Content Coding of Psychotherapy Transcripts Using Labeled Topic Models

Garren Gaut, Mark Steyvers, Zac E. Imel, David C. Atkins, and Padhraic Smyth, *Member, IEEE*

***Abstract*—Psychotherapy represents a broad class of medical interventions received by millions of patients each year. Unlike most medical treatments, its primary mechanisms are linguistic; i.e., the treatment relies directly on a conversation between a patient and provider. However, the evaluation of patient–provider conversation suffers from critical shortcomings, including intensive labor requirements, coder error, nonstandardized coding systems, and inability to scale up to larger data sets. To overcome these shortcomings, psychotherapy analysis needs a reliable and scalable method for summarizing the content of treatment encounters. We used a publicly available psychotherapy corpus from Alexander Street press comprising a large collection of transcripts of patient–provider conversations to compare coding performance for two machine learning methods. We used the labeled latent Dirichlet allocation (L-LDA) model to learn associations between text and codes, to predict codes in psychotherapy sessions, and to localize specific passages of within-session text representative of a session code. We compared the L-LDA model to a baseline lasso regression model using predictive accuracy and model generalizability (measured by calculating the area under the curve (AUC) from the receiver operating characteristic curve). The L-LDA model outperforms the lasso logistic regression model at predicting session-level codes with average AUC scores of 0.79, and 0.70, respectively. For fine-grained level coding, L-LDA and logistic regression are able to identify specific talk-turns representative of symptom codes. However, model performance for talk-turn identification is not yet as reliable as human coders. We conclude that the L-LDA model has the potential to be an objective, scalable method for accurate automated coding of psychotherapy sessions that perform better than comparable discriminative methods at session-level coding and can also predict fine-grained codes.**

***Index Terms*—Clinical communication, conversation analysis, labeled latent Dirichlet allocation (L-LDA), machine learning, multilabel document classification.**

I. Introduction

Across medical specialties, the basic medium of information gathering and intervention between the provider (i.e., MD, psychologist, nurse) and patient is a conversation. The patient describes problems and the provider listens, asks questions, and recommends solutions and specific treatment strategies. The content of this conversation can be useful across a broad variety of contexts, such as helping a primary care provider to detect and prevent suicide [1], promoting patient adherence to treatment recommendations [2], reducing cold severity and duration [3], and predicting a surgeon's history of malpractice lawsuits [4].

Psychotherapy (sometimes called counseling or behavioral treatment) represents a particular class of interventions that has a special focus on the provider–patient interaction. With psychotherapy, the interaction contains the treatment's active ingredients rather than simply being a means of developing rapport and forming a diagnosis. Psychotherapy ranges from brief, single session interventions [5] to multisession interventions over weeks or months [6] and research suggests that psychotherapy is effective for a broad range of mental health disorders [7].

The typical method of summarizing the content of this conversation is based on the provider's recollection and self-report of what happened as they record it in the medical record. Many methods exist for obtaining summary measures from transcribed text—e.g., by separating a transcript into broad semantic topics [8]–[14], detailed behavioral features (such as requests for clarification [15]) or syntactic parts of speech [16], among others. These summary measures can be used as context to extract and evaluate treatment information, including patient diagnosis, analysis of client communication, and evaluation of suicide risk [17]–[22].

At present, the evaluation of psychotherapy sessions and other types of patient–provider communication relies on human raters who summarize sessions by attaching codes (also called labeling or annotating) in order to quantify the information in treatment encounters [23]. The process of attaching these codes, called observational coding, provides theory-driven organizational systems through which complex linguistic data can be structured

Manuscript received August 6, 2015; revised October 23, 2015; accepted November 21, 2015. Date of publication November 25, 2015; date of current version March 3, 2017. This work was supported by the National Institutes of Health/National Institute on Alcohol Abuse and Alcoholism (NIAAA) under Award R01/AA018673 (D. C. Atkins/M. Steyvers, co-PIs) and National Institute on Drug Abuse under Award R34/DA034860. This paper was presented at ABCT 2014, 48th Annual Convention of the Association for Behavioral and Cognitive Therapies, Philadelphia, PA, USA, November 21–24, 2014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

G. Gaut and M. Steyvers are with the Department of Cognitive Science, University of California Irvine, Irvine, CA 92697-5100 USA (e-mail: gbius06@gmail.com; mark.steyvers@uci.edu).

Z. E. Imel is with the Department of Educational Psychology, University of Utah, Salt Lake City, UT 84112 USA (e-mail: zac.imel@utah.edu).

D. C. Atkins is with the Department of Psychiatry and Behavioral Science, University of Washington, Seattle, WA 98105 USA (e-mail: datkins@u.washington.edu).

P. Smyth is with the Department of Computer Science, University of California, Irvine, CA 92697 USA (e-mail: padhraics@gmail.com).

This paper has supplemental material available at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2015.2503985

for further analysis. Codes can represent the subject of conversation (e.g., medications, spousal relationships), symptoms expressed (e.g., depression, anxiety, anger), or specific verbal behaviors in individual utterances or talk-turns for providers (e.g., open or closed questions by the therapist, degree of empathy) or patients (e.g., signaling intent to change or maintain behavior).

Observational coding has critical shortcomings, including intensive labor requirements, coder error, nonstandardized coding systems (new codes require new training), and inability to scale up to larger coding projects [12]. Each hour of therapy takes roughly 10 h to code and the number of alcohol and drug abuse sessions in the U.S. healthcare system alone runs into the hundreds of thousands per year. The burden of human coding leads typical psychotherapy research studies to be small, which contributes to the incredible heterogeneity across studies investigating the relationships between therapist behavior and patient outcome [24]. Accordingly, human-based coding is not a feasible method for evaluating the content of treatment encounters on a large scale. An objective, scalable method for summarizing the content of actual treatment encounters is needed.

We can describe the implementation of coding systems for text as multiple-label classification problems, where multiple codes are attached to each document [25]. Machine learning approaches for automatic multiple-label document classification have been successfully used in various domains [26]–[30], including medical applications for disease diagnosis and medical error detection [31]–[33]. One such class of tools called topic models [34] has been used to assess the fidelity of therapist treatment [12] through prediction of behavioral codes, compare type of psychotherapy treatment [13], and predict therapy outcomes in schizophrenic patients [35], [36].

In this paper, we illustrate the ability of one specific type of topic model, labeled latent Dirichlet allocation (L-LDA) [11], [37], to semiautomatically infer subject and symptom codes from a large heterogeneous psychotherapy corpus; i.e., what topics and symptoms were discussed during treatment. Every session in the corpus was manually annotated with general discussion content and patient symptom codes such that the observable outcomes of the manual annotation process are codes for the session as a whole. However, implicit in the coding process is a fine-grained, or local, evidence-accumulation process, where each word, utterance or talk-turn in a session affects the decision to attach a given code. Establishing a link between specific within-session passages of text and overall codes for the session (session-level codes) is fundamental to understanding the coding procedure. We implement a model that, in addition to learning session-level coding systems, can localize specific passages of text representative of a session code. In other words, the model is able to infer codes at a local (talk-turn) level from codes that were provided at the global (session) level.

Previous work on computational analysis of psychotherapy transcripts used topic models to summarize therapy corpora and extract features for use in predictive models for therapy type [13] or as a stand-alone model to predict behavioral codes [12]. Our current work expands upon past research by using topic models to predict session content, by providing a detailed quantitative evaluation of predictive performance that includes comparisons to baseline models, and by developing methodology for talk-turn annotation using session-level metadata.

The model is evaluated and compared against a baseline discriminative model [lasso logistic regression (LLR)] using standard performance measure—the receiver operating curve and area under the curve (AUC). Additionally, we provide R-precision [38] scores for talk-turn prediction. Session-level R-precision scores can be found in the supplementary files. For all performance evaluation, we use tenfold cross validation at the session level to emphasize the models' ability to predict novel data.

As we will discuss in the experimental results section, the accuracy of the proposed techniques, in terms of code prediction, are not yet at the level of human annotators. Thus, these approaches are not yet ready to be used for fully automated annotation of therapy transcripts in an off-the-shelf manner. Nonetheless, as we outline in the discussion section later in the paper, the current techniques could potentially be used as components within a semiautomated approach, for example, to assist in therapist training, using the model to rank and present to a supervising therapist specific talk-turns within a trainee session in terms of the talk-turns likelihood of containing specific codes. There are multiple publicly available L-LDA software packages [39]–[41] that could be used to support such efforts. More broadly, the work described in this paper represents the next step toward a long-term goal of fully automatic code prediction for psychotherapy transcripts.

II. Data

The primary source of data comes from a psychotherapy corpus maintained by Alexander Street Press and made available via library subscription. At the time of the present analyses, the corpus contained 1181 therapy sessions with approximately eight million words. Each session was conducted with a unique therapist and client. On average each session contains 250 talk-turns, which are defined as uninterrupted passages of time during which either the patient or therapist speaks. Talk-turn length ranges from several words to several sentences. Sessions were conducted by prominent psychotherapists and serve as exemplars of different treatment approaches. Each session includes metadata such as patient age, patient gender, type of psychotherapy, and two types of nominal content codes (i.e., labels) referring to subjects discussed in the session (161 possible codes) and patient symptoms discussed in the session (48 possible codes). We use subject and symptom codes because we are interested in the relationship between language and the codes' semantic meanings (as opposed to codes for type of therapy, client gender, etc.). The list of symptom and subject codes was derived from the DSM-IV manual and other primary psychology/psychiatry texts. All codes annotated in the psychotherapy corpus are session-level codes, meaning that a single label is applied (as a binary present/absent label) to the entire session, and the original corpus did not include any labels for specific subunits such as sentences, talk-turns or paragraphs. Each session is annotated with multiple codes (min = 1 code, max = 17 codes) and the average session is annotated with approximately five codes. Prior to the analysis, we applied a number of preprocessing steps, including stop word removal and n-gram extraction to convert the original corpus into a form suitable for text analysis. We chose stop words from standard lists used in natural language processing and augmented these lists with words from the corpus that were not on standard stop word lists, but that contain little semantic content (e.g., "mm-hmm") (see Models section for details on preprocessing and supplementary files for full stop word lists). Stop words were removed from both patient and therapist speech. In the case of a talk-turn comprised completely of stop words, we removed the talk-turn from the data. The resulting representation of the text consisted of sparse vector counts of terms for each document, including unigrams (single words such as "medicine," "anger"), bigrams (e.g., "side effect"), and trigrams ("it sounds like").

In order to evaluate the ability of the model to find representative talk-turns, we conducted additional coding to generate labels for talk-turns within selected sessions. The aim of the additional coding was to generate data for specific within-session sections of text (in this case talk-turns) against which to test our model. These coded talk-turns were only used for model evaluation, not for model training. We focused on five symptom codes: anger, anxiety, depression, low self-esteem, and suicide. These codes were chosen first for their therapeutic importance and second for their high frequency of annotation in order to provide a sufficient amount of additional data. We restricted the number of symptom codes to limit the amount of human coding required for talk-turn annotation. For each of these symptoms, we randomly selected 200 client talk-turns of at least 50 characters in length (before stop word removal) from sessions that had the symptom code attached. On average, the selected talk-turns were approximately 277 characters in length before stop word removal. The process led to a total of 993 talk-turns. Each talk-turn was rated in terms of the representativeness of the symptom on a scale of 1 (atypical) to 7 (very typical) by each of six graduate students or postdoctoral fellows with training in clinical/counseling psychology.

III. Models

We approach the problems of session coding and identifying representative talk-turns through the use of L-LDA [11], [37], a semi-supervised extension of Latent Dirichlet Allocation (LDA). We first present the LDA model and then the L-LDA model. The model presentation is aimed at readers who have some experience with topic models. For readers new to topic modeling, we recommend reading a tutorial introduction [42]. Then, we show how these models can be used for document classification and how to apply the models to predicting codes and talk-turns in the general psychotherapy corpus. Finally, we present LLR as a baseline model against which to compare L-LDA.

A. Latent Dirichlet Allocation

LDA is an unsupervised modeling approach that learns a set of latent topics across a corpus of text. As opposed to L-LDA, there are no labels that are part of the data to learn from. The only data provided to LDA are a set of documents, where documents are treated as a "bag-of-words"; i.e., sparse vectors of word counts for each document. Thus, the order of words is not relevant for the model. We use both individual words and multiword terms (n-grams) in the vocabulary for our model—but for simplicity will refer to both as "words."

Standard applications of topic models assume that the text corpus can be naturally divided into documents. For example, a corpus of scientific articles is naturally divided into documents according to paper. In the case of spoken dialogue, choosing a rule for partitioning a corpus into documents is less straightforward. Documents can be defined as sentences, paragraphs, entire sessions or through any type of feasible partitioning. As in past research [12], [13], for the General Psychotherapy Corpus, we define documents to be individual talk-turns (although other definitions are possible as well). Using talk-turns to define documents yields a larger set of documents with more localized word co-occurrences compared to defining documents at the session level. We have found in our experiments that these localized word co-occurrences tend to result in more specific topic-word distributions and improve classification performance.

LDA specifies a generative process for the creation of text documents. From this generative process, we learn a predictive model by reverse-engineering the process—i.e., learning the parameters most likely to have generated the data. In LDA, each document (in this case talk-turn) is represented as a mixture of topics, where each topic is defined as a multinomial distributions over words. The creation of each document begins by sampling a document-specific distribution over topics. To generate each word in the document, a topic is sampled from the document specific distribution over topics and a word is sampled from that topic. Formally, let T be the total number of topics in the model and V be the size of the vocabulary (number of unique words in the corpus). Then, we can specify the marginal distribution over words for a document d as

$$P(w) = \sum_{t=1}^{T} P(w|z_w = t)P(z_w = t|d)$$

where z_w indicates the topic from which word w was drawn, $P(w|z_w = t)$ is a V-dimensional distribution over words for topic t, and $P(z_w|d)$ is a T-dimensional distribution over topics for document d. To simplify notation, we will let $\boldsymbol{\phi}^{(t)} = P(w|z_w = t)$ represent the distribution over words for topic t and $\boldsymbol{\theta}^{(d)} = P(z_w|d)$ represent the distribution over topics for each document d.

LDA incorporates *a priori* knowledge about topics likely to occur in a document by placing a Dirichlet prior on the distribution over topics $\boldsymbol{\theta}^{(d)}$ for each document. The Dirichlet prior is the conjugate prior of the multinomial distribution and is used to express the prior probability of observing a topic in a given document before observing any data. The Dirichlet distribution is parameterized by the vector $(\alpha_1, \ldots, \alpha_T)$, where α_t can be interpreted as the prior observation count for the number of times topic t is sampled in a document before having observed any actual words from that document. Thus, we can view the distribution over topics for a document d as a sample from this group-level prior distribution over topics.

In a similar manner, LDA also incorporates prior information about which words are likely to occur in a given topic. LDA does this by placing another Dirichlet prior on the distribution over words $\boldsymbol{\phi}^{(t)}$ for each topic t. This second Dirichlet distribution is parameterized by the vector $(\beta_1, \ldots, \beta_V)$, where β_w represents the prior observation counts of word w before observing any documents. Here, we can interpret each topic as a sample from this group-level prior distribution over words. We follow the common practice of setting the Dirichlet parameters uniformly (i.e., $(\beta_1, \ldots, \beta_V) = (\beta, \ldots, \beta)$), which corresponds to the assumption that each word is equally likely *a priori*.

B. Labeled LDA Model

L-LDA is a semisupervised variant of LDA in which some topics are placed in correspondence with labels that can be associated with a document. Documents in the training phase are assumed to have been preassigned to a subset of labels from a larger lexicon of possible labels. In the context of the psychotherapy corpus, possible labels include symptom and content codes and L-LDA model infers a unique topic for each code. These topics are learned by restricting inference to only the word tokens in documents annotated with the topic's corresponding label. We use a separate unsupervised set of topics, called background topics, to account for words not associated with the known codes. These background topics allow the model to capture some of the linguistic variability in the data that is not directly related to subject and symptom codes. Without these background topics many words would have to be explained by the topics associated with the symptom and content codes, which would decrease the generalizability of those topics. During training of the L-LDA model, when sampling the topic for a word token in a document (as describe later), only topics that belong to labels associated with a document (including background labels) can be sampled. All other topics have zero probability of being expressed in the document.

Formally, let $T = T_c + T_b$ be the total number of topics. A subset of T_c topics are in one-to-one correspondence with the labels associated with documents. The remaining T_b topics capture background information. During the generative process, for each document d, we restrict the space of possible document mixtures by restricting the hyperparameters of the Dirichlet prior on $\boldsymbol{\theta}$ according to a binary topic assignment vector $\boldsymbol{\Lambda}^{(d)} = (\Lambda_1^{(d)}, \ldots \Lambda_T^{(d)})$. We define

$$\boldsymbol{\Lambda}_t^{(d)} = \begin{cases} 1 : (\text{code } t \text{ is attached to document } d) \text{ or } (t > T_c) \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad : \text{ otherwise.} \end{cases}$$

We then define the hyperparameters for document d as $\boldsymbol{\alpha}_d = (\alpha_{d1}, \ldots, \alpha_{dT}) = \boldsymbol{\Lambda}^{(d)} \times \boldsymbol{\alpha}$. Note that the only topics that can be expressed for a particular document are topics corresponding to a code associated with the document or background topics.

Letting D be the number of documents in the collection, the generative process of the L-LDA model can be described as follows:

1) For topic $t \in 1, \ldots, T$:
    a) sample a multinomial distribution over words $\boldsymbol{\phi}^{(t)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_V)$.
2) For document $d \in 1, \ldots, D$:
    a) use the labels associated with document d to set the hyperparameters $\boldsymbol{\alpha}_d = \boldsymbol{\Lambda}^{(d)} \times \boldsymbol{\alpha}$;
    b) sample a multinomial distribution over topics $\boldsymbol{\theta}^{(d)} \sim \text{Dirichlet}(\boldsymbol{\alpha}_d = (\alpha_{d1}, \ldots, \alpha_{dT}))$.
    c) For each term $i \in 1, \ldots, N_d$:
        i) sample a topic indicator $z_i \sim$ Categorical $(\boldsymbol{\theta}^{(d)})$;
        ii) sample a word token $w_i \sim$ Categorical $(\boldsymbol{\phi}^{(t=z_i)})$,

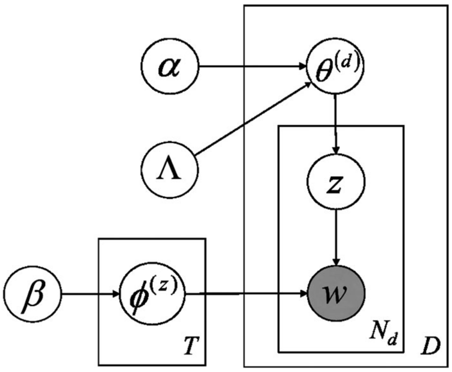

Fig. 1. Graphical model of L-LDA.

where N_d is the number of word tokens in document d. Note that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are hyperparameters for the model. The graphical model for L-LDA is presented in Fig. 1.

C. Training the L-LDA Model

The variables we would like to infer are the topic assignment variables z_w for each word w, the document mixtures $\boldsymbol{\theta}^{(d)}$, and the topic distributions $\boldsymbol{\phi}^{(t)}$. For sampling, we use a collapsed Gibbs sampler [43], which integrates out $\boldsymbol{\phi}^{(t)}$ and $\boldsymbol{\theta}^{(d)}$ so that we only sample the topic assignments z_w. The topic assignments z_w are then used to generate point estimates of $\boldsymbol{\phi}^{(t)}$ and $\boldsymbol{\theta}^{(d)}$.

The Gibbs sampling procedure considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the current topic assignments to all other word tokens. From this conditional distribution, we sample a topic assignment for the current word token. We write the conditional distribution as $P(z_i = t|\mathbf{z_{-i}}, w_i, d, \cdot)$, where $z_i = t$ represents the topic assignment of token i to topic t, $\mathbf{z_{-i}}$ refers to the topic assignments of all other word tokens, and "$\cdot$" refers to all other known or observed information such as all other word indices $\mathbf{w_{-i}}$, distributions over topics for all other documents, and hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The conditional distribution can be calculated as follows [43]:

$$P(z_i = t|\mathbf{z_{-i}}, w_i, d, \cdot) \propto \frac{C_{w_i t}^{VT} + \beta_{w_i}}{\sum_{w=1}^{V} (C_{wt}^{VT} + \beta_w)} \cdot \frac{C_{dt}^{DT} + \alpha_{dt}}{\sum_{j=1}^{T} (C_{dj}^{DT} + \alpha_{dj})} \qquad (1)$$

where t is restricted to the set of topics defined by the union of 1) codes t attached to document d, and 2) background topics $t > T_c$. All other topics have probability 0 for document d (as specified by the generative model) and are not eligible to be sampled. In the aforesaid equation, C^{VT} and C^{DT} are matrices of counts with dimensions $V \times T$ and $D \times T$, respectively; $C^{\mathrm{VT}}_{w_i t}$ contains the number of times word i occurred in a document with topic t and C^{DT}_{dt} contains the number of times a word token in document d was assigned to topic t. These matrices are incremented using the sampled topic assignment variables at each step of the Gibb's sampler for every word w.

The Gibbs sampling algorithm is initialized by assigning each word token in document d randomly to one of the set of eligible topics for document d (i.e., the codes t attached to document d or the background topics $t > T_c$). For each word token, the count matrices C^{VT} and C^{DT} are first decremented by one for the entries that correspond to the current topic assignment. Then, a new topic is sampled from the distribution in (1) and the count matrices C^{VT} and C^{DT} are incremented with the new topic assignment. Each Gibbs sample consists of the set of topic assignments for all N word tokens in the corpus, achieved by a single pass through all documents.

The sampling algorithm gives us samples for the topic assignment variables z_w for each word w. However, we are interested in estimating the word-topic distributions $\boldsymbol{\phi}^{(t)}$ and topic-document distributions $\boldsymbol{\theta}^{(d)}$. We can approximate the probability of choosing the kth word from the distribution over words for topic t, $\boldsymbol{\phi}^{(t)}$, using the word-topic count matrix (computed from the sampled topic assignment variables) as follows:

$$\hat{\phi}^{(t)}_k = \frac{C^{\mathrm{VT}}_{w_k t} + \beta_{w_k}}{\sum_{w=1}^{V} (C^{\mathrm{VT}}_{wt} + \beta_w)}.$$

Here, $\hat{\phi}^{(t)}_k$ can be interpreted as the estimated probability of choosing word w_k from topic t. We can also estimate the probability of choosing the tth topic from the distribution over topics for document d, $\boldsymbol{\theta}^{(d)}$, using the count matrix C^{DT} (also computed from the sampled topic assignment variables) as follows:

$$\hat{\theta}^{(d)}_t = \frac{C^{\mathrm{DT}}_{dt} + \alpha_{dt}}{\sum_{j=1}^{T} (C^{\mathrm{DT}}_{dj} + \alpha_{dj})}.$$

Here, $\hat{\theta}^{(d)}_t$ can be interpreted as the estimated probability of expressing topic t in document d. We later use $\hat{\boldsymbol{\phi}}^{(t)}$ to qualitatively examine topics corresponding to session codes and $\hat{\boldsymbol{\theta}}^{(d)}$ to estimate the topics (and therefore symptom and content codes) expressed in document d.

D. Prediction With the L-LDA Model

We evaluate the model by predicting labels for documents unseen by the model during training using the word-topic counts (C^{WT}) learned during training. The goal for prediction is to infer a document-topic count vector $C^{\mathrm{DT}}_{d't}$ for each new document d' , where the inferred count vector contains information about the likely topics (and associated codes) for d'.

For a new document d', we set $\Lambda^{(d')}_t = 1 \; \forall \, t \in \{1, \ldots, T\}$ so that any topic can be part of the document mixture. We run a Gibbs sampling procedure, where we compute the posterior distribution over topic assignments

$$\begin{aligned} P(z_i = \; & t|\mathbf{z}_{-\mathbf{i}}, w_i, d', \cdot) \\ & = \frac{C^{\mathrm{WT}}_{w_i t} + \beta_{w_i}}{\sum_{w=1}^{W} (C^{\mathrm{WT}}_{wt} + \beta_w)} \cdot \frac{C^{\mathrm{DT}}_{d't} + \alpha_t}{\sum_{j=1}^{T} (C^{\mathrm{DT}}_{d'j} + \alpha_j)} \end{aligned} \quad (2)$$

where $\alpha_t = \alpha$. The posterior for this sampling procedure is similar to the posterior used in the sampling procedure during training except that the word-topic count matrix C^{WT} is not updated. Holding C^{WT} constant formalizes the assumption that the word-topic counts are learned and that prediction consists of learning just the document-topic counts. Another difference from the sampling procedure used during training is that we sample the posterior probabilities $P(z_i = t|\mathbf{z}_{-\mathbf{i}}, w_i, d_i, \cdot)$ at each iteration (after burn-in) instead of the word-topic count assignments (that were sampled during training). While either word-topic counts or posterior probabilities can be used to compute prediction scores, we found that using posterior probabilities provided more accurate code predictions and required less samples for accurate prediction. We use the posterior samples to compute topic scores that represent the likelihood that a document should be annotated with the code corresponding to each topic. We compute a score $\eta_{t,d}$ for each topic t and test document d as follows:

$$\eta_{t,d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \gamma_{t,d,i}$$

where the variable $\gamma_{t,d,i}$ estimates the probability in (2) that the tth topic was assigned to the ith word token in document d. Thus, $\eta_{t,d}$ can be interpreted as the average probability of assigning a word from document d to topic t. To calculate each word's posterior estimate of topic assignment ($\gamma_{t,d,i}$), we average over the posterior samples of the probability of assigning word i to topic t. We compute $\gamma_{t,d,i}$ as follows:

$$\gamma_{t,d,i} = \frac{1}{K} \sum_{k=1}^{K} p(z_{t,d,i})^k$$

where $p(z_{t,d,i})^k$ is the kth sample of the posterior probability expressed in (2) and K is the total number of samples.

IV. Learning Topics From Labeled Sessions With L-LDA

A. Text Preprocessing

The original corpus contained eight million word tokens and 40 000 unique words. Before fitting the L-LDA model, we applied a number of preprocessing steps on the corpus. We first removed any words that occur five or fewer times in the entire corpus on the assumption that these words are unlikely to be useful in general for categorization. This step reduced the number of unique words from 40 000 to 27 000 unique words. After removing infrequent words, we removed words that we thought contained little semantic content. We performed a preliminary filtering using a common stop word lists to remove words (see unigram stop word list in supplementary files) Next, we applied a part-of-speech tagger [16] that we used to remove determiners, adverbs, pronouns, interjections, particles, modal words, punctuation, and numbers. We used part-of-speech tags to create additional stop word lists for bigrams and trigrams, and performed a second stop word filtering using these lists. A final stop word filtering was done for interjections that are common in psychotherapy, but were not identified by the part-of-speech tagger. See supplementary files for a full list of stop words. The final corpus contained 28 000 unique words (including generated bigrams and trigrams) and 1.4 million word tokens.

B. Model Parameters for Training L-LDA

The L-LDA model requires a number of decisions to be made and parameters to be selected before training the model, including the number of background topics T_b, the settings for the priors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the number of iterations, and the number of burn-in samples.

For the number of background topics T_b, we chose $T_b = 50$ for the results reported in this paper, and found that the model was not particularly sensitive to the number of background topics as long as T_b was at least 20. We used uniform $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where each element was set to $1/50$ and $1/100$, respectively. These are typical values used in training LDA models and we found that that the method was reasonably robust to small perturbations in these values. Our results are from a model using 100 training iterations, and 20 iterations for prediction, where the last $S = 10$ iterations are used for generating prediction scores. We ran several models that varied the number of iterations and burn-in samples and found results similar to the model we report.

C. Inferred Topics

Prior to assessing predictive performance measures, we qualitatively examined the topics generated by the L-LDA model (see Table I). We examined three types of topics corresponding to subjects, symptoms, and background content. For the subject and symptom labels, we illustrate the topics learned by the model for the five most common labels. For the background topics, we picked an illustrative set of five topics. Qualitatively, subject and symptom topics are very interpretable—e.g., the medications topic consists of examples of medications, words used to describe administration of medication, and words used to describe the effects of medication. The background topics shown in Table I also have intuitive interpretations and contain words that are not covered by the content codes in the psychotherapy corpus. For example, there are background topics that explain word usage related to people, jobs, and sleeping (background topics 9, 36, and 39, respectively). Without these background topics, the high-probability words associated with them would have to be redistributed over the content topics for subjects and symptoms, potentially decreasing their interpretability and predictive power.

TABLE I
THE MOST LIKELY TERMS INFERRED FOR THE TOPICS ASSOCIATED WITH THE FIVE MOST COMMON SUBJECT AND SYMPTOMS AND AN ILLUSTRATIVE SET OF BACKGROUND TOPICS

Subject	Inferred Topic Distribution
medications	medicine, mg, dose, wellbutrin, medicines, lamictal, prescription, effects, side_effects, ability
relationships	relationship, women, feels, friend, relationships, boyfriend, date, position, example, react
parent-child relations	mother, father, love, remember, relationship, parents, brother, emotional, loved, needed
depressive disorder	depression, medication, doctor, medicine, prozac, depressed, zoloft, generic, wellbutrin, add
spousal relationships	wife, marriage, married, husband, relationship, mhm, children, attitude, divorce, got_married
Symptom	**Inferred Topic Distribution**
anxiety	anxiety, anxious, panic, nervous, depression, worried, worst, fine, experience, helps
depression	depressed, depression, doctor, pain, die, needed, drugs, low, xanax, mg
anger	angry, feelings, anger, express, get_angry, be_angry, reaction, feels, pissed, 'm_feeling
low self-esteem	love, teaching, boyfriend, positive, stupid, attractive, fit, negative, sorta, criticism
irritability	annoyed, irritable, message, safe, dishes, cause, wife, skin, irritated, cats
Background	**Inferred Topic Distribution**
background 9	friends, family, mom, dad, close, sister, brother, daughter, men, lives
background 13	care, stop, took, weight, takes, ready, lose, take_care, amount, body
background 23	house, room, walk, bed, door, walking, rid, front, throw, clean
background 36	job, wants, work, business, works, office, busy, baby, buy, paper
background 39	morning, sleep, hours, Friday, sleeping, Monday, tomorrow, Saturday, wake, bed

For each topic, the ten most likely n-grams are shown.

V. SESSION-LEVEL PREDICTION

A. Cross-Validation and Scoring

To test generalizability of the model to new data, we use tenfold cross validation where for each fold the sessions are partitioned into two disjoint sets: 1) a training set with 90% of the sessions used to train an L-LDA model and 2) a validation set with the other 10% of sessions used to evaluate the trained model. We compute an AUC score for each validation set (for each code) and report the average of the AUCs across the ten validation sets.

To compute an AUC score for a topic corresponding to a particular code and a particular validation set we proceed as follows. For each session in a validation set, we predict scores at the talk-turn level (as described earlier) and then aggregate scores for all talk-turns in a session. We define the session likelihood score for session s and topic t as

$$\eta_{t,s} = \frac{1}{D_s} \sum_{d \in d^{(s)}} \eta_{t,d}$$

where D_s is the number of talk-turns in session s and $d^{(s)}$ is the set of all talk-turns (documents) in session s. For each topic t, using these scores, we rank the sessions in the validation set and compute the AUC using the known subject and symptom codes attached to each session.

B. Results for Session-Level Predictions

For each subject and symptom code, we computed the AUC for each cross-validation fold and took the average across folds to measure classification performance. Values of the AUC range in theory from 0.5 (chance level performance, e.g., randomly generated rankings) to 1 (perfect predictive accuracy). In practice, performance that is above the level of chance can occur even from models, where the scores are randomly distributed (and unrelated to the content of the sessions). This is especially the case with codes that occur infrequently. In order to assess the significance of the predictive accuracy of our model relative to chance performance, we calculated a set of AUC scores for each code using 1000 randomly generated rankings and computed the corresponding 90% confidence intervals.

Additionally, we compare the L-LDA model to a standard machine learning classifier, LLR. LLR is often used in classification settings, where the number of features is larger than the number of observations because of its ability to force feature weights to zero for uninformative features. For more details about LLR, see supplementary files.

The results of the AUC analysis are shown in Fig. 2. The widths of the 90% chance confidence intervals for each code correspond closely with the inverse of the code frequencies (lower frequencies are associated with larger chance confidence intervals). The L-LDA model showed higher predictive accuracy than the LLR model and both models performed significantly better than the chance model for a large number of codes. For the L-LDA model, the average AUC score over all codes is 0.789 (SD = 0.137) and average AUC for subject and symptom codes are 0.800 (SD = 0.131) and 0.753 (SD = 0.150), respectively.

All but ten of 209 codes had AUC scores above 0.5. The five codes with lowest AUC scores are gender roles, withdrawn, recollections, general pain, and self-fulfilling prophecy. The language associated with each of these codes contains a broad spectrum of variation that may have contributed to poor model performance. The five codes with highest AUC are hallucinogen abuse, drug addiction, alcohol dependence, passiveness, and attraction.

For the LLR model, average AUC score over all codes is 0.702 (SD = 0.145) and average AUC for subject and symptom codes are 0.713 (SD =0.146) and 0.667 (SD =0.137), respectively. There were 29 codes with AUC scores below 0.5. Overall, the LLR model performed significantly worse on average than L-LDA ($p < 0.001$ in a Wilcoxon sign test).

A common goal of document classification is to identify the relationships between specific classifiers and characteristics of data that lead to high-classification performance. Previous comparisons between L-LDA and discriminative models have shown that the L-LDA model can outperform discriminative models on low-frequency codes [11]. We analyzed this relationship on the general psychotherapy corpus and found only a weak correlation (R = 0.22) between the AUC difference for the two models and code frequency. This correlation showed that L-LDA model performs slightly better in comparison to LLR at predicting low-frequency codes than high-frequency codes. Posthoc qualitative analysis suggests that highly predictable codes contain unique language that facilitate prediction. For example, sessions that discuss hallucinogen abuse and drug addiction contain a range of drug-specific terms that are highly specific. Conversely, we expect that hard to predict codes, such as gender roles, are attached to sessions containing a broad spectrum of language.

VI. Talk-Turn Prediction

A. L-LDA Talk-Turn Prediction

As a second test of performance, we assessed the ability of the L-LDA model to find talk-turns that are representative of a session-level code. This comparison is novel in that the L-LDA model is trained using only session-level codes, but can then generalize the topics learned to identify representative talk-turns within each session. The evaluation procedure tests the models' abilities to distinguish the most representative talk-turns (as judged by human raters) from all other talk-turns.

We had six human coders generate ratings at the talk-turn level for 993 talk-turns using five symptom codes chosen from the set of general psychotherapy codes. The codes used were *anger, anxiety, depression, low self-esteem, and suicidal behavior*. Each talk-turn was assigned a continuous rating from 1 (atypical) to 7 (very typical) by each of the six coders. To keep model performance measures on the same scale as the session-level performance measures, we converted the continuous human ratings to binary scores (thus allowing us to compute classification performance measures). To binarize ratings, we chose a rating threshold and considered any ratings above the threshold to be representative of a symptom and any ratings below to be not representative. While there are many ways of choosing this threshold, we chose the threshold such that the top $c\%$ of ratings would be considered representative. We computed performance for $c = \{5, 10, \text{ and } 20\}\%$ to emphasize the model's ability to predict highly representative talk-turns.

Since raters did not rate talk-turns for the other symptom codes in the psychotherapy corpus, we created a mapping from the more detailed labels in the psychotherapy corpus to the five selected symptom codes. The motivation behind creating these code mappings is that a single symptom code (e.g., depression) might be aptly described by multiple codes in the psychotherapy corpus (e.g., depression, depressive disorder, hopelessness, ...). To create the mappings, we had a clinical psychologist mark that codes from the general psychotherapy corpus are related to each of the five symptom codes. See Appendix for more detail.

In addition to AUC scores, we report the R-precision. R-precision is a measurement of precision at the threshold at which precision is equal to recall. To generate model predictions, AUC scores, and R -precisions at the talk-turn level we proceeded as follows:

1) An L-LDA model was trained on each of the ten training data sets used for session-level cross-validation. For each training set, any session that contained any of the coded talk-turns was removed (making the prediction problem somewhat more difficult by not allowing the model access to the coded talk-turn nor any other talk-turns from the same session). We remove these talk-turns to avoid optimistic performance results since in application the model would be identifying codes for talk-turns from novel sessions.
2) Each of the ten trained models made predictions on the 993 labeled talk-turns. The $\eta_{t,d}$ scores were computed for each general psychotherapy code t and each talk-turn (document) d as described earlier, using each model's word-topic count matrix. To compute a score for a

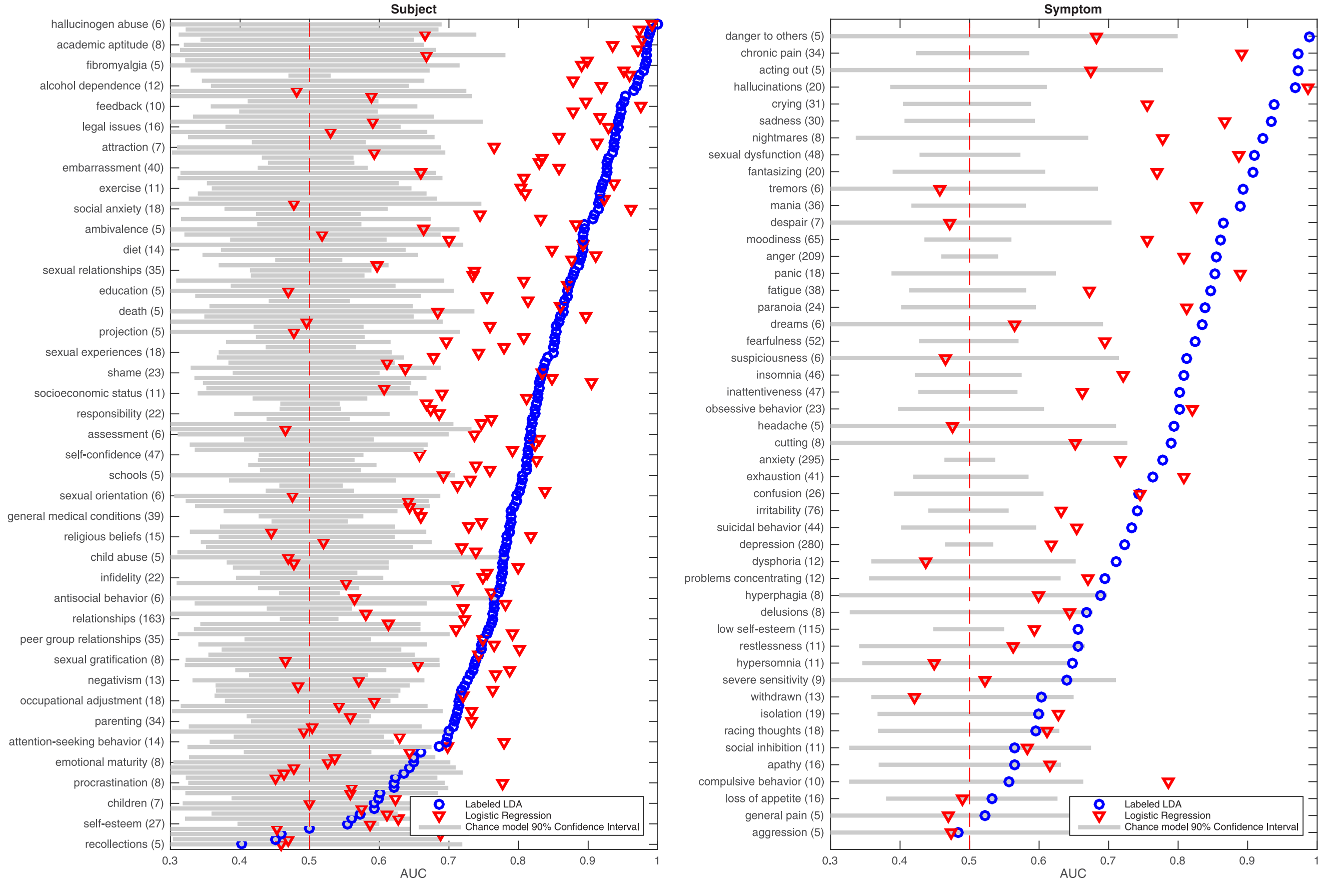

Fig. 2. Session-level AUC scores for the labeled topic model, the LLR model, and chance performance. Codes are reported along the *y*-axis and are ordered by labeled topic model performance. For subject codes, one in every four codes names is shown.

symptom code, we averaged the model scores from each of the related general psychotherapy codes (as defined by the code mapping described earlier). The ten scores for each code and each talk-turn were then averaged across the ten trained models.

3) For each code, AUC scores were generated as follows. The 993 talk-turns were ranked by their averaged model-based scores. These rankings were then compared to the ratings from each individual rater, where the ratings were binarized by using the highest 5%, 10%, top 20% of that individual's ratings, leading to three different AUC scores, one for each percentile cutoff. Overall AUC scores for the model, for each of the three cutoffs and each of the five codes, were then computed by averaging across the model's AUCs computed relative to each individual rater.

4) For each code, R-precisions were generated as follows. The 993 talk-turns were ranked by their averaged model-based scores. These rankings were then compared to the ratings from each individual rater, where the ratings were binarized by using the highest 5%, 10%, and 20% of that individual's ratings. For each rater and rating cutoff $c \in \{5, 10, 20\}$, we compute the R-precision as the number of true positives in the top $c\%$ of ratings divided by the number of talk-turns in the top $c\%$ of ratings. The R-precision ranges from 0 to 1 and it can be shown that the R-precision is equal to recall for the top $c\%$ of ratings. We compute overall R-precision scores as the average of model R-precision scores computed relative to each individual rater.

B. Results for Talk-Turn Predictions

Table II shows example talk-turns for all five symptoms tested. Talk-turns are ordered by model representativeness score. We also report the human representativeness rating (1–7 scale) averaged across raters. Several talk-turns illustrate that the model learns words associated with a symptom and not just the symptom keyword itself. For example, the first example talk-turn for depression in Table II is rated by the model as most representative and is also judged by humans as highly representative. This talk-turn does not contain the word depression but only expressions related to depression (i.e., "I am crying"). The first talk-turn for anxiety presents another interesting example. It is given the highest representativeness score by the model but only received a low human rating. The model may have learned to associate the word "roommate" with anxiety (through the other sessions in the training set), resulting in a high likelihood.

Again, we compare the L-LDA model to LLR. Table III shows a table of AUC scores for the L-LDA and LLR model with 5%, 10%, and 20% cutoffs used to create binary scores for the human representativeness ratings. For the L-LDA and LLR models, we computed an AUC score for the model relative to each individual rater and then averaged the AUCs across raters. To compare the models against human raters, we also calculated a human

TABLE II
MODEL PREDICTIONS FOR MOST REPRESENTATIVE TALK-TURNS FOR EACH SYMPTOM CODE

Symptom	Average Rating	Talk-Turn
Anger	6.2	Nobody every got angry; they never got angry. I don't ever remember my parents screaming at each other, ever. I mean throughout all my childhood I can't remember them having a yelling fight. It was never that way. and I just never knew how to scream at anybody.
	7	I have occasionally felt bad about the things I've told you about, I have. But it's interesting that this is the first time that a lot of anger has come out. You know, there's another side that I really affirm here, that there's a lot of anger that I have toward her that she's always been able to seem to get out and express at me.
	6.8	I don't know. But I didn't get mad at Harold when he gave me genital warts. I felt mad, I mean, I felt betrayed and lied to and cheated on, but I didn't - I just dealt with it, I just deal with things, and I've always thought that that was a positive quality, I mean, I just-i don't think that anger is necessarily productive. But I guess in some ways it can be. I just-I work through things, I talk through things, I'm calm, I don't get mad or yell and scream. If i-you know, I can argue with people if I don't - you know, it's not like I won't express my opinions or, you know, talk about something that bothers me. But I don't yell and scream and I don't get angry.
	7	At night, and then I take my zyprexa and I fall asleep in two hours. The one thing I'd say I notice about her is she will be talking like this and then all of a sudden I don't what happens, something happens and she just gets real angry, real fast, like that. We will be talking and all of sudden she will think of something that got her angry and it will be like boom.
	3.6	Even just now, when you ask me that, I don't know, it just feels like, why are you asking me these questions? I don't understand them. I feel like ... it's just really uncomfortable.
Anxiety	2	The only-the only thing I can-like I thought back to this. When I was a senior in college I met a girl who was a roommate of my roo-well, my roommate's - my roommate and I-my roommate had his fiance and she had a roommate. And this, anyhow, to make it all work ....
	3.2	When I got there, he said that I needed to go to the hospital. So I went, he sent me to ... when I got to ..., they did another ekg. They told me I had a heart attack 2 weeks before that.
	7	And, um, had a little anxiety about it. I go 2 nights a week, monday and wednesday from 6 to 10, and yeah, had a little anxiety attack about it'cause just the whole like possible failure, and like oh god, I'm like I really want, 'cause I really want it and I'm really, you know, I'm good at it, but it's like oh god, it's pressure, you know, that type of thing. Well, I ended up calling ... remember I was seeing him?
	5	I don't really know. I've always been kind of just like - I'm always just really scared of - I don't really have like a lot of, in my family there's not really a lot of people that would help me if something like that was to happen. So I think that just kind of like fuels this like fear in me with like employment in general. It's just kind of like, 'well what if there is a cutback or what if somebody buys us out or ... ' just kind of like, I just want to be okay if that happened.
	2.2	Yeah. I think I just ... I never ... I don't know. Since probably before I came to shimer was the last time I actually like really either showed interest in a guy. Like even if I was interested, I haven't within the past three years, like done anything about it really. Brendan, close, like we've actually kissed and ... but like ... that was.
Depression	6	As like two to three months ago, I was crying and this was more or less yknow I didn't want to be doing this but now I'm crying, I'm like, I don't care that I'm crying, yknow.
	5	Well I think one of the things I wanted to ask you about was what we talked about last week the matter of guilt when I touched on that briefly. I'm a little bit confused because it seems to me that a person has desires to kind of change their ways as it were that one of the motives of them wanting to do that is them some feeling of guilt or something approximating guilt about the way they're presently acting. And yet you said that you thought that I should feel that way, people in general too, but in this particular case me should not feel guilty for example about vanity because I've done that. so you think that they should feel that way but it seems to me that one of the motivating forces for me is a certain sense of guilt. or not exactly guilt but maybe something like ... well I suppose it is guilt the guilt of throwing away a good part of my life. I feel guilty about that even in a moral sense as well as a practical one. so what do you mean by that? how do you work around ... .
	6.6	Um, the pamelor 50, I take that at night, that seems to be doing okay. I mean I'm still a little depressed but, you know, basically that seems to be doing okay. nr : doctor, patients are leaving.
	3.6	It might be. I guess I feel the anger because it, like a given situation turned out unhappy or sad instead of happy.
	3.6	And yknow, be supportive for my nephews and my nieces and I found myself kind of leaning with my dad in just being sad, just being, yknow, it was just different.
Low Self-Esteem	5	Umm ... well certainly if I'm not being obsessive and worrying over things. because the truth is, it's obvious when I'm in that state, even if I don't tell people. you can see it all over my face. and he notices. and so if I'm not that way, and if I'm confident and mature.
	4.2	Which probably isn't a good thing. but I just don't ' do it. and they tell me," you should do it. you should do it. you should do it. ". and I say, notes don't work for me. I make up excuses. I tell them that, no, that's not me. leave me alone. let me do my own thing. and that's sort of me not taking criticism well.
	3.4	It's funny, because when I talk about my relationship with my parents with cole it's always that I'll say something negative and I'll say, "but I know they love me, but-but-but-but," you know? like I'll always have an aside for "yeah, like, but it's okay because ... " right, I guess I feel like it's not okay to be. I know it is okay to be angry at times. and of course now I'm thinking, but I don't blame them, they're who they are, you know? yeah, I always have to have an excuse for them. but -.
	2.6	I think I mentioned it a little bit earlier but like, I was talking to my friend about it. I mean he - we were sort of sitting in the lunchroom and he saw a girl that he thought was attractive walk by and he was like, "wow, she's a really attractive girl. ". and I was like, "eh. ". he's like "what, you don't think she is? ". and I was like, "yeah, but who cares? ". that's just sort of been my sort of feeling lately, that it's like yeah, she's an attractive girl. whatever.
	2.6	No.. if someone is looking yeah because well first of all one of the answers that I think that I can't find is like this question of what is a man looking for?

Talk-turns are ordered by model score Average human Likert rating (1–7) is reported to compare model scoring versus human scoring.

reliability score to serve as a measure of interannotator agreement. These scores give us an upper bound on performance against which to compare our model (assuming that human raters are performing optimally). To calculate this reliability score, we compared each individual rater against each of the other raters by computing pairwise AUC scores. We express human-reliability as AUC scores so that L-LDA performance and human reliability are expressed in the same units and fair comparisons can be made. For an individual rater, we calculate AUCs using the ratings of the individual rater (analogous to the

TABLE III
TALK-TURN CODING PERFORMANCE FOR THE L-LDA MODEL AND HUMAN RELIABILITY SCORES

		AUC at 5%			AUC at 10%			AUC at 20%		
Code	No. Talk-Turns	L-LDA	Human	L1 LR	L-LDA	Human	L1 LR	L-LDA	Human	L1 LR
anger	197	0.89	0.94	0.91	0.84	0.94	0.85	0.75	0.87	0.73
anxiety	200	0.76	0.80	0.77	0.72	0.78	0.72	0.66	0.72	0.64
depression	198	0.73	0.87	0.74	0.67	0.84	0.73	0.66	0.84	0.70
low self-esteem	200	0.70	0.82	0.74	0.64	0.81	0.68	0.60	0.78	0.63
suicidal behavior	198	0.78	0.96	0.77	0.70	0.87	0.70	0.67	0.82	0.67
average		0.77	0.88	0.79	0.71	0.85	0.73	0.67	0.81	0.67
		R Precision at 5%			R Precision at 10%			R Precision at 20%		
Code	No. Talk-Turns	L-LDA	Human	L1 LR	L-LDA	Human	L1 LR	L-LDA	Human	L1 LR
anger	197	0.57	0.58	0.67	0.56	0.72	0.59	0.44	0.68	0.41
anxiety	200	0.22	0.33	0.26	0.23	0.43	0.19	0.31	0.46	0.28
depression	198	0.10	0.39	0.26	0.23	0.53	0.24	0.31	0.56	0.30
low self-esteem	200	0.08	0.36	0.14	0.11	0.44	0.14	0.25	0.45	0.25
suicidal behavior	198	0.42	0.78	0.12	0.34	0.64	0.13	0.38	0.60	0.178

AUC and R-precision scores are shown for the top 5%, 10%, and 20% of talk-turns as rated by human coders. Human reliability is expressed in AUC and R-precision scores to enable direct comparison to model performance.

model scores) to predict the binarized ratings (at 5%, 10%, and 20% cutoffs) of each of the other raters. We compute human reliability as the average of all AUCs calculated from each pair of raters and report the computed scores in Table III. To compute human reliability in terms of R-precision, we perform an analogous computation using R-precision instead of AUC.

The table shows that both L-LDA and LLR perform well at identifying representative talk-turns relative to human reliability. On average, L-LDA AUC scores are between 10–18% lower than average inter-rater AUC scores. The L-LDA model performs distinctly better at identifying talk-turns representative of anger than talk-turns representative of the other tested symptoms. The unique lexicon of words used to express anger may influence the model's performance. In addition, the other four symptoms may be expressed in a broader language that is more difficult to capture through uni-, bi-, and trigrams. In addition to variation in performance by symptom, the model performs better when identifying the top 5% of representative talk-turns as compared to the top 10%. Therefore, the model is able to identify the most relevant talk-turns in a session with reasonable precision. The comparison between L-LDA and the baseline model shows that the LLR model performs about the same or marginally better than the L-LDA model on each of the three cutoffs ($p = 0.28$, $p = 0.11$, $p = 0.78$, respectively, in pairwise t-tests).

VII. DISCUSSION AND CONCLUSION

In this study, we have presented the L-LDA Model as a method for the semiautomatic code annotation of psychotherapy sessions. L-LDA outperforms standard discriminative methods at identification of session-level codes, replicating results from prior psychotherapy process research and general applications in multidocument classification. In addition to session-level coding, machine-learning methods show promise for annotation of psychotherapy transcripts at fine-grained levels of detail, such as for talk-turn annotation. L-LDA and LLR can identify talk-turns representative of session-level codes with accuracy close to that of trained human coders.

Machine learning methods for document classification often focus either on topic-based classification involving large documents and many topics, or sentiment classification involving a small set of sentiment labels and often shorter documents [28]. Our work involves both topic-based classification (for session level prediction) and analysis more similar to sentiment classification (talk-turn prediction for a small set of class labels). The generative nature of L-LDA provides a natural bridge between these two types of document classification problems by inferring labels for talk-turns based on session-level metadata. Topic-based classification is performed by integrating topic information over constituent parts of a document (in our case talk-turns), and sentiment classification is performed using a mapping between topic-based class labels to sentiment labels. In this way, L-LDA provides richer information than many sentiment classification methods and more flexibility than some topic-based classification models. Examining the relationships between the mapping from topic-based classes to sentiment classes is an interest for future work and we suspect that incorporating this information will lead to improved predictive performance.

Promising results in annotation of psychotherapy transcripts suggest potential for application to clinical settings in addition to reducing labor costs and improving the scalability of observational coding. For example, in the process of training junior therapists, supervising therapists review records of the junior therapist's sessions. Supervising therapists are often in charge of many junior therapists and are in need of tools that make the review process more efficient. One method for making this process more efficient would be to use text-based models that predict important topics discussed in the sessions (such as depression, suicide, etc.). The supervisor can get a quick summary of session content and can locate specific passages in the session by content labels. Additionally, the supervisor can provide feedback to the model on which passages were relevant to that topic and, thus, improve future code annotation.

L-LDA is a model for the semantics of language that, like all models, provides an approximation to the true underlying process of generating speech to convey meaning. L-LDA makes several simplifying assumptions about the process of text generation that could provide starting points for further model development. The "bag-of-words" assumption disregards information about temporal characteristics of language and their relation to semantics. L-LDA also ignores syntactic dependencies. An important direction for semantic analysis of psychotherapy sessions would be to incorporate sequential information and context into our analysis. This would involve significant feature engineering, but could benefit from already existing text processing techniques such as word and sentence embedding.

The work presented earlier analyzes the relationship between semantic information contained in spoken language and subjects and symptoms that encompass not just semantics, but emotion, and behavior. To gain a deeper understanding of psychotherapy, semantic language models need to be extended to encompass behavior. Considerable information is contained in behavioral cues such as tone, laughter, or body language that encompass the semantic meaning of a statement. While these behavioral cues are most likely correlated with language, we think that jointly analyzing behavior and language will lead to deeper understanding of the psychotherapy process and its effect on patient outcome.

In conclusion, we used data from the patient–provider interactions in psychotherapy to illustrate the potential of machine learning methods to automate coding of key aspects of clinical conversation and to understand the linguistic processes behind psychotherapy. L-LDA is a robust automated coding method that outperforms a baseline logistic regression discriminative method at predicting codes at the session level and that can be used to localize information using only session-level metadata.

Appendix
Codes Used in Talk-Turn Prediction

Several of the symptoms for which we performed additional local coding are closely associated with more than a single code in the psychotherapy corpus (e.g., the anger symptom is closely associated with anger and frustration). The human raters who judged the representativeness of the five symptoms were unaware of the variety of content codes used in the psychotherapy corpus and, therefore, the rater's concept of suicide might not map onto the (narrow) concept of suicide in the psychotherapy corpus. We, therefore, had a clinical psychologist create associated code sets by selecting from the list of psychotherapy symptom codes (See Table IV) with the constraint that the code set contain the matching code term. These meta code sets were created prior to any evaluation of the model.

For each of the five symptoms in the ratings experiment, we take the set of codes from the psychotherapy corpus that are closely associated with the symptom (e.g., for the anger suicide, we take the set anger and frustration) and average the model predictions across the codes in the set. This creates a model representativeness score for each talk-turn in the ratings experiment that can be compared to the binarized human ratings (highly representative/not highly representative). This approach of combining predictions among closely related labels can be viewed as a simple implementation of the idea that labels in multilabel document classification are often dependent and leveraging such dependences is worthwhile and can improve predictive performance [25].

TABLE IV
Symptom Codes and Sets of Associated Codes

Symptom Code	Code Set
anger	anger, frustration
anxiety	anxiety, fear, nervousness, social anxiety, stress, death anxiety, fearfulness, panic, paranoia, restlessness
depression	depression, grief, guilt, hopelessness, loneliness, shame, crying, depressive disorder, despair, dysphoria, loss of appetite, problems concentrating, sadness, suicidal behavior, withdrawn
low self-esteem	low self-esteem, self-esteem
suicidal behavior	hospitalization, suicide, cutting, dysphoria, death

References

[1] M. D. Feldman, P. Franks, P. R. Duberstein, S. Vannoy, R. Epstein, and R. L. Kravitz, "Let's not talk about it: Suicide inquiry in primary care," *Ann. Fam. Med.*, vol. 5, no. 5, pp. 412–418, 2007.

[2] C. E. Golin, H. Liu, R. D. Hays, L. G. Miller, K. Beck, J. Ickovics, A. H. Kaplan, and N. S. Wenger, "A prospective study of predictors of adherence to combination antiretroviral medication," *J. Gen. Intern. Med.*, vol. 17, no. 10, pp. 756–765, 2002.

[3] D. Rakel, B. Barrett, Z. Zhang, T. Hoeft, B. Chewning, L. Marchand, and J. Scheder, "Perception of empathy in the therapeutic encounter: Effects on the common cold," *Patient Educ. Couns.*, vol. 85, no. 3, pp. 390–397, 2011

[4] N. Ambady, D. LaPlante, T. Nguyen, R. Rosenthal, N. Chaumeton, and W. Levinson, "Surgeons' tone of voice: A clue to malpractice history," *Surgery*, vol. 132, no. 1, pp. 5-9. 2002.

[5] W. R. Miller and S. Rollnick, *Motivational Interviewing: Preparing People for Change*. 3rd ed. New York, NY, USA: Guilford, 2002.

[6] A. Christensen, D. C. Atkins, S. Berns, J. Wheeler, D. H. Baucom, and L. E. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *J. Consult. Clin. Psychol.*, vol. 72, no. 2, pp. 176–191, Apr. 2004.

[7] United States. Public Health Service. Office of the Surgeon General. (1999). Surgeon general's report [Online]. Available: http://profiles.nlm.nih.gov/ps/access/NNBBHS.pdf.

[8] M. Purver, "Topic segmentation," in *Spoken Language Understanding: Systems for Extracting Semantic Information From Speech*. Hoboken, NJ, USA: Wiley, 2011, pp. 291–317.

[9] M. Dowman, V. Savova, T. L. Griffiths, K. P. Kording, J. B. Tenenbaum, and M. Purver, "A probabilistic model of meetings that combines words and discourse features," *IEEE Trans. Audio. Speech Lang. Process.*, vol 16, no. 7, pp. 1238–1248, Sep. 2008.

[10] C. Howes, M. Purver, and R. McCabe. (2013). Investigating topic modeling for therapy dialogue analysis. presented at the IWCS [Online]. Available: http://www.ling.uni-potsdam.de/iwcs2013/Papers/CSCT-4.pdf

[11] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Mach. Learn.*, vol. 88, nos. 1/2, pp. 157–208, 2012.

[12] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification," *Implement Sci.*, vol. 9, no. 49, pp. 1–11, 2014.

[13] Z. E. Imel, M. Steyvers, and D. C. Atkins. (2014). Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions. presented at the Psychotherapy [Online]. Available: http://www.psiexp.ss.uci.edu/research/papers/Imel_Atkins_Steyvers_2014.pdf.

[14] M. Mitchell, K. Hollingshead, and G. Coppersmith. (2015). Quantifying the language of schizophrenia in social media. presented at the CLPsych [Online]. Available: http://m-mitchell.com/clpsych2015/pdf/CLPsych02.pdf

[15] M. R. J. Purver, "The theory and use of clarification requests in dialogue," Ph.D. dissertation, Dept. Comput. Sci., KCL, London, U.K., 2004.

[16] T. Kristina and C. D. Manning. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. presented at the EMNLP/VLC [Online]. Available: http://www-nlp.stanford.edu/cmanning/papers/emnlp2000.pdf.

[17] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, "Diagnosis code assignment: models and evaluation metrics," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 2, pp. 231–237, 2014.

[18] E. Mayfield, M. B. Laws, I. B. Wilson, and C. Rose, "Automating annotation of information-giving for analysis of clinical conversation," *J. Am. Med. Inform. Assoc.*, vol. 21, no. e1, pp. e122–e128, 2014.

[19] S. Iver, R. Harpaz, P. LePendu, A. Bauer-Mehran, and N. H. Shah, "Mining clinical text for signals of adverse drug-drug interactions," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 2, pp. 353–362, 2014.

[20] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–551, 2011.

[21] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner, "Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 540–543, 2011.

[22] C. Poulin, B. Shiner, P. Thompson, L. Vepstas, Y. Young-Xu, B. Goertzel, B. Watss, L. Flashman, and T. McAllister, "Predicting the risk of suicide by analyzing the text of clinical notes," *PloS One*, vol. 9, no. 1, art. no. e85733, 2014.

[23] Z. E. Imel, S. A. Baldwin, J. S. Baer, B. Hartzler, C. Dunn, D. B. Rosengren, and D. C. Atkins, "Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients," *J. Consult. Clin. Psychol.*, vol. 82, no. 3, pp. 472. 2014.

[24] C. A. Webb, R. J. DeRubeis, and J. P. Barber, "Therapist adherence/competence and treatment outcome: A meta-analytic review," *J. Consult. Clin. Psychol.*, vol. 78, pp. 200–211, 2010.

[25] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," in *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: Inf. Sci. Ref., 2007, ch. 6, pp. 64–74.

[26] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[27] G. V. Cormack, "Email spam filtering: A systematic review," *Found. Trends Informat. Retrieval*, vol 1, no. 4, pp. 335–455, 2007.

[28] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Informat. Retrieval*, vol. 2, nos. 1/2, pp. 1–135, 2008.

[29] S. Dumais and H. Chen. (2000). Hierarchical classification of web content. presented at the 23rd ACM SIGIR [Online]. Available: http://www.msr-waypoint.com/en-us/um/people/sdumais/sigir00.pdf.

[30] D. Billsus and M. Pazzani. (1999). A hybrid user model for news story classification. presented at the UM. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.5846&rep= rep1&type=pdf.

[31] Q. Li, K. Melton, T. Lingren, E. S. Kirkendall, E. Hall, H. Zhai, Y. Ni, M. Kaiser, L. Stoutenborough, and I. Solti, "Phenotyping for patient safety: Algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 776–84, 2014.

[32] Y. Ye, F. R. Tsui, M. Wagner, J. U. Espino, and Q. Li, "Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 871–875, 2014.

[33] B. J. Marafino, J. M. Davies, N. S. Bardach , M. L. Dean, R. A. Dudley, and J. Boscardin, "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 815–823, 2014.

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[35] C. Howes, M. Purver, R. McCabe, P. G. Healey, and M. Lavelle. (2012). Predicting adherence to treatment for schizophrenia from dialogue transcripts. presented at SIGDIAL [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.361.9153&rep= rep1&type =pdf#page=97.

[36] C. Howes, M. Purver, and R. McCabe, "Using conversation topics for predicting therapy outcomes in schizophrenia," *Biomed. Inform. Insights*, vol. 6, no. suppl. 1, pp. 39–50, 2013.

[37] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. presented at EMNLP [Online]. Available: http://ai.stanford.edu/nmramesh/emnlp09.pdf.

[38] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[39] D. Ramage and E. Rosen. (2009). Stanford Topic Modeling Toolbox [computer software] [Online]. Available: http://nlp.stanford.edu/software/tmt/tmt-0.4/.

[40] M. Ott. (2013). JGibbLabeledLDA. [computer software] [Online]. Available: https://github.com/myleott/JGibbLabeledLDA

[41] N. Shuyo. (2010). Labeled Latent Dirichlet Allocation. [computer software]. Cybozu Labs Inc. [Online]. Available: https://github.com/shuyo/iir/blob/master/lda/llda.py

[42] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook of Latent Semantic Analysis*. London, U.K.: Psychol. Press, 2007, ch. 21, pp. 424–440.

[43] T. Griffiths and M. Steyvers. (2004). Finding scientific topics. presented at the PNAS [Online]. Available: http://marketingpedia.com/Marketing-Library/Network%20Science/PNAS_2004%20Article/PNAS-2004-Griffiths-5228-35.pdf

Authors' photographs and biographies not available at the time of publication.