation of the text into meaningful units (e.g., client/therapist speech turns). One of the main advantages of the LDA method is that it reduces data dimensionality considerably by grouping words into topics and associates each document with a

set of topics that characterize it. A “document” is a cohesive unit of text; for example, all words spoken during a session, only words spoken by the client, or (as in our case) words spoken in a single dialogue turn. A “topic” is a distribution over words in the document, which groups those words that the model has found to be related to each other, and indicates the centrality of each word to the topic. The algorithm infers topics and their association with the document on the assumption that (a) there is a fixed set of topics that covers all the documents analyzed; (b) each document is composed of a limited subset of the topics; and (c) each topic is characterized by a limited number of words. Based on these constraints, the model assigns words to topics and topics to documents to best “explain” the data, in a probabilistic sense. In this process the model infers (“discovers”) latent topics and trends in the data. Inferred topics are not named, but they can often be assigned meaning by looking at the most dominant words in each. The inputs to the LDA algorithm are the predetermined number of topics and the set of documents. The outcome of the algorithm is a set of topics, many of which are interpretable, and the associations of these topics to documents (see Blei et al., 2003). Thus, in addition to reducing the dimensionality of the text, the main advantage of LDA is that the topics are typically highly interpretable, which makes them valuable in clinical practice as well as research where an understanding of the underlying dimensions of the linguistic data is important. Atkins et al., (2012) highlighted two key potential uses of topic modeling for psychotherapy researchers and clinicians. Psychotherapy researchers can use the extracted topics to predict variables related to the psychotherapy process and outcomes. Associating topics with coherent semantic concept labels means that higher-level concepts, rather than words, can be treated as the main feature of analysis. This allows for both qualitative and quantitative analysis of the data, which can be harnessed to determine which topics are most strongly associated with positive and negative processes and outcomes in

psychotherapy. By considering each document as a set of topics rather than a set of words, statistical tools can deal with a manageable number of predictors.

In addition, the resulting topics can serve as a useful summarization of the semantic themes in the linguistic therapy session data. The researcher can achieve a broader picture of the data, such as what themes were discussed in a session or tended to dominate in specific clients across sessions, what changes emerged in specific topics for specific clients or over an entire sample, or whether changes in specific themes were associated with changes in outcomes. This information can be made available to the clinician through a specialized user interface that allows for a conceptual exploration of the therapy process.

Recently, Gaut, Steyvers, Imel, Atkins, and Smyth (2015) showed that topic models can identify specific talk turns representative of symptom codes. Atkins et al. (2012) demonstrated the utility of topic modeling to predict productive processes in couples psychotherapy. Imel et al. (2015) reported that topic models with large psychotherapy transcript data could identify clinically relevant content, including symptoms and relation-related topics. They suggested that to expand what is known about what actually happens during psychotherapy sessions and what specific content is associated with better outcomes, more studies using computational methods such as topic modeling to predict central processes and outcomes should be conducted. Using LDA-derived topics to identify clients' functioning levels and the occurrence of alliance ruptures presents numerous advantages, since these are two of the most central and frequently investigated variables in psychotherapy research.

Client functioning is often used to assess session outcomes and provide feedback to therapists (e.g., Hatfield & Ogles, 2006; Shimokawa, Lambert, & Smart, 2010), usually based on clients' self-reports that are assessed before or after the session. Unfortunately, therapists are

often limited in their ability to track changes in clients' functioning and also tend to generally over or-underestimate them (e.g., Bar-Kalifa et al., 2016). Computerized linguistic analyses may be able to capture subtle, albeit diagnostically significant functioning markers. These analyses can serve as an implicit way of studying clients' functioning and may be used as a complement to standard monitoring systems to evaluate clients' progress (Pace et al., 2016). Clients' level of functioning may affect or be affected by session content; using topic models to extract topics that identify clients' functioning may reveal which content does so, and how.

The vast clinical literature on the therapeutic alliance and the robust association frequently found between alliance and treatment outcome underscore the importance of the therapeutic relationship in the psychotherapy process (Flückiger, Del Re, Wampold, and Horvath, 2018). Alliance ruptures are relational events in the interpersonal space between therapists and clients that correspond to tensions or breakdowns in their collaborative relationship (Safran & Muran, 2006). Alliance ruptures can be meaningful therapeutic events (Bordin, 1979), providing therapists and clients with new access to the client's inner world in the here-and-now of therapy. This access allows implementation of corrective experiences, (hopefully) leading to repair and relief of symptoms (Eubanks-Carter, Gorman, & Muran, 2012).

Alliance ruptures have received substantial empirical attention recently (for a review, see Eubanks, Muran, & Safran, 2018). Most studies have explored ruptures at relatively low time resolution (once each session, typically weekly) using self-reports. However, ruptures may occur at higher time resolutions within a session (Coutinho, Ribeiro, Sousa, & Safran, 2014). Recent studies have used within-session coding tools to detect ruptures moment-by-moment during a session, yielding important insights into the within-session processes that lead to ruptures (e.g., Eubanks-Carter, Muran, & Safran, 2009). These insights have been used to train therapists to

recognize ruptures when they happen (Eubanks, Muran, & Safran, 2015). However, these coding systems are time-consuming and labor-intensive. Topic modeling may be able to recognize ruptures and capture subtler and more implicit content associated with their occurrence. In addition, discovery-oriented text-mining procedures such as topic modeling to explore ruptures are likely to contribute to a better understanding of the actual content that leads to these events.

In this study we aimed to assess whether topic models could identify clients' functioning levels and alliance ruptures. In addition, we aimed to examine whether and to what extent the topics identified would change over the course of treatment and whether this change was associated with treatment outcome. Finally, we explored the potential value of topic modeling to clinicians by analyzing the patterns of change in topics over the course of one good and one poor outcome cases. Specifically, two hypotheses were formulated:

Hypothesis 1. Topic models will identify clients' functioning and rupture occurrence above chance.

Hypothesis 2. Changes in topics over the course of treatment will be associated with changes in outcome over the course of treatment (2a). The topics will demonstrate a different pattern of change over the course of a case of a good vs. a case of a poor outcome (Exploratory hypothesis 2b).

Method

Clients

The analyses were based on 873 sessions from 58 clients in individual psychotherapy at a large university outpatient clinic located in a central city in Israel. Data were collected naturalistically between August 2014 and August 2016 as part of the clinic's regular practice of monitoring clients' progress. The language of therapy was Modern Hebrew (MH). The clients

were all above age 18 (Mage= 39.06, SD = 13.67, range 20-77), and most were women (58.9%). Of the clients, 53.5% had at least a bachelor's degree; 53.5% were single, 8.9% were in a committed relationship but unmarried; 23.2% were married, and 14.2% were divorced or widowed. Diagnoses were based on the Mini International Neuropsychiatric Diagnostic Interview for Axis I DSM-IV diagnoses (MINI 5.0; Sheehan et al., 1998). The interviews were conducted before the actual therapy began, by well-trained independent clinicians. All intake sessions were audiotaped, and a random 25% of the interviews were sampled and rated again by an independent clinician. The mean Kappa value for the Axis I diagnoses was excellent (k = 0.9).

Of the clients, 22.9% had one diagnosis, 20.0% had two, and 25.7% had three or more. The most common diagnoses were comorbid anxiety and affective disorders (25.7%), followed by other comorbid disorders (17.1%), anxiety disorders (14.3%), and affective disorders (5.7%). Several clients (31.4%) reported relationship concerns, academic/occupational stress, or other problems that did not meet criteria for any Axis I diagnosis.

Therapists and Therapy

Clients were treated by 52 therapists. All were MA or Ph.D. students at different stages of clinical psychology training (1-5 years of experience). Twenty therapists were first-year graduate students, and had no previous clinical experience. The remainder had a range of 50-250 previous clinical hours. Clients were assigned to therapists in an ecologically valid manner reflecting therapist availability and caseload. Most therapists treated one client each but some (6) treated two. Each therapist received 1 hour of individual supervision and 4 hours of group supervision on a weekly basis. All therapy sessions were audiotaped for use in supervision with senior clinicians. The individual and group supervision focused heavily on the review of the audiotaped case material and appropriate therapist interventions.

Individual psychotherapy consisted of once-weekly sessions of primarily psychodynamic psychotherapy organized, aided, and informed (but not prescribed) by a short-term psychodynamic psychotherapy treatment model (e.g., Shedler, 2010; Summers & Barber, 2009). The key features of this model are (a) focus on affect and experience and expression of emotions, (b) exploration of attempts to avoid distressing thoughts/feelings, (c) identification of recurring themes and patterns, (d) emphasis on past experiences, (e) focus on interpersonal experiences, (f) emphasis on the therapeutic relationship, and (g) exploration of wishes, dreams, or fantasies (Shedler, 2010). Treatment was open-ended in length, but was often restricted to 9 months to 1 year, reflecting the trainee clinicians' program and the university calendar.

Instruments and Data Collection

Outcome Rating Scale (ORS; Miller, Duncan, Brown, Sparks, & Claud, 2003). The ORS is a 4-item visual analog scale developed as a brief alternative to the OQ-45. It assesses change in three areas of client functioning that are widely considered to be valid indicators of progress in treatment: functioning, interpersonal relationships, and social role performance. Respondents complete the ORS before each therapy session by rating four statements on a visual analog scale anchored at its respective ends by the words Low and High. This scale yields four separate scores between 0 and 10 (for a total score of 0–40), with higher scores indicating better functioning. According to the ORS manual, a score of 24 represents the cutoff for clinical status. The Reliable Change Index (RCI) for the ORS is 5; thus, cases with a gain score of 5 and above are classified as improved. The ORS has strong reliability estimates (α =.87–0.96) and moderate correlations between the ORS items and the OQ-45 subscale and total scores (ORS total - OQ-45 total: r = .59). The reliability levels in the current study were considered excellent (within α = 0.88, between α = 0.95). The ORS has been translated into many languages, including Hebrew

(Hafkenscheid, Duncan, & Miller, 2010). In our sample the ORS mean score was 24.397 (SD = 7.95).

PSQ (Post-Session Questionnaire) (PSQ: Muran, Samstag, Safran, & Wilson, 2004). Alliance ruptures were assessed after each session with one question to the therapist: "Did you experience any tension, misunderstanding, conflict or disagreement in the relationship with your patient?". This item is answered subjectively on a 5-point Likert-type scale from 1 ("not at all") to 5 ("constantly"). Following Muran et al. (2009), a rupture was defined as any rating higher than 1 on the scale. The PSQ has been widely used in psychotherapy research and demonstrates sound psychometric properties, including predictive validity with a variety of process indices (Muran et al., 2009), such as the Working Alliance Inventory (WAI, Tracey & Kokotovic, 1989). The PSQ has been translated into many languages, including Hebrew (Wiseman & Tishby, 2014). In our sample the PSQ mean score was 2.06 (SD = 1.43).

The Hopkins Symptom Checklist-Short Form (HSCL-11; Lutz, Tholen, Schürch, & Berking, 2006). This 11-item self-report inventory assesses symptomatic distress, and is a brief version of the SCL-90-R (Derogatis, 1977). The items are rated on a 4-point Likert scale ranging from 1 (not at all) to 4 (extremely). The mean of the 11 items represents the client's level of global symptomatic distress during the preceding week. The score was found to be highly correlated with the SCL-90-R's global severity index (r = 0.91) and has high internal consistency (α = .92; Lutz et al., 2006). In the current study, the between- and within-person reliabilities for the scale were high (within α = .83, between α = .92). In our sample the HSCL mean was 1.74 (SD = 0.59).

Transcription. To capture the treatment processes from session to session, and since the transcription process is highly expensive, transcriptions were conducted alternately (i.e., sessions 2, 4, 6, 8, etc.). Where the material was incomplete (e.g., due to low recording quality or failure

to complete questionnaires for a specific session), the next session was transcribed instead. The transcriber team was composed of seven transcribers, all of whom were graduate students in the University's Psychology Department. The transcribers went through a one-day training workshop and monthly meetings were held throughout the transcription process to supervise the quality of their work. Their training included specific guidelines on how to handle confidential and sensitive information and the transcribers were instructed to replace names by pseudonyms and to mask any other identifying information. The transcription protocol followed general guidelines as described in Mergenthaler and Stinson (1992) as well as in Albert, MacWhinney, Nir, and Wintner, (2013). The audiotape was transcribed in its entirety and provided a verbatim account of the session.

There were 873 transcripts in total (the mean transcribed sessions per client was 11.79; SD = 3.08). Each transcript incorporated metadata such as the client's code, which allowed the client data to be linked across sessions and for hierarchical analysis. The transcriptions totaled about five million words, and over 240,000 talk turns (i.e., switching between speakers). On average, there were 5842 words in a session, of which 4525 (77%; SD=1407.07; range 416-8176) were client utterances and 1317 (23%; SD=728.12; range 160-6048) were therapist utterances with a mean of 278.08 (SD=140.59; range 47-972) talk turns per session.

Procedure. The procedures were part of the routine assessment and monitoring process in the clinic. 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. The clients completed the ORS and the HSCL before each session and the therapists completed the PSQ at the end of each session.

Selection of good and poor outcome cases. To illustrate changes in topics over the course of treatment (Exploratory Hypothesis 2b), we selected one case of a good and one case of a poor outcome. Specifically, from the initial sample of 58 clients, we excluded 33 who did not meet the ORS clinical cutoff at the start of treatment (i.e., the average of their first 3 sessions' ORS scores exceeded 24). We then computed the average of each client's last three reports on the ORS to determine whether they showed a reliable change during treatment (an increase of at least 5 points on the ORS from beginning to end of treatment; Miller et al., 2003).

Using this procedure, 65% of clients were identified as responders and the rest as non-responders. From the responders group, we randomly selected one client, Noya (pseudonym), who was diagnosed with dysthymia at the beginning of treatment. Her average ORS score for the first three sessions was 21 and for the last three sessions was 28, indicating a reliable change and a shift to the non-clinical range by the end of treatment. To match Noya, we chose another female client from the nonresponder group who was diagnosed with dysthymia at the beginning of treatment. Gali's average ORS score for the first three sessions was 18, and did not change in the last three sessions. We thus treated these clients as coming from two qualitatively distinct groups, where we differentiated client groups as a function of treatment response (as per, e.g., Montesano, Gonçalves, & Feixas, 2017).

Data Analysis and Results

In this section, we describe each step of our analytic strategy and the key results. We start with the preliminary stage of preprocessing that is needed before the main text analysis; then we describe the extraction of topics from the transcription data using LDA . Next, we describe how

the transcription topic distributions were used as input features to identify session-level clients' functioning and alliance ruptures (Hypothesis 1). Figure 1 provides a visualization of the flow of these stages, which are described in detail below.

Finally, we tested whether change trajectories in topics were related to change trajectories in clients' symptoms (Hypothesis 2a). We then discuss two cases to illustrate how topics changed throughout one good and one poor outcome (Exploratory Hypothesis 2b).

Preprocessing

Text analysis requires some standard preprocessing steps to make data machine readable and easy to use. Thus, we applied the following preprocessing steps.

Cleaning. We were only interested in content words uttered during the session; thus, other information (punctuation, linguistic comments by the transcribers, interpretations, etc.) was omitted.

Lemmatization. We extracted base forms for all the words ("lemmas"; e.g., לחכות "wait" (infinitive), חכה "wait" (imperative) מחכה "waits" (present, 3rd person, singular), מחכים "wait" (present, other persons and/or plural), חיכה "waited" (simple past), etc. were all lemmatized to מחכה "wait").

POS annotation. Syntax and structure go hand in hand: specific rules, conventions, and principles usually govern how words are combined. Parts of speech (POS) are lexical categories to which words are assigned based on syntactic context and role (e.g. "noun," "verb," "adjective," "adverb," "pronoun," "prepositions," etc.). We associated each word with its part of speech, and retained only words in specific categories (see 1.3 Word Filtering below).

Hebrew language. Our sessions are in Hebrew, a language in which extracting lemmas and parts of speech is more challenging than in English, due to the high morphological and

syntactic ambiguity (Adler, 2007; Goldberg, 2011). We used the specialized YAP tool (More & Tsarfaty, 2016; https://github.com/OnlpLab/yap) to handle these challenges. YAP provides full morphological disambiguation with relatively high accuracy, from which we extracted the lemma and part-of-speech information. There are many other parsing tools that support other languages and can be implemented with programming languages such as Python, R, and Java. For example: CoreNLP (https://stanfordnlp.github.io/CoreNLP/), spaCy (https://spacy.io/), NLTK (https://www.nltk.org/).

Topic Modeling

Background on LDA. Here, a brief description of LDA and topic modeling is provided (technical details are in Blei et al., 2003). The basic assumption of LDA is that a document (as defined below) represents a mixture of *k* topics (where *k* is a user-predefined hyperparametric number) in different proportions (Blei et al., 2003). Using Bayesian probabilistic modeling, LDA finds clusters of terms (i.e., topics) that tend to co-occur in subsets of the documents. Thus, topics are defined as a distribution over a fixed vocabulary of *n* words for a given corpus (Blei, 2012). LDA defines a generative process in which two kinds of probabilities are drawn from Dirichlet distributions (i.e., a distribution over multinomial distributions): (1) distributions of words given topics, whose skewness is governed by a Dirichlet prior (β); and (2) distributions of topics given documents, whose skewness is governed by a Dirichlet prior (α). "Prior" means that the α and β hyperparameters must be set prior to analysis. Lower values of α result in documents containing fewer topics, and lower values of β result in more separate topics.

The learning process of the model concludes with 3 outputs: (1) a list of *k* topics, where each topic represents ideally a cluster of same-themed vocabulary items (this output is described in Figure 1, stage 4a, and the output of the results appears in Table 1; note that words can belong

to multiple topics). Every topic is a sum-to-1 distribution vector over the entire vocabulary of *n* words. (2) Assignment of each of n_d words in each document to a topic (Figure 1, stage 4b); (3) a list of *d* vectors, one for each document, where each vector is a sum-to-1 distribution vector over all *k* topics, indicating the topic proportion in the document (Figure 1, stage 4c).

Document definition. To apply topic modeling, a "document" in the data must be defined, with several considerations in mind depending on the dataset. A maximum number of documents is preferable to yield semantically cohesive topics; however, document length should not be too short or too long. The basic assumption is that every document is semantically coherent and consists of a small number of topics. Based on these criteria, and in line with previous studies (e.g., Atkins, 2012), we defined a document to correspond to a client talk turn. Documents (turns) with fewer than 50 or more than 1000 words were omitted to stabilize document length to a mean of 138 words (SD=103.91), which yielded 28,323 documents.

Word filtering. We omitted semantically empty words, rare words and uninformative words. Words were omitted from each document if they corresponded to one or more of the following 3 conditions: (1) The word is a function word, defined as any word whose POS is not a noun, verb, adjective, or adverb, and hence it does not carry topical semantic information (e.g., "the," "and," "yes"). (2) The word is a rare word, appearing fewer than 10 times over the whole documents. The high frequency of rare words (the "long tail" effect) can undermine the model's regularization and lead to poor semantic-interpretable topic outcomes. (3) The word is common, appearing in more than 90% of the documents. High-frequency words are subsumed by most topics and thus detract from the richness of the meaning of the topics.

Determining the number of topics. The model should be provided with a number k of topics as a hyperparameter (these are set heuristically, for a discussion on the number of topics,

see Blei et al., 2003 and Griffiths & Steyvers, 2004). In this study we tested the model with a set of common values of k = {50, 100, 200, 300} and manually examined the resulting topic distributions. Consistent with previous studies (e.g., Atkins et al., 2012), after running the model with each of these values, each of the resulting set of topics was visually inspected and the set of topics that was the most interpretable was selected. If most of the original documents have a very small number of topics, this implies that the number of topics needs to be increased; on the other hand, if a large number of words appear in multiple topics, this indicates that there has been an over-allocation of k. By applying this procedure, we settled on k=200 topics, since this set of topics was the most interpretable. When choosing too few topics, the model cannot deal with noise, and may be forced to combine several topics into a single one. Using more topics allows the model to better deal with noise, by assigning more weight to meaningful topics, while delegating the "noise" to idiosyncratic topics (i.e., certain topics may reflect the idiosyncratic content of specific individuals in the corpus). An overly high number of topics may result in many idiosyncratic topics (in our case, this was observed with k=300).

Topic modeling implementation. The documents were analyzed with an unsupervised probabilistic topic model using the Machine Learning for Language Toolkit (MALLET; McCallum, 2002), based on standard LDA (Blei et al., 2003). We used the default setting of 1000 Gibbs sampling iterations. The number of iterations between re-estimating Dirichlet hyperparameters was set to 20, and the number of iterations before first estimating Dirichlet hyperparameters to 50. Automatic hyperparameter optimization (for α and β) was enabled to allow prominence of topics and skewness of associated word distributions to vary to best fit the data. The model resulted in a table of 200 topics, where each topic is a group of related words sampled from the documents' corpus dictionary. Table 1 presents 35 selected topics (out of a

total of 200). A good outcome here, which is indicative of the model's successful learning, is a meaningful topic division. Table 1 makes it clear that the high probability words in each topic capture semantically related content which reflect aspects of the therapeutic encounter that we might expect clients and therapists to discuss. Based on Atkins et al., (2012) and Imel et al., (2015) and to contribute to better interpretability, the topics containing the 10 most frequent words were provided labels and organized into 7 macro clusters that represent their most salient themes. For example, topic 152 labelled "enjoyment' contains words such as *fun moment quiet produce hang-out enjoy love learn live benefit*, and topic 89 labelled 'humor' contains the words *laugh hilarious tell funny do fun mischievous humor wow luck*. Hence, both topics seem to have a theme related to a cluster labelled 'Positive Experience'. By contrast, topics 81 labelled 'Loneliness' contained the words *people fight crazy loneliness quiet shit lie friends competition circular*, and topic 150 labelled 'anxiety' contained the words *fear scary afraid frightening anxiety person getting safe thoughts aggression,* where both topics seem to have a theme related to the 'Negative Experience' cluster.

These results underscore the ability of the model to generate both abstract terms topics such as topic 22, labelled 'communication' which was assigned to the 'Treatment' cluster and contains the words *place therapist life patient say ask listen person process,* and more concrete topics such as topic 140 –labelled 'Smoking' which contains the words *smoker drugs fume cigarettes smell alcohol smoke stop drink toilet*. Since the model is unsupervised, some topics may not correspond to interpretable concepts. Topic 91 (not shown in the table) clustered the words *come, say, arrive, friend, money, back, sister, evening, girl, park.*

The second output was *d* document distribution vectors V_{doc_i}, where each document is a client speech turn. We were interested in topic distribution at the session level (in order to

associate topics with session level labels), and thus aggregated turn topics into session topics by averaging all m_j document vectors derived from every session j; thus, $V_{session_j} = \frac{1}{m_j}\sum_{i=1}^{m_j} V_{doc_i}$. This amounts to a uniform linear interpolation of the turn-based topic distribution into a session topic distribution.

Using the Topics to Identify Clients' Functioning levels and Alliance Rupture

Our first goal (Hypothesis 1) was to determine which of the topics extracted in the previous stage are most associated with clients' levels of functioning and alliance ruptures. To do so, we used a common machine learning approach where the model is trained to predict[1] a particular outcome (in our case, clients' functioning labels and alliance ruptures) from a given set of feature inputs (in our case, 200 topic distribution vectors). The prediction models were repeatedly updated during a training process in which model-based predictions were compared to corresponding known outcomes. During the training period, incorrect predictions caused the model to adjust how it predicted the outcome, and continuously improve its predictive performance by a strictly data-driven approach. To test model generalizability, after the training period the models underwent a testing period, where their predictive performance was empirically tested with new data not used during training. Specifically, we trained a Sparse Multinomial Logistic Regression (SMLR; Yamashita, Sato, Yoshioka, Tong, & Kamitani, 2008) model to predict the label classes (ORS and PSQ). The resulting trained model predicted the labels as substantially above-random accuracy. The SMLR model simultaneously performs feature selection and training of model parameters for classification by determining the importance of each feature while estimating the corresponding parameter value. This process

[1] Note that our use of prediction terminology does not imply a causal relationship between the variables, but rather that the topics are trained to identify the labeled data (e.g., high vs. low reported functioning).

selects only a few features as important and prunes away the others. We then examined which features (topics) were informative for the model's prediction tasks.

Processing ORS outcomes. The *ORS* outcome score is a continuous value in the range of 0–40, converted to a binary outcome of {0, 1}, based on the clinical cutoff of 24 (Miller et al., 2003), where any score below 24 was set to 0 and any score above 24 was set to 1. The rationale for this procedure is that ratings below and above the cutoff are the most informative for differentiating between low and high functioning, and allow for a balanced binary outcome (for a similar approach, see Atkins et al., 2012). The division resulted in 400 samples of 0-class labels and 449 samples of 1-class labels.

Processing PSQ outcomes. The *PSQ* outcome score is a discrete value in the range of 1-5, converted to a binary class of {0,1}, where a choice of 0 means no rupture has occurred (the score was 0) and a choice of 1 means any degree of rupture has occurred (the score was 1,2,3 or 4) (for a similar approach, see Muran et al., 2009). This division resulted in 432 samples of 0-class labels and 397 samples of 1-class labels. (The inconsistency between the counts of sessions (873), ORS sample (849), and PSQ sample (829) reflects the incompleteness of material in some specific sessions.)

Model implementation. Two datasets were created: The first dataset for the ORS: $S_{ORS} = \{(x_1, y_1), \ldots, (x_{849}, y_{849})\}$, consisted of 849 samples where each instance x_i was a vector of 200 elements that represented the 200 topic probabilities $x_i \in \mathbb{R}^{200}$; and each instance of y_i was a label from a set of 2 possible classes $y_i \in Y = \{0,1\}$. The second dataset for the PSQ: $S_{PSQ} = \{(x_1, y_1), \ldots, (x_{829}, y_{829})\}$, consisted of 829 samples where each instance x_i was a vector of 200 elements that represented the 200 topic probabilities, and each instance of y_i was a label from a set of 2 possible classes $y_i \in Y = \{0,1\}$.

We trained an SMLR model on each dataset with default model parameters (Python implementations available at https://github.com/KamitaniLab/smlr).

Cross-validation was implemented on 80% of the training data (i.e., data that the model learned) and 20% of the test data (i.e., data not used during the training process, and used here for test purposes for the first time). This data division was performed while keeping clients distinct across the train and test datasets, such that a client's data were only in either the train or the test set. This led to attributing 46 client labels to the training dataset and 12 client labels to the test dataset.

Model performance was assessed by a simple accuracy measure of the percentage of correctly predicted labels, which was compared to the performance of the majority baseline accuracy (i.e., the binary labels' distribution over the entire dataset). The results are presented in Table 2, where, as can be seen, the model predicted clients' functioning labels with an accuracy of 75% and alliance rupture labels with an accuracy of 65%, which is superior to the majority baseline accuracy results, indicating meaningful model learning. As is typical in supervised ML models, the training accuracy was higher than the testing accuracy; however the difference was relatively small (3.22% for the ORS and 7.72% for the PSQ) and the test-set accuracies (75% for functioning and 65% for alliance rupture) reflected the true behavior of the model on the test data.

We then explored which topics in each session best predicted functioning level and alliance rupture. The logistic model ranks input features (i.e., topics) by effectiveness for the model's learning process. Table 3 presents the top four topics for predicting classes 0 and 1 for the ORS and class 1 for the PSQ (class 0 for the PSQ was not included since non-occurrence of rupture is not a distinct event but rather the typical situation). As seen in the table, topics 72, 15,

152, and 171 describe "celebration," "leisure experience," "enjoyment," and "choice," which intuitively seem to be related to positive experiences and to high functioning. On the other hand, topics such as 81, 199, 166, and 61 seem to be about "loneliness," "suffering," "physical difficulties," and "anger," which intuitively seem related to negative experiences and to low functioning. For PSQ labels, topics 22, 165, 133, and 48 included themes of "communication," "goal setting," "needing help," and "problems," which appear to be related to the therapeutic relationship and to an experience of rupture.

Exploring the Association between Topic and Outcome

The second goal (Hypothesis 2) was to explore how topics that were most closely associated with functioning and rupture level changed over the course of treatment, and whether client-level differences in these trajectories were associated with changes in clients' symptomatology. For this purpose, we ran a series of unconditional multi-level growth models, in which the outcome was the weighted average of the four topics that were found to be most closely associated with each of the three labels (i.e., classes 0 and 1 for the ORS and class 1 for the PSQ) and the predictor was session number. An additional unconditional growth model was run to estimate changes in clients' symptoms (i.e., HSCL). We then extracted the empirical Bayes estimates of the session random effects (representing client-level differences in the change trajectories) from each model to assess the association between change in topics and change in clients' symptoms. None of the three topics showed a linear[2] change (*ps* > .602); however, as Figure 2 shows, and in line with our hypothesis (2a), changes in topics over the course of treatment were associated with change in outcome. Specifically, an increase in the high functioning topics (clustered under *positive experience*s) was associated with a decrease in

[2] The same null pattern was found when we tested quadratic and log10 trajectories

clients' symptoms (r = -.29, p = .03), whereas an increase in the topics most closely associated with rupture occurrence (clustered under *treatment*) was associated with an increase in clients' symptoms (r = .32, p = .03). No association was found between the trajectory of the low functioning topics (clustered under *negative experience*s) and clients' symptoms (r = .12, p = .34).

Next, in order to illustrate one potential use of topic models for clinicians and researchers, we explored how topics most closely associated with functioning and ruptures changed over the course of one good and one poor outcome case (Exploratory hypothesis 2b). To do so, the proportional value of the four topics (for both ORS and PSQ) of the 200 topics was used for each session. Figure 3 depicts the topics that were found to be most closely associated with the ORS (i.e., topics 81, 199, 166, and 61 for low functioning and 72, 160, 152, 171 for high functioning) for Noya (good outcome) and Gali (poor outcome) session by session throughout their treatment. As seen in the figure, in Noya's case, low functioning topics (negative experiences cluster) tended to decrease during treatment, whereas high functioning topics (positive experiences cluster) tended to increase. In contrast, Gali's low functioning topics tended to increase throughout therapy, and signs of high functioning topics did not appear.

In addition, Figure 4 shows the topics that were found to be most closely associated with the PSQ (i.e., topics 22, 165, 133, and 48 for ruptures; all under the "treatment" cluster). As seen in the figure, Noya's rupture topic signal peaked mid-treatment and was low at beginning and late stages; in contrast, in Gali's case, a strong signal for rupture topics tended to appear repeatedly throughout treatment as a whole.

Discussion

Advanced machine learning techniques are relatively novel in psychotherapy research, but emerging evidence suggests the value of integrating them into traditional measures commonly applied to therapy (Dwyer, Falkai, & Koutsouleris, 2018). We used topic modeling, a data-driven machine learning technique that extracts latent topics from textual data to examine which topics best identify clients' functioning and alliance ruptures in psychotherapy sessions, and whether changes in these topics were associated with changes in treatment outcome.

Topic modeling yielded semantically meaningful topics that were then used to identify session level clients' functioning and rupture. Consistent with our first hypothesis, the SMLR models with topic models features identified labels above chance, at 65% (alliance ruptures) to 75% (clients' functioning) test accuracy. The model performance in the current study replicates the successful application of topic modeling in previous studies (e.g., Atkins et al., 2012; Imel et al., 2015). As was the case for our results, Atkins et al. (2012) found that Sparse Logistic Regression models with topic models features predicted behavioral codes drawn from couples' psychotherapy with between 65% to 70% accuracy, and Imel et al. (2015) used features that are derived from topic models to identify therapists' interventions with high accuracy (87%).

A possible explanation for the higher model performance identifying clients' functioning than identifying alliance ruptures could be related to the difference in the nature of these tasks. Machine learning algorithms, like humans, tend to perform better with more concrete, less ambiguous tasks. Level of functioning appears to be more concrete and explicit, whereas alliance ruptures are more abstract and implicit. In fact, rating alliances is also challenging for human coders, as shown in previous studies that have revealed weak correlations between clients' ratings and those of observers (Tichenor & Hill, 1989). An alternative explanation is that both text and functioning ratings are "products" of the same person (the client), while alliance

ruptures are rated by a different person (the therapist). Future studies would benefit from examining textual data from both clients and the therapists along with ratings of process and outcome measures from the perspectives of both members of the dyad to obtain a fuller picture of which content is associated with these labels.

In our results, the topics most closely associated with both functioning level and alliance ruptures seem to make intuitive sense in terms of their associated labels. Since topic modeling is a data-driven inductive measure, the results may shed light on what themes are most associated with low and high functioning. For example, the four topics most closely associated with low functioning—loneliness, suffering, physical difficulties, and anger—clearly relate to negative experiences, while the topics most closely associated with high functioning - enjoyment, leisure experience, celebration, and choice seem related to positive experiences. These findings are consistent with Atkins et al., (2012) who used LDA to demonstrate the ability of topic models to capture themes associated with positive and negative emotional experiences in couples' therapy.

Recent work using observational coding measures has shown the value of focusing on clients' language to assess treatment outcome (Poulin, Button, Westra, Constantino, & Antony, 2019). The ability of topic models to capture themes associated with level of functioning without requiring human coding may suggest that using these measures can be considered alongside standard monitoring systems in order to evaluate clients' progress.

Similarly, the three topics most closely associated with alliance rupture - communication, goal setting, needing help, problems - seem to capture themes that correspond to Bordin's (1979) definition of alliance and to Safran and Muran's definition of alliance rupture as consisting of tension in one or more alliance components (bond, task, and goals). This finding is consistent with Imel et al., (2015) who reported topic modeling's ability to identify semantically

meaningful topics and use them to predict sessions that were labeled as focusing on client-therapists relationships. Though it may be unsurprising that themes related to client-therapist relationships were common in sessions with rupture occurrence, these topics were extracted automatically from the session text, supporting the ability of topic modeling to capture important aspects of the clinical encounter. The identification of the specific content that led to rupture occurrence may contribute to a better understanding of these events. Future studies may benefit from employing clinical judges' information about resolved versus unresolved ruptures to learn which topics best identify rupture resolution.

In line with our second hypothesis (2a), change trajectories in topics were associated with change trajectories in clients' symptoms over the course of treatment. Specifically, the findings indicated that increases in the high functioning topics which included themes related to positive experiences were associated with decreased symptoms. Interestingly, changes in the low functioning topics which included themes related to negative experiences were not associated with changes in outcome. These findings are consistent with contemporary psychotherapy theories that highlight the important role of positive experience as a main transformational agent in psychotherapy (e.g., Fosha, 2004; Stalikas et al., 2015). While in classical psychodynamic psychotherapy models there was a tendency to neglect positive experience and to focus on expanding clients' ability to tolerate negative and painful experience, in recent years there has been increasing recognition that broadening clients' ability to experience both positive and negative emotions has an important curative effect (Fosha, 2004; Roten, Drapeau & Michel, 2008).

Our findings are also in line with recent studies indicating that positive outcomes are associated with increased positive experiences (e.g., Atzil et al., 2019; Atzil, Wiseman & Tishby,

2015; Stalikas et al., 2015). Whereas these previous studies relied on self-reports or coding systems to assess clients' experiences, the current study contributes by introducing a computerized method to identify topics discussed by clients during psychotherapy sessions.

Our findings also showed that the trajectory of linear decrease in rupture topics was associated with the trajectory of decrease in symptoms. This finding is consistent with previous studies that showed an association between linear progress in the therapeutic alliance during treatment and improved outcome (e.g., Kramer, de Roten, Beretta, Michel, & Despland, 2009). However, other studies have shown that a V-shaped deflection in the alliance (interpreted as rupture-repair sequences) were associated with greater gains (e.g., Stiles et al., 2004). Given the importance of examining nonlinear changes in psychotherapy (Hayes et al., 2007), we also explored the associations between non-linear patterns of change in both functioning and rupture topics on the one hand and symptoms on the other, but did not find significant associations. More research with intensive repeated measurements is needed to further examine whether non-linear changes in topics are associated with treatment outcomes.

In order to demonstrate the utility of topic modeling to summarize content discussed in psychotherapy sessions visually in a way that may be valuable to clinicians and researchers, we also focused on specific clients in our sample and explored whether topics identified in the previous stage as most closely associated with clients' functioning and alliance rupture would show different patterns of change in one good versus the one poor outcome case (Hypothesis 2b). The heatmaps (Figure 3 and 4) provide a visualization of how these topics evolved over treatment and show that in Noya's case (good outcome), the negative experience topics tended to decrease during treatment, whereas the positive experience topics tended to increase. In contrast,

in Gali's case (poor outcome) the negative experience topics tended to increase throughout therapy, and signs of positive experience topics did not emerge.

A different pattern of change in these two cases was also seen in topics most associated with alliance rupture. Noya's rupture topics showed a strong signal in the middle of treatment, while at the beginning and end of treatment the signal was weak; in contrast, Gali's rupture topics showed a high signal repeatedly throughout treatment. While the pattern of change in Noya's and Gali's treatment was somewhat different than the results found for the entire sample, the presentation of these specific cases shows how therapists can use topic modeling as a summary of the content that was discussed by their clients in specific sessions or throughout treatment.

Topic modeling vs. other Machine Learning Models

This work employed topic modeling, specifically LDA, to infer a small number (n = 200) of latent interpretable semantic topics from psychotherapy data. These latent inferred topics were then used either as features for a regularized classifier (to test Hypothesis 1 which used the extracted topics to predict clients' functioning and rupture occurrence), or as a mechanism for high-level data session summarization and exploration (which is demonstrated in Hypothesis 2a which shows that changes in specific themes were associated with change in outcome, and hypothesis 2b which visually shows changes in specific content for specific clients throughout treatment). From a technical perspective, the role of the LDA algorithm is to achieve dimensionality reduction by grouping words together into semantically meaningful clusters. Other dimensionality reduction methods to achieve similar goals include Latent Semantic Indexing (LSA; Landauer & Dumais, 1998), probabilistic Latent Semantic Indexing (pLSI, Hoffman, 1999) and Non-negative Matrix Factorization (NMF, Lee & Seung, 2000).

Historically, pLSI is a probabilistic extension of LSA, whereas LDA is a Bayesian extension of pLSI that introduced modeling with Dirichlet priors. Practically, the use of the Dirichlet prior in LDA allows the modeling to encourage a small number of topics in each document, in contrast to LSA, pLSI and NMF, where no such prior is given. Consequently the topics produced using LDA are often more coherent and more closely associated with each document.

This work examined which topics best identified clients' functioning and alliance ruptures using ML techniques. If the goal were solely predictive performance (to obtain a classifier that can reliably predict clients' functioning and alliance ruptures from text), LDA fed into a highly regularized classifier may not be the best approach. In particular, a set of predictive models could be used, such as (a) classifiers that operate directly on the words of the text as symbolic units, such as a Naive-Bayes SVM (Wang & Manning, 2012), Decision Forests, (e.g., Ho, 1998), or gradient boosted trees (Friedman, 2001) or (b) classifiers that operate over a neural-encoding of the text using neural network approaches such as word embeddings (e.g., Mikolov, Chen, Corrado & Dean), recurrent encoders (Elman, 1990) attention-based encoders (Vaswani et al., 2017), or pre-trained attention based language models (BERT, Devlin, Chang, Lee, Toutanova, 2018). These predictive models often produce very strong predictive accuracy. However, they also require large amounts of training data, and are opaque black boxes, whose inner working cannot be interpreted. By contrast, high-quality LDA topics can be inferred from relatively small amounts of data. Our intention was not to produce the best classification model for clients' functioning and alliance ruptures, but to provide insights into the therapeutic process. Using LDA-derived features fed into a highly regularized classifier, as done here, makes it possible not only to obtain predictions, but also to associate the prediction with specific interpretable topics, and to link topics with outcomes.

Limitations and Future Directions

Machine learning methods in general and NLP methods in particular have a notable practical limitation: for significant insights they require big data. Though in our research clinic we have a very large corpus of audio-taped sessions, only 873 sessions were transcribed for this study due to the cost of transcription and the labor involved. While the study included a fairly large sample of 28,323 documents, model performance was likely reduced due to small sample size which may limit the generalizability of the results. However, with advances in speech-to-text technologies (e.g., Google Transcribe) it seems likely that transcription will soon be conducted automatically, allowing for the analysis of much larger datasets.

Furthermore, in our sample the functioning level was relatively high (56% were in the nonclinical range), compared to rates in other outpatient samples (for example in Miller et al., 2006, 25% were in the nonclinical range) and the therapists were trainees in a program that emphasizes a psychodynamic model of treatment. Both these factors could limit the generalizability of the findings. The data reported here were collected in a training clinic that screens out many clients with more complicated diagnostic profiles. Although this sample resembles samples from studies conducted in other university counseling centers (for example in Minami et al., 2009, only 34% of the clients were above the clinical cutoff) and though previous studies have shown that clients' speech behavior often tends to be similar and demonstrates parallel patterns across orientations and modalities (Mergenthaler, 2015), future studies are required to explore whether the results can be replicated in other settings.

Another limitation is that our topic model was based on the "bag of words" approach that does not consider sequence of words or topics. However, word sequences are likely to be very important for human communication. Using more advanced, dialogue-related models that take

into account word and topic sequential structure (e.g. Purver, Körding, Griffiths, & Tenenbaum, 2006; Park et al., 2019) should help produce models that better describe the psychotherapy interaction and identify what aspects of client–therapist dialogue lead to better treatment outcomes. In addition, this work used clients' texts alone; therapist speech turns were not analyzed. However, therapist texts are clearly a main component of the therapeutic process. Future work could examine which therapists' topics are associated with clients' functioning and rupture occurrence. This could have significant implications for clinicians, since they are not always aware of their clients' experience, but they do have more control over their own speech.

Note as well that the rupture occurrence labels were based on therapists' reports derived from only one item from the PSQ questionnaire, consistent with previous studies on ruptures in the therapeutic process (e.g., Muran et al., 2009). However, therapists may often be unaware of the occurrence of a rupture (Chen et al., 2018). Future studies may benefit from using rupture measures that are based on clients' or observers' perspectives, which may allow for a more objective assessment of ruptures. Furthermore, using moment-by-moment measures of ruptures (e.g., Eubanks, Muran & Safran, 2015) may provide richer and larger data to train the models on and may lead to a better grasp of alliance ruptures that occur outside therapists' awareness.

Another limitation is the usage of current functioning labels. The predictive models were trained over labels derived from clients' self-reports completed at the beginning of the session. Future studies would benefit from assessing whether topic modeling can capture session outcomes and train models over labels derived from therapists' assessment of clients' functioning.

Finally, the model was restricted to text and did not have access to acoustics or facial expressions during treatment, which are also important (Imel et al., 2014). Much information

about the therapeutic interaction lies in the non-verbal and para-verbal facets of communication between clients and therapists. Recently computerized methods to explore small units within psychotherapy sessions such as non-verbal behavioral methods (e.g., Altmann et al., 2019; Ramseyer, 2019) are starting to be used in psychotherapy research. Future studies could explore how components acting on different time scales and measured by different methods (e.g., textual, vocal, physiological measures) interrelate and their association with treatment outcome.

With respect to clinical implications, our results may enable therapists to be better attuned to specific topics that may signal important events in therapy. The information provided by topic model output and the associations between the resulting topics and other process and outcome variables can be made available to clinicians through a specialized user interface that allows for conceptual exploration of the therapy process. This information could allow therapists to access a summary of topics discussed in a session, locate specific themes associated with rupture or with clients' deterioration (or other important events in therapy), and direct interventions to improve the situation. Future studies could use topic model output alongside existing monitoring tools to inform therapists of meaningful linguistic processes that occur within psychotherapy sessions.

References

Adler, M. (2007). *Hebrew morphological disambiguation: An unsupervised stochastic word-based approach*. Beersheba: Ben-Gurion University of the Negev.

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology, 26*(5), 816–827.

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science, 9*(1), 49.

Atzil-Slonim, D., Bar-Kalifa, E., Fisher, H., Lazarus, G., Hasson-Ohayon, I., Lutz, W., ... & Rafaeli, E. (2019). Therapists' empathic accuracy toward their clients' emotions. *Journal of consulting and clinical psychology, 87*(1), 33.

Atzil–Slonim, D., Wiseman, H., & Tishby, O. (2016). Relationship representations and change in adolescents and emerging adults during psychodynamic psychotherapy. *Psychotherapy Research, 26*(3), 279-296.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice, 16*(3), 252–260.

Bucci, W., Kabasakalian, R., & The RA Research Group (1992). *Instructions for scoring referential activity (RA) in transcripts of spoken narrative texts*. Ulm: Ulmer Textbank.

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering, 23*(5), 649–685.

Chen, R., Atzil-Slonim, D., Bar-Kalifa, E., Hasson-Ohayon, I., & Refaeli, E. (2018). Therapists' recognition of alliance ruptures as a moderator of change in alliance and symptoms. *Psychotherapy Research, 28*(4), 560–570.

Coutinho, J., Ribeiro, E., Sousa, I., & Safran, J. D. (2014). Comparing two methods of identifying alliance rupture events. *Psychotherapy, 51*(3), 434–442.

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

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of Brief Therapy*, *3*(1), 3–12.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91–118.

Elman, J. L. (1990). Finding structure in time. Cognitive science, 14(2), 179-211.

Eubanks, C.F., Muran, J.C., & Safran, J.D. (2015). Rupture Resolution Rating System (3RS): Manual. Unpublished manuscript, Mount Sinai-Beth Israel Medical Center, New York.

Eubanks-Carter, C., Muran, J. C., & Safran, J. D. (2015). Alliance-focused training. *Psychotherapy, 52*(2), 169–173.

Eubanks, C. F., Muran, J. C., & Safran, J. D. (2018). Alliance rupture repair: A meta-analysis. *Psychotherapy, 55*(4), 508–519.

Eubanks-Carter, C., Gorman, B. S., & Muran, J. C. (2012). Quantitative naturalistic methods for detecting change points in psychotherapy research: An illustration with alliance ruptures. *Psychotherapy Research, 22*(6), 621–637.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy, 55*(4), 316–340.

Gaut, G., Steyvers, M., Imel, Z. E., Atkins, D. C., & Smyth, P. (2015). Content coding of psychotherapy transcripts using labeled topic models. *IEEE Journal of Biomedical and Health Informatics, 21*(2), 476–487.

Gelo, O. C. G., Salcuni, S., & Colli, A. (2012). Text analysis within quantitative and qualitative psychotherapy process research: introduction to special issue. Research in Psychotherapy: Psychopathology, Process and Outcome, 45–53.

Goldberg, Y. (2011). *Automatic syntactic processing of Modern Hebrew.* Beersheba: Ben Gurion University of the Negev.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Hafkenscheid, A., Duncan, B. L., & Miller, S. D. (2010). The Outcome and Session Rating Scales: A cross-cultural examination of the psychometric properties of the Dutch translation. *Journal of Brief Therapy*, *7*(1), 1-12.

Hatfield, D. R., & Ogles, B. M. (2006). The influence of outcome measures in assessing client change and treatment decisions. *Journal of Clinical Psychology*, *62*(3), 325–337.

Hayes, A. M., Laurenceau, J. P., Feldman, G., Strauss, J. L., & Cardaciotto, L. (2007). Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy. *Clinical psychology review*, *27*(6), 715-723.

Hill, C. E., & Lambert, M. J. (2004). *Methodological issues in studying psychotherapy processes and outcomes*. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed.; pp. 84–135). Hoboken: Wiley.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266.

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57).

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

Kazdin, A. E. (2016). Single-case experimental research designs. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 459–483). Washington, DC: APA.

Kramer, U., de Roten, Y., Beretta, V., Michel, L., & Despland, J. N. (2009). Alliance patterns over the course of short-term dynamic psychotherapy: The shape of productive relationships. *Psychotherapy Research*, *19*(6), 699-706.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*(2), 211.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788-791.

Lutz, W., Tholen, S., Schürch, E., & Berking, M. (2006). Die entwicklung, validierung und reliabilität von kurzformen gängiger psychometrischer instrumente zur evaluation des therapeutischen fortschritts in psychotherapie und psychiatrie. Diagnostica, 52, 11–25. http://dx.doi.org/ 10.1026/0012-1924.52.1.11.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. Retrieved from http://mallet.cs.umass.edu

Mergenthaler, E. (1996). Emotion–abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology, 64*(6), 1306–1315.

Mergenthaler, E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research, 18*(2), 109–126.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv

Miller, S. D., Duncan, B. L., Brown, J., Sorrell, R., & Chalk, M. B. (2006). Using formal client feedback to improve retention and outcome: Making ongoing, real-time assessment feasible. *Journal of Brief Therapy, 5*(1), 5-22.

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

Minami, T., Davies, D. R., Tierney, S. C., Bettmann, J. E., McAward, S. M., Averill, L. A., ... & Wampold, B. E. (2009). Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. *Journal of Counseling Psychology, 56*(2), 309.

Montesano, A., Gonçalves, M. M., & Feixas, G. (2017). Self-narrative reconstruction after dilemma-focused therapy for depression: A comparison of good and poor outcome cases. *Psychotherapy Research, 27*(1), 112–126.

More, A., & Tsarfaty, R. (2016, December). Data-driven morphological analysis and disambiguation for morphologically rich languages and universal dependencies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 337–348). Osaka: COLING Organizing Committee.

Muran, J. C., Safran, J. D., Gorman, B. S., Samstag, L. W., Eubanks-Carter, C., & Winston, A. (2009). The relationship of early alliance ruptures and their resolution to process and outcome in three time-limited psychotherapies for personality disorders. *Psychotherapy, 46*(2), 233–248.

Muran, J. C., Safran, J. D., Samstag, L. W., & Winston, A. (2004). *Patient and therapist postsession questionnaires, Version 2004.* New York: Beth Israel Medical Center.

Park, J., Kotzias, D., Kuo, P., Logan IV, R. L., Merced, K., Singh, S., … Tai-Seale, M. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-

provider interactions. *Journal of the American Medical Informatics Association, 26*(12), 1493–1504.

Peek, N., Combi, C., Marin, R., & Bellazzi, R. 2015. Artificial intelligence in medicine thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial Intelligence In Medicine, 65*(1), 61–73.

Poulin, L. E., Button, M. L., Westra, H. A., Constantino, M. J., & Antony, M. M. (2019). The predictive capacity of self-reported motivation vs. early observed motivational language in cognitive behavioural therapy for generalized anxiety disorder. *Cognitive Behaviour Therapy, 48*(5), 369–384.

Purver, M., K. Körding, T. Griffiths, & J. Tenenbaum (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL), Sydney, Australia* (pp. 17–24). Stroudsburg, PA: Association for Computational Linguistics.

Safran, J. D., & Muran, J. C. (2006). Has the concept of the therapeutic alliance outlived its usefulness? *Psychotherapy, 43*(3), 286–291.

Salvatore, S., Gelo, O. C. G., Gennaro, A., Metrangolo, R., Terrone, G., Pace, V., … Ciavolino, E. (2017). An automated method of content analysis for psychotherapy research: A further validation. *Psychotherapy Research, 27*(1), 38–50.

Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *American Psychologist, 65*(2), 98–109.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International neuropsychiatric interview

(MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry, 59*, 22-33.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (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.

Stalikas, A., Fitzpatrick, M., Mistkidou, P., Boutri, A., & Seryianni, C. (2015). Positive emotions in psychotherapy: Conceptual propositions and research challenges. In O. C. G. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy research: Foundations, process, and outcome* (pp. 331–349). Vienna, Austria: Springer Vienna. http://dx.doi.org/10.1007/978-3-7091-1382-0_17

Steyvers, M., & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis: A road to meaning* (pp. 5221–5228). Hillsdale, NJ: Laurence Erlbaum.

Stiles, W. B., Glick, M. J., Osatuke, K., Hardy, G. E., Shapiro, D. A., Agnew-Davies, R., … Barkham, M. (2004). Patterns of alliance development and the rupture-repair hypothesis: Are productive relationships U-shaped or V-shaped? *Journal of Counseling Psychology, 51*(1), 81–92.

Summers, R. F., & Barber, J. P. (2009). *Psychodynamic therapy: A guide to evidence-based practice*. New York: Guilford Press.

Tichenor, V., & Hill, C.E. (1989). A comparison of six measures of working alliance. *Psychotherapy, 26*(2), 195–199.

Tracey, T. J., & Kokotovic, A. M. (1989). Factor structure of the Working Alliance Inventory. *Psychological Assessment, 1*, 207–210.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Wang, S. I., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2) (pp. 90-94).

Whipple, J. L., & Lambert, M. J. (2011). Outcome measures for practice. *Annual Review of Clinical Psychology, 7*, 87–111.

Wiseman, H., and Tishby, O. (2014). Client attachment, attachment to the therapist and client–therapist attachment match: How do they relate to change in psychodynamic psychotherapy? *Psychotherapy Research, 24*, 392–406. DOI:10.1080/10503307.2014.892646.

Yamashita, O., Sato, M. A., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, *42*(4), 1414–1429.

Table 1
Topic Model Output - Selection and Characterization of 35 Topics

Cluster	Topic Name	Topic Number	Top 10 Words in the Topic
Positive Experience	Enjoyment	152	fun, moment, quiet, produce, hang-out, enjoy, love, learn, live, benefit
	Humor	89	laugh, hilarious, tell, funny, do, fun, mischievous, humor, wow, luck
	Celebration	72	birthday, present, Sabbath (Saturday), gift, bless, luck, surprise, friends, party, celebrate
	Choice	171	choice, change, behavior, approach, pattern, situation, think, positive, good, reverse
	Hope	44	life, period, happiness, years, happy, age, good, live, joy, happiness
Negative Experience	Loneliness	81	loneliness, quiet, people, fight, crazy, shit, lie, friends, competition, circular
	Suffering	199	suffer, feel, person, people, cause, others, hate, poor, shame, act
	Anger	61	nerve, fight, know, time, patience, shout, suffer, furious, quiet, quarrel
	Anxiety	150	fear, scary, afraid, frightening, anxiety, person, getting, safe, thoughts, aggression
	Pain	131	crying, weep, cry, mother, tears, see, happen, sensitive, pain, hysteria
Relationships	Intimacy	37	love, like, fun, real, hug, live, get, kiss, hate, getting-closer
	Romantic	97	egalitarian, couple, relationship, date, favor, find, serious, date, opportunity, love
	Family	187	family, mother, aunt, children, sister, uncle, brother, grandmother, parents, grandchildren
	Children	68	children, kid, mother, time, age, small, baby, family, life, house
	Marriage	64	marriage, wedding, life, children, bring, honeymoon, pregnancy, split, ups,
Treatment	Communication	22	communication, place, therapist, life, patient, say, ask, listen, person, process
	Goal Setting	165	purpose, target, middle, reach, distance, aim, far-away, far, away, safe-haven, stunning
	Needing Help	133	help, ask, please, want, assist, give, facilitate, situation, give, request
	Problems	48	problem, speak, want, issue, solve, truth, story, solution, end, psychologist
	Process	83	treatment, therapist, think, psychological, continue, meeting, truth, tell, process, patient
Health	Physical Difficulties	166	surgery, stomach, painful, foot, hurt, body, wheelchair, doctor, inflammation, fracture
	Mind-Body	13	soul, mind, health, course, service, rehabilitation, human, hostel, accompaniment, association
	Smoking	140	smoker, drugs, fume, cigarettes, smell, alcohol, smoke, stop, drink, toilet
	Health Care	161	doctor, patient, examination, physician, ER, turn, medication, private, dialysis, emergency-room
	Nutrition	134	weight, sport, kilogram, eat, gained, diet, lost, thin, fat, healthy
Everyday Life	Writing	127	write, compose, letter, story, book, record, diary, email, Facebook, words
	Work	58	employee, working, office, people, director, business, boss, customer, domain, marketing
	School	149	grade, school, teacher, pupils, children, high-school, year, education, lesson, well, educator
	Studies	10	learn, studies, degree, course, university, test, domain, profession, year, math
	Finances	30	money, pay, invoice, month, bank, price, apartment, cost, amount, salary
Miscellaneous	Family background	123	Holocaust, land, country, war, Israel, Jews, family, world, election, Arabic
	Geography	27	country, city, place, center, travel, living, region, street, dwell, north
	Drinks	19	water, coffee, glass, up, jump, wine, bottle, beer, drink, Coca-Cola
	Household duties	177	dishes, laundry, kitchen, water, shower, wash, sink, dishwasher, clothing, washing-machine

Weather	103	air, air-conditioner, light, winter, weather, rain, heat, summer, sun, cold

Note: Thirty-five selected topics were sampled out of the 200 output topics. Each topic was labeled and assigned to seven clusters by the authors to ease interpretation. Each topic is a probability distribution over all the words in the corpus; the top 10 words are listed here for purposes of convenience. Note that the source data were transcribed in Hebrew; in the table, the words were manually translated into English.

Table 2

Sparse Logistic Regression Model Accuracy Results for ORS and PSQ Label Prediction

Label	Majority Baseline	Train		Test	
	Accuracy (%)	Accuracy (%)	Δ^a (%)	Accuracy (%)	Δ^b (%)
ORS	52.88	78.81	25.93	75.59	22.71
PSQ	52.11	77.02	24.91	65.32	13.21

Note: Train and test accuracy result over the ORS (Outcome Rating Scale; Miller et al., 2003) and PSQ (Post-Session Questionnaire; Muran et al., 2004) models were compared to majority baseline accuracy. Δ^a represents the difference between the train accuracy and the baseline accuracy. Δ^b represents the difference between the test accuracy and baseline accuracy. The train and test data were derived from the entire dataset according to a 80%–20% division, while keeping clients distinct across the train and test datasets.

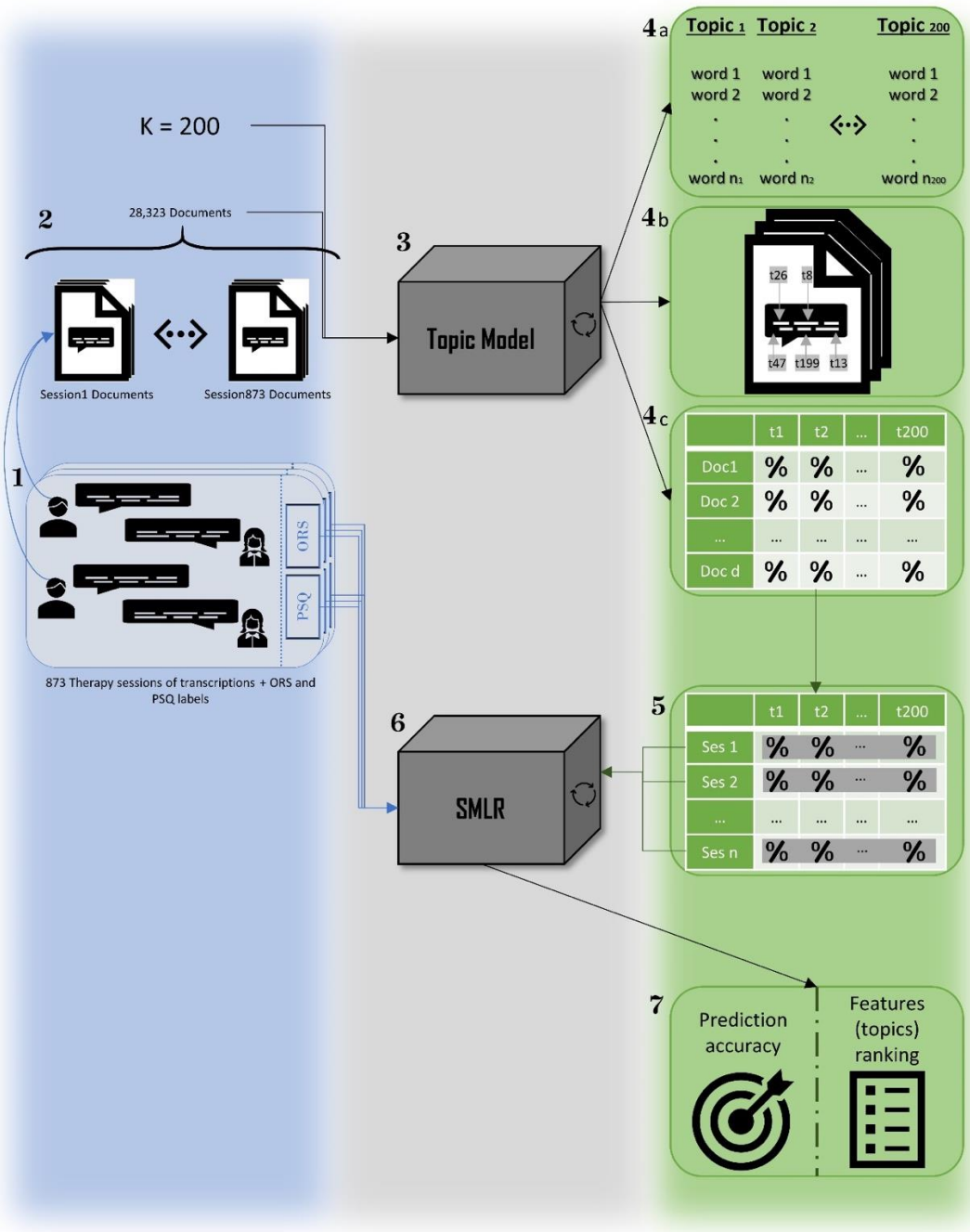
Table 3

Most Effective Top 10 Words per Topic for Label Class Identification

Label		Topic			
Name	Class	Number	Name	Cluster	Top 10 Words
ORS	0	81	Loneliness	NE	loneliness quiet people fight crazy shit lie friends competition circular
		199	Suffering	NE	suffer feel person people cause others hate poor shame act
		166	Physical Difficulties	Health	surgery stomach painful foot hurt body wheelchair doctor inflammation fracture
	1	61	Anger	NE	nerve fight know time patience shout suffer furious quiet quarrel
		72	Celebration	PE	birthday present Sabbath (Saturday) gift bless luck surprise friends party celebrate
		15	Leisure Experience	PE	play game football like basketball world sport card artist team
		152	Enjoyment	PE	fun moment quiet produce hang-out enjoy love learn live benefit
		171	Choice	PE	choice change behavior approach pattern situation think positive good reverse
PSQ	1	22	Communication	Treatment	communication place therapist life patient say ask listen person process
		165	Goal Setting	Treatment	purpose target middle reach distance aim far-away far away safe-haven stunning
		133	Needing help	Treatment	help ask please want assist give facilitate situation give request
		48	Problems	Treatment	problem speak want issue solve truth story solution end psychologist

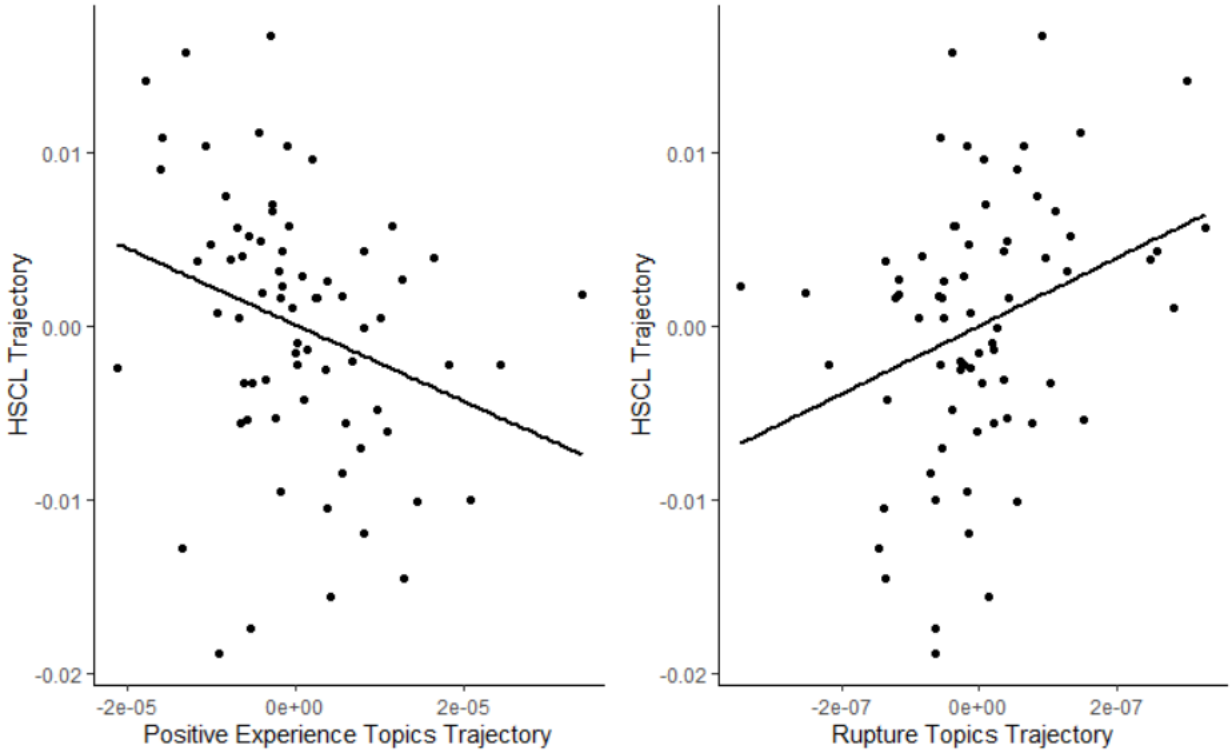
Note: Top 4 most effective topics. The topics are listed in their order of effectiveness by SMLR model ranking. ORS = Outcome Rating Scale (Miller et al., 2003); PSQ = Post Session Questionnaire (Muran et al., 2004). Class 0 for ORS = low functioning; Class 1 for ORS = high functioning; class 1 for PSQ = rupture occurrence; NE = Negative Experience; PE = Positive Experience.

Figure 1. End-to-end workflow visualization.



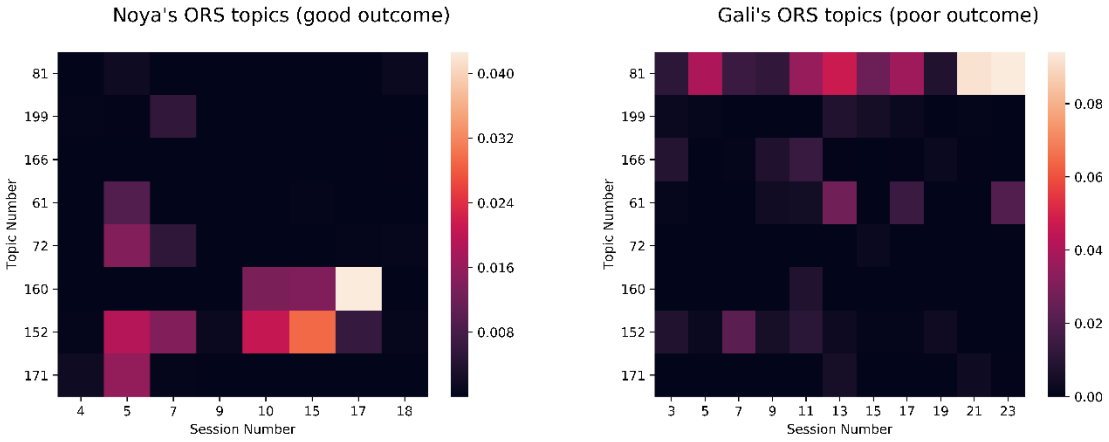
Note: Illustration of the topic modeling process, which starts by processing the input data (blue) through the analytic models used (grey) and their output results (green). The 873 transcriptions (1) were processed into 28,323 documents (2) that formed the input to the Topic model (3), which provided 3 outputs (4). The output in document-topic proportions was transformed to session-topic proportions (5). The vectors of these proportions were imported as X's, along with the transcription labels (1) as Y's, to the SMLR model (6) which resulted in a prediction outcome and feature ranking (7).

Figure 2. Association between change trajectories in topics and symptoms.



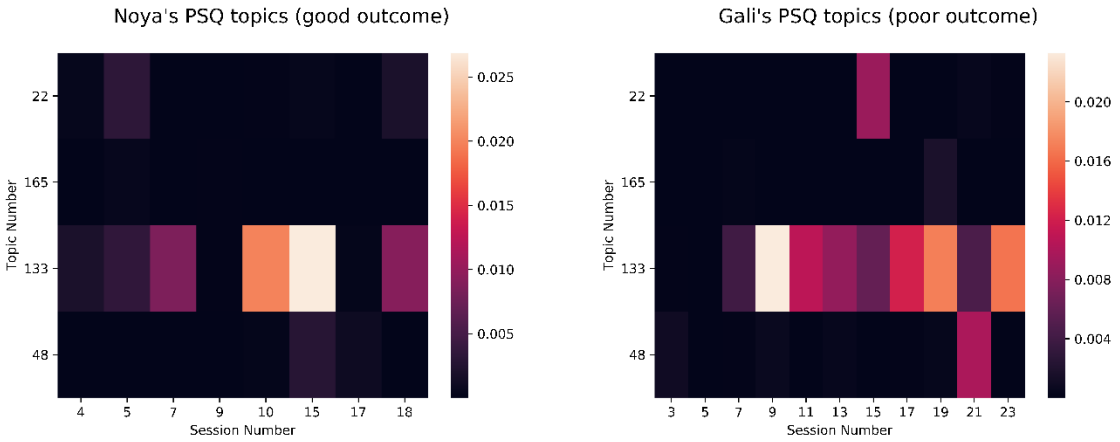
Note: HSCL= Hopkins Symptom Checklist (Lutz et al., 2006).

Figure 3. Change in topics most closely associated with ORS during good and poor outcomes cases.



Note: Heatmap plot of the proportion of sampled topics (most closely associated with ORS) from transcripts for Noya (good outcome) and Gali (poor outcome). Rows of the plot are defined by topic number, and columns are defined by session number. The lighter the topic representations, the stronger the signal. The color mapping was scaled for easier interpretation. Only 8 topics out of 200 are represented here for intuitive visualization; when all 200 topics are represented, the signal (probability) sums to 1.

Figure 4. Change in topics most closely associated with *PSQ* during good and poor outcomes cases.



Note: Heatmap plot of the proportion of sampled topics (most closely associated with PSQ) from Noya's and Gali's transcripts. Rows of the plot are defined by topic numbers, and columns are defined by session number. The lighter the topic representation, the stronger the signal. The color mapping was scaled for easier interpretation. Only 4 topics out of 200 are represented here for intuitive visualization; when all 200 topics are represented, the signal (probability) sums to 1.