HHS Public Access

Author manuscript

Behav Res Methods. Author manuscript; available in PMC 2022 October 01.

Published in final edited form as:

Behav Res Methods. 2021 October ; 53(5): 2069–2082. doi:10.3758/s13428-020-01531-z.

How do you feel? Using Natural language processing to automatically rate emotion in psychotherapy

Michael J. Tanana[a], **Christina S. Soma**[b], **Patty B. Kuo**[b], **Nicolas M. Bertagnolli**[c], **Aaron Dembe**[b], **Brian T. Pace**[d], **Vivek Srikumar**[e], **David C. Atkins**[f], **Zac E. Imel**[b]

[a]University of Utah, Social Research Institute

[b]University of Utah, Department of Educational Psychology

[c]empathy.rocks

[d]Lyssn.io

[e]University of Utah, School of Computing

[f]University of Washington, Department of Psychiatry and Behavioral Sciences

Abstract

Emotional distress is a common reason for seeking psychotherapy, and sharing emotional material is central to the process of psychotherapy. However, systematic research examining patterns of emotional exchange that occur during psychotherapy sessions is often limited in scale. Traditional methods for identifying emotion in psychotherapy rely on labor-intensive observer ratings, client or therapist ratings obtained before or after sessions, or involve manually extracting ratings of emotion from session transcripts using dictionaries of positive and negative words that do not take the context of a sentence into account. However recent advances in technology in the area of Machine Learning algorithms, in particular Natural Language Processing, have made it possible for mental health researchers to identify sentiment, or emotion, in therapist-client interactions on a large scale that would be unattainable with more traditional methods. As an attempt to extend prior findings from Tanana et al (2016), we compared their previous sentiment model with a common dictionary-based psychotherapy model - LIWC - and new NLP model - BERT. We used the human ratings from a database of 97,497 utterances from psychotherapy to train the BERT model. Our findings revealed that the unigram sentiment model (kappa = 0.31) outperformed LIWC (kappa = 0.25), and ultimately BERT outperformed both models (kappa = 0.48).

Keywords

emotion; natural language processing; psychotherapy process; emotion coding; sentiment analysis

Psychotherapy involves goal-directed conversations where people are able to explore their emotions, experiences, and distress. For over a century, researchers and practitioners have consistently acknowledged the central role emotions play in psychotherapy (Freud & Breuer, 1895; Lane, Ryan, Nadel, & Greenberg, 2015). Emotion is directly involved in psychotherapeutic process and outcome, including the formation of the therapeutic alliance (Safran & Muran, 2000; Chui et al., 2016), client decision making (Bar-On, Tranel, Denburg, & Bechara, 2004; Isen, 2008), behavior change (Lang & Bradley, 2010), personality style (Mischel, 2013), and happiness (which of course, is an emotion; Diener, Scollon, & Lucas, 2009). More generally, emotion is implicated in human memory (Schacter, 1999), and its expression and perception are building blocks of empathy (Elliott, Bohart, Watson, & Greenberg, 2011; Zaki et al., 2008). The particular role of emotion in different psychotherapy theories varies from accessing and releasing suppressed emotions (as in Psychoanalysis; e.g. Kohut, 2013), to identifying the impact of cognitions on emotions (as in Cognitive Behavioral Therapy; Beck, Rush, & Shaw, 1979), to deepening understanding and accepting that emotions as a fundamental part of life (i.e., Emotion Focused Therapy; Greenberg, 2015). Researchers have employed various methods to examine the relationship between communication of emotions, therapy processes, and outcomes. Many of these methodologies focus on examining emotional valence, as processing and experiencing of both positive and negative affect is often a crucial component of therapy that spans different treatment modalities (Sloan & Kring, 2007).

Self-report ratings are often used to study client and therapist emotion (e.g., the Positive and Negative Affect Scale - PANAS - Crawford & Julie, 2004; Self-Assessment Manikin - SAM - Bradley & Lang, 1994). For example, the PANAS has been used to assess positive and negative affect of clients, which are important indicators of symptom improvement for clients with anxiety and depression (e.g. Krings, Persons, & Thomas, 2007; Watson, 2005). Similarly, researchers have used the SAM to assess for processing of traumatic memories in clients engaged in exposure based treatments (Sloan, Marx, & Epstein, 2005). Therapist detection of client emotional valence is important, as lower therapist empathic accuracy for positive emotions can predict greater symptom severity in following sessions (Atzil-Slonim et al, 2018). Similarly, researchers have also used observer ratings of different theorized emotional processes to detect positive and negative valenced emotions (e.g., see Kramer, Pascual-Leone, Despland, & Roten, 2015; Kramer, Pascual-Leone, Rohde, & Sachse, 2016; Luedke, Peluso, Diaz, Freund, & Baker, 2017). Across client self-report and observer measures of examining emotions, a recent meta-analysis found that both client and therapist emotional expression were associated with improved treatment outcomes (Peluso & Freund, 2018). Furthermore, researchers have utilized physiological measures such as skin conductivity (Messina et al., 2013) and vocal tone (Imel et al., 2014) as measures of emotional arousal in psychotherapy.

While self-report measures, behavioral coding, and physiological measures have been important methods in understanding degree, arousal, and valence of emotional expression, there are problems with these methods for assessing emotions: 1) Self-report measures are easy to obtain and allow access to internal private experiences, but rely on retrospective client or therapist insight (see Joormann & Stanton, 2016 for a review of emotion reporting for individuals with depression), which do not capture moment-to-moment fluctuations in

emotion during a session; 2) Observer ratings can be more objective, and can capture these finer-grained changes during a session, but require time-intensive coding (see Gottman et al., 2002 for an example of coding procedures); 3) Physiological measures have been historically impractical, and typically only capture magnitude of arousal rather than emotional valence (e.g., Juslin & Scherer, 2005; Imel et al., 2014). Thus, until very recently, the exploration of emotion in psychotherapy has been limited by the lack of methodology for examining emotion during sessions in direct and scalable ways. However, recent advances in machine learning and natural language processing provide potential solutions to facilitate coding of emotion during psychotherapy (Gonçalves, Araújo, Benevenuto, & Cha, 2013), which may better account for within-session valence that has otherwise been unexplored.

Natural Language Processing and Sentiment Analysis

Machine learning is a field of computer science that includes the process of creating algorithms such that computers are able to learn patterns of inputted data without being explicitly programmed with large collections of manually developed rules (see Samuel, 1962). Machine learning has provided innovative and critical methodologies to support various domains of mental health research (Aafjes-van Doorn, Kamsteeg, Bate, & Aafjes, 2020). For example, machine learning algorithms have been applied to session notes to assess treatment of post-traumatic stress disorder among veterans (Shiner et al, 2013). Furthermore, researchers have used network analysis to examine symptom clusters of individuals whose depression and anxiety symptoms relapsed or went into remission (Lorimer, Delgadillo, Kellett, & Brown 2019). Similarly, machine learning algorithms have been used to estimate alliance-outcome estimates for individual patients (Rubel, Zilcha-Mano, Giesemann, Prinz, & Lutz 2018). The emergence of research focusing on machine learning has laid the foundation to exploring different ways of examining and improving mental health care.

Natural Language Processing (NLP) is a subfield of machine learning whose goal is to computationally “learn, understand, and produce human language content” (Hirschberg & Manning, 2015, p. 261; Hladka & Holub, 2015). For example, researchers implemented automated speech analysis and machine learning methods to predict the onset of schizophrenia (Bedi et al, 2015), and produced language in the form of conversational dialogue (Vinyals & Le, 2015). NLP techniques have already been used to extract topics of conversation between therapists and clients (Atkins et al, 2012; Imel at al, 2015), and examine empathy of therapists (Xiao et al, 2015). Currently, a major focus in NLP is developing methods that correctly identify the emotion related phenomenon in passages using only the written words - often called sentiment analysis in computer science (for a review, see Pang & Lee, 2008). This field is broad, including classification of emojis (Read, 2005), tone of movie reviews (Socher, Pennington, Huang, Ng & Manning, 2011), and product reviews (Nasukawa & Yi, 2003).

A common approach before the more widespread usage of NLP techniques was to rely solely on hand-compiled lists of positive or negative words (e.g., texts with more positive words have more positive sentiment; Baccianella, Esuli, & Sebastiani, 2010; see Linguistic Inquiry and Word Count; LIWC; Pennebaker et al, 2003). Word count based

programs such as LIWC have been utilized to investigate the relationship of word usage in populations with mental health diagnoses. For people suffering from depression, research has shown first person-singular pronoun usage to be positively correlated with symptoms of depression following treatment (Zimmermann, Brockmeyer, Hunn, Schauenburg, & Wolf, 2017). Similarly, another study demonstrated that for participants diagnosed with anorexia and bulimia nervosa, usage of first person-singular pronouns during the recall of negative memories was positively correlated with self-reported symptoms of depression and anxiety (Brockmeyer et al, 2015). LIWC is a dictionary-based classification method, whereby the emotion word categories are based on a list of 915 positive and negative affect words (Pennebaker, Booth, Boyd, & Francis, 2015). LIWC has been utilized in a variety of ways, including understanding people's emotional reaction to the terrorist attack on September 11, 2001 (Cohn, Mehl, & Pennebaker, 2004), illustrating changes in emotional expression in published books over time (Acerbi, Garnet, Lampos, & Bentley, 2013), and predicting elections with Twitter data (Tumasjan et al, 2010). Emotion word dictionaries can identify positive and negative affect at a level competitive with human coding of emotional responses (Tausczik & Pennebaker, 2010; see also Kahn, Tobin, Massey, & Anderson, 2007). These dictionary-based techniques benefit from simplicity and interpretability, but require researchers to compile the word lists to create a comprehensive inventory of all positive and negative words. In addition, this technique does not allow a model to improve with more data.

Since the creation of dictionary-based programs, a number of new methods have been developed for performing text analysis. Using a dataset with sentences labeled by humans as positive or negative, these statistical models can predict whether the presence of words or phrases increased the likelihood of a sentence being labeled as positive or negative. Specifically, researchers have begun to use statistical techniques to attempt to model how the presence of words and phrases changes the probability of a passage being labeled as positive, negative or neutral (Jurafsky & Martin, 2008; Gonçalves et al, 2010; Pak & Paroubek, 2015). In practice, statistical NLP methods have been shown to be superior to lexical-based dictionary methods such as LIWC (Gonçalves et al, 2013), which are typically used by psychology researchers. For example, Bantum and Owen (2009) demonstrated that when analyzing an Internet-based psychological intervention for women with breast cancer, LIWC, in comparison with human raters, overidentified emotional expression. As larger collections of text labeled for sentiment have become available, NLP researchers have begun to use models that rely on advanced machine learning methods to identify sentiment (e.g., Recursive Neural Networks, see data analysis section for description; Socher et. al, 2013).

Detecting Emotion in Psychotherapy with Natural Language Processing

At present, text-based methods for evaluating emotion in psychotherapy are reliant on dictionary-based methods. Mergenthaler (1996) was one of the first researchers to create a quantitative method for measuring emotional expression in psychotherapy. Mergenthaler and Bucci (1999) hypothesized that key moments in the psychotherapy process involved client expression of both high emotional content and high verbal abstraction. To test this hypothesis, Mergenthaler used a dictionary-based method similar to LIWC; that is, a list of words that expressed either positive or negative emotional tones that were specific to

psychotherapy based text (Mergenthaler, 1996). In a similar study, Anderson and colleagues (1999) assessed therapist verb usage in high versus low affect segments in therapy sessions and found that therapists who used more cognitively oriented verbs in high affect sessions had worse outcomes. Similar to the limitations of LIWC, described above, these methods use *a priori* identification of positive and negative words, as opposed to empirical measurement learned from human ratings.

There are existing, publicly available tools that use statistical NLP tools to rate the valence of passages of text. However, these tools have been trained on immense text corpora obtained from domains other than psychotherapy such as classic literature (Yussupova, Bogdanova, & Boyko, 2012; Qiu, Liu, Bu, & Chen, 2011; Liu & Zhang, 2012), news articles (see Pang, & Lee, 2008 for a list of databases), and social media text (for examples see Bohlouli et al, 2015; Gokulakrishnan et al, 2012; Pak & Paroubek, 2015). Additionally, researchers have used a variety of models to harvest data from the Internet, including a live feed of tweets and posts from social media outlets as Twitter and Facebook (Bohlouli et al, 2015). Given the availability of these trained models, it is reasonable to wonder whether it might be useful to simply utilize one of these models to label emotion in psychotherapy - emotion in one domain is not necessarily different from emotion in another. However, this seems unlikely as 'domain adaptation' is a major subfield of machine learning, wherein researchers explore if models developed in one domain can be meaningfully applied to another (e.g., Is a sentence parser developed on newspaper articles accurate with conversational text; see, Dredze, Blitzer, Talukdar, Ganchev, Graca & Pereira, 2007). Pang and Lee (2008) have argued that sentiment analysis is quite likely domain specific. For example, if one were reviewing a movie and wrote that "the movie was very effective emotionally, deeply sad", the review might be rated as a very positive statement. But in a therapy session, the word "sad" would be more likely to be used in the context "I am feeling very sad". Moreover, there are many emotion-relevant words that might be extremely rare in other datasets, but are very common in psychotherapy. For example, "Zoloft" (an anti-depression medication) may never occur in a movie review corpus, but it is said 381 times in the collection of transcripts we use in the current study (http://alexanderstreet.com/). Moreover, psychotherapy text comes from spoken language, not written communication. Modeling strategies that work well on written text may perform poorly on spoken language (Jurafsky & Martin, 2008). For example, methods that require that language be structured based upon grammar (i.e., parse tree), may have difficulties analyzing disfluencies, or fillers and fragments that occur frequently in dialogue. Virtually all of the databases for training sentiment analysis models are written and none come from mental health domains.

With large and highly flexible deep neural networks, performance improvements are limited not by the model selection but by the quantity of labeled training data. To address these issues researchers have investigated how to extract meaningful representations from unlabeled textual data. Some early work achieved reasonable success by structuring the problem as learning a representation for words based on their context (e.g., Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014). Recently, strides have been made in combining a number of ideas and recent advances in NLP into one system - the Bidirectional Encoder Representations and Transformations (BERT; Devlin et al, 2018). BERT utilizes massive quantities of unlabeled data to learn useful representations

of language and linguistic concepts by masking portions of the input and trying to predict which word was in fact masked. As such, BERT can learn powerful representations of human language from billions of sentences. This massive pre-training makes it possible to fine-tune BERT on specific tasks introducing only minor task tweaks to the model and leveraging the knowledge acquired through extensive pertaining. Additionally, such extensive pre-training allows for BERT to outperform traditional models.

Given the abundance of linguistic information present in psychotherapy transcripts, modern NLP techniques have the potential to become highly efficient alternatives to relying solely on human ratings to gather emotion data present during psychotherapy. Statistical programs designed to conduct sentiment analysis can consume hundreds of hours of data, and create analyses of sessions almost immediately, whereas behavioral coding often takes months and requires a multitude of human resources. Mental health researchers have already demonstrated the capacity of developing and training more complex NLP models (see Imel et al, 2019). These methods may allow researchers to explore new and more complex questions about the role of emotion in psychotherapy and how it interacts with other psychotherapy processes (e.g., alliance, cultural discussions) and client outcomes (e.g., distress indices, satisfaction, dropout).

In summary, the role of emotions is clearly important to the process of psychotherapy, but there has been a lack of empirical research on this subject. One of the primary reasons is the shortcomings in methodology. Self-report is easy to obtain but coarse, and lexical methods have been limited to dictionary-based techniques, which can be insensitive to context. Advances within the field of NLP have provided methods that can improve the way that emotions are measured in a psychotherapy session. In an effort to use NLP to rate emotion in psychotherapy, Tanana et al (2016) compiled nearly 100,000 human labeled utterances and developed a model to identify, test, and compare four different sentiment models. However, the initial study was brief, and these previous models have not been compared to an existing psychotherapy dictionary-based model and requires an update with emergent NLP technology. In the current study, we extended findings from Tanana et al (2016) by comparing the sentiment model to LIWC and an innovative NLP model, BERT. We hypothesized that the prior NLP models from Tanana et al (2016) would outperform LIWC, and that BERT would outperform all models.

Method

Data Sources

The raw data was obtained from a large corpus of psychotherapy transcripts that were published by Alexander Street Press (http://alexanderstreet.com). Tanana et al (2016) provides only a brief description of the data and coding procedure, so the following description provides more detail. The original transcripts come from a variety of different therapists (e.g., Carl Rogers, Albert Ellis) and theoretical perspectives (Person-Centered, Rational Emotive, Psychodynamic, Experiential/Humanistic, Cognitive Behavioral and also include Drug Therapy/Medication Management sessions). Data was from real therapy sessions that were anonymized for confidentiality. The data included patient demographics, number of sessions, patient symptoms, and general topic of each therapy session. These

transcripts are available through a public university library subscription. As a result, they can be accessed more easily than typical psychotherapy datasets (see Imel, Steyvers, & Atkins, 2015 for additional details). The Alexander Street Press corpus has been utilized in prior research to train machine learning models on psychotherapy data (e.g., Xiao, Imel, Georgiou, Atkins, & Narayanan, 2015; Xiao, Huang, Imel, Atkins, Georgiou, & Narayanan, 2016; Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016; Gaut, Steyvers, Imel, Atkins, & Smyth, 2017).

At the time of developing the initial sentiment model, there were 2,354 session transcripts available, with 514,118 talk turns. The dataset includes speaker-identified talk turns, which are continuous periods where one speaker talks until the other speaker interrupts or responds. Before sampling from the dataset, we segmented talk turns on the punctuation indicating sentence boundaries (e.g. periods, exclamation and question marks indicated by the transcriber). We refer to these discrete units as utterances. We also excluded any talk turns that were shorter than 15 characters (a large part of the dataset consists of short filler text like ‘mmhmm’, ‘yeah’, ‘ok’ that are typically neutral in nature). We retained nonverbal indicators that were transcribed, like ‘(laugh)’ or ‘(sigh),’ because they might be useful indicators of the sentiment of the sentence. We randomly sampled 97,497 (19%) from the entire dataset of utterances that met the criteria for length, without any stratification by session.

Tanana et al (2016) used naïve coders to identify basic valence of emotion (e.g. positive, negative, neutral) among a large corpus of utterances from psychotherapy sessions. Valence has been identified as a primary component of emotion; according to the widely researched circumplex theory of emotion, emotions can be conceptualized as occurring on a continuum of positive and negative valence and high and low arousal (Russell, 2003). Focus on valence allowed for comparison with existing computer science models of sentiment (Pang & Lee, 2008), and positive and negative emotion categories in dictionary-based programs (Tausczik & Pennebaker, 2010). Naïve coding was utilized because previous research studies suggest that they are viable alternatives to identifying basic aspects of emotions like valence, and require less training than expert coders. Naïve coders are used, almost exclusively, in the field of computer science for tasks involving coding of positive/ negative emotions in text (Pang and Lee, 2008). In the field of psychology, naïve coders who received little training in identifying common emotions have been found to have adequate interrater reliability (Albright, Kenny, & Malloy, 1984; Ambady & Rosenthal, 1993), and have ratings similar to those of trained coders (Waldinger et al. 2004).

Amazon Mechanical Turk (MTurk) workers were recruited to code the dataset for sentiment. Researchers have found that workers on MTurk are more diverse than typical college samples, as well as other types of internet samples (Buhrmester, Kwang, & Gosling, 2011). Workers were limited to individuals in the United States to reduce the variability in the ratings to only US English speakers. In addition, it was required that workers were all ‘master’ certified by the system, meaning they had a track record of successfully performing other tasks. Each utterance was with a set of 7 others that were all completed at the same time (though all were selected randomly and were not in order). Workers were told that the utterances came from transcripts of spoken dialogue, and as a result are sometimes

messy, but were to try their best to rate each one. For each rating, workers were given the following five options: Negative, Somewhat Negative, Neutral, Somewhat Positive, Positive. Each utterance in the main dataset was rated by one person.

From the overall collection of 97,497 ratings, utterances were randomly split into training, development, and test subsets. This is a standard approach in machine learning in order to prevent overfitting the model to the training data. Tanana et al (2016) allocated 60% of the data to the training set (58,496 ratings), 20% to the development (19,503) and 20% to the test set (19,498). The training set was used to estimate model parameters, and the development set is used to periodically monitor performance and compare model variations on data that was not used for training. To ensure that the model did not begin to capitalize on chance in the development set, the model was run once on the test set to ensure that the final model performance was an accurate representation of how the model would perform on similar, unseen data (Hastie, Tibshirani, & Friedman, 2009).

Interrater Dataset—In addition to the main dataset, where one worker rated each utterance, a separate dataset was created where a random selection of 100 utterances were each rated by 75 workers (i.e., a single utterance was rated 75 times by different people). The purpose of this dataset was 1) to estimate interrater reliability of human coding of sentiment in psychotherapy, and 2) to estimate the distribution of sentiment ratings for different utterances, providing a direct estimate of the inherent uncertainty in making judgments on sentiment. Several decisions were made that resulted in a more conservative - though we would argue more accurate - estimate of inter-rater reliability. In contrast to many other studies identifying emotions from therapy sessions or text (e.g. Herrmann et al, 2016; Greenberg et al, 2007; Bantum & Owen, 2009; Choi et al, 2016), the sample of utterances was not restricted to those pre-determined to have high emotional content. Moreover, utterances with little content (less than 15 characters) were removed, which prior work suggests individuals tend to agree are neutral in valence, thus artificially inflating reliability. Finally, interrater reliability was examined at the utterance, rather than the session level, with the first four raters, utilizing Intraclass Correlation Coefficient (ICC) (Shrout & Fleiss, 1979). Using a two-way random effects model for absolute agreement, treating the data as ordinal, ICC = .81. (95% CI [.74, .86])[1].

In this study, we report the interrater reliability for individual utterances. It should be noted that this is a very different approach from other studies that typically report on ratings aggregated over longer passages or time periods (Auszra et al., 2013; Bantum & Owen, 2009; Greenberg et al., 2007; Herrmann et al., 2014), or entire sessions (Denecke & Deng, 2015). As a result, the interrater reliability may appear to be lower than other studies. Despite this choice, our interrater reliability remained in the moderate range. This finding suggests that we should not expect perfect performance from sentiment analysis models because not even humans completely agree on these types of ratings. However, it is reasonable to expect the best models to approach this level of performance.

[1]Many studies will use an ICC that is the average of k raters, which progressively increases as the number of raters increases. The ICC using ratings was estimated from the first 4 raters, and estimated an ICC(32,k) of .81. Due to the fact that we had 80 raters, the ICC(3,k) for this data was .99. Due to the fact that one rarely uses 80 raters, a much more conservative ICC(3,1) estimate was used.

Procedure

For the current study, we compared LIWC and BERT (described below) to the previous four NLP models from Tanana et al (2016) - unigram, bigram, trigram, and recursive neural net (RNN) models. We provide a more comprehensive description of each model below than the original study to aid with the comparison of otherwise complex models. Prior results from Tanana et al (2016) are presented in Table 1, along with results from the current study.

Features / Predictors—A common strategy for classifying language into categories in NLP is to break a sentence into what are called N-Grams (for a comprehensive tutorial, see Jurafsky and Martin, 2008). N-Grams represent individual vocabulary words and short phrases as a set of indicator variables, which are often several thousand in total. As a result, this type of NLP model represents any sentence as a very large list of possible n-grams (i.e., sparse vector). We tested several n-gram combinations in this study. N-Grams were created by parsing on word boundaries, without separating out contractions. For example, the word "don't" would be left as a single gram. Each model was tested with 1) unigram features 2) unigram + bigram features 3) unigram, bigram and trigram features. (Note: in the results when we write "tri-gram" this is short for uni-grams + bi-grams + tri-grams).

Classifier Models

N-Gram Models.: Tanana et al (2016) used a Maximum Entropy (MaxEnt) classifier model (Jaynes, 1990). This is a predictive model that is mathematically identical to logistic or multinomial regression. Specifically, we used the N-Grams present in a given sentence to identify the probability that a new sentence example fits into one of three categories (Negative, Neutral, Positive). The original variable had five categories (very negative, negative, neutral, positive, and very positive), but was reduced due to the lack of usage of the very negative and very positive codes. However, in this situation there are several important differences between the use of logistic regression and what might be typical in a classic statistical scenario. First, the number of possible predictors in a typical NLP scenario are much, much larger than in a typical statistical analysis (e.g., 40,374 predictors for the unigram model). Second, in a typical statistical scenario most psychology readers would be familiar with, logistic regression is fit simultaneously (though iteratively) to the entire dataset at once, which would take a prohibitive amount of time with so many predictors. To solve this problem, an approach called Stochastic Gradient Descent (SGD; see Bottou, 2012) was used, which considers the error for one case, or a small number of cases, makes a small update to the model parameters, and then iterates throughout the entire training set. Third, with NLP the statistical significance of any one predictor is not of interest, so no confidence intervals are estimated around each of the features. Finally, when estimating a model with thousands of predictors, it is easy to overfit the training data. Overfitting refers to a model that learns how to classify a phenomenon accurately in the training data set, but perform poorly on new and unseen data. To solve this problem, researchers use what is called a regularizer. Instead of training a model to minimize the error in a set, the model makes a compromise between minimizing error and minimizing the size of each parameter. This helps prevent the model from overfitting the training data and improves performance on new data (Tibshirani, 1996). All MaxEnt models were estimated using R version 3.3.2, with the *RTextTools* package (Collingwood, Jurka, Boydstun, Grossman, & van Atteveldt, 2013)

and the *maxent* package (Jurka et al, 2013). Tanana et al (2016) demonstrated that though the trigram model had only slightly higher accuracy, and was outperformed by the unigram model (kappa = 0.308).

Recursive Neural Network (RNN).: Tanana et al (2016) also tested a version of the sentiment model from the Stanford NLP toolkit (Manning, et al, 2014), which was based on an RNN, but was trained on movie reviews instead of psychotherapy transcripts (Socher et al, 2013). This model used neural networks organized in a tree structure to predict the sentiment of a sentence from the bottom of the tree (words) up (combinations of words). This model was one of the best performing models in the NLP literature that was publicly available and could be tested on the dataset. Thus, it is a state-of-the-art NLP model for sentiment, but one that had not been adapted to the psychotherapy domain. The RNN was outperformed by the n-gram models, and in particular the unigram model. See Figure 1 for an example of how a therapy session annotated for sentiment appears.

Comparison Models

We compared the performance of the four MaxEnt models trained on the Alexander Street Press dataset to two other models. 1) The LIWC coding method, which is commonly used in the psychology literature (Tausczik & Pennebaker, 2010) and similar in function to dictionary based methods that have been used in psychotherapy (Mergenthaler, 1996), and 2) an innovative deep learning approach called the Bidirectional Encoder Representations and Transformations (BERT; Devlin et al, 2018), also trained on written text (e.g., English Wikipedia).

Linguistic Inquiry and Word Count (LIWC; Pennebaker, Frances, & Booth, 2001).—LIWC, as mentioned in the literature review, uses the frequency of words in a document to classify the text on a number of different dimensions (e.g., affect, cognition, biological processes). We used the positive and negative emotion dimensions to categorize whether a statement was generally positive, negative or neutral. The LIWC coding system does not give any specific guidelines of how to turn the continuous rating into the categories of positive, negative and neutral (Tausczik & Pennebaker, 2010). To create these categories we subtracted the negative emotion dimension from the positive dimension. Any statement with a positive value was then classified as positive, a negative value as negative and a zero value as neutral.

Bidirectional Encoder Representations and Transformations (BERT).—Finally, in order to test a more recent innovation in NLP we used the Bidirectional Encoder Representations and Transformations (BERT; Devlin et al, 2018). BERT is a type of deep neural network which attentends to other words in a particular sentence depending on the current state of the network (i.e., attention mechanism). We combined 12 submodels, and initially trained on the Books Corpus (800 million words; Zhu et al, 2015) and English Wikipedia (2,500 million words). By utilizing prior language knowledge via a vast text corpus, BERT provides a highly advanced language detection model due to the quantity of training texts. It should be noted that BERT is not pretrained on *labeled* datasets, it is simply

learning from unlabeled english text by masking random words in a sentence and trying to predict which word best fits in this blank.

We used a pre-trained version of BERT that can be downloaded by other researchers from TensorFlow Hub (https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1). This model comes with a tokenizer that uses Byte-Pair Encoding (BPE) to convert words into discrete numeric representations. We used a process known as 'fine-tuning' where we begin with the pre-trained BERT model weights, but then allow them to change as they learn from our dataset. For the experiments in this paper, we allowed all layers of the model to learn from the data (not just the final classification layer). In using BERT to predict our sentiment classes we used a softmax classifier stacked above the 768 hidden units that were output above the <cls> token (a word that signals to the model that we are performing a classification task). Our training used the Adam optimizer (Kingma & Ba, 2014). We used 10% dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) as a regularizer to prevent our model from overfitting the data. The model was trained for 9140 steps with a batch size of 32. We used a 10% linear warm-up period to condition the model to the new dataset before moving on to a larger learning rate. The warm-up period followed a linear schedule of (step/warmup_period * lr). Over the course of 914 steps the learning rate linearly increased from 2.2e-8 to 2e-5. At which point the Adam optimizer was used with a beta_1=0.9, beta_2=0.999, weight decay of 0.01, and an epsilon of 1e-6.

Evaluation

All of the models were evaluated on how well they predicted the coarse human-generated sentiment labels, which were 'positive', 'negative' and 'neutral'. We used several metrics: 1) Overall accuracy predicting labels. This measure is the percentage of correctly predicted labels divided by the total number of guesses. 2) F1 score for each of the individual labels. F1 is a common metric used in machine learning that is designed to be a compromise between recall and precision. Recall is the percentage of a group or label that is correctly identified. Precision is the percentage of a label that a model identifies that is correct. F1 is the harmonic mean between these two metrics. 3) Cohen's Kappa (weighted) which gives a measure of correctly identified labels after correcting for chance guessing. Because the base rate for neutral was high in our dataset, the Kappa metric gives the best overall measure of the performance of these models. Although accuracy is reported, Kappa is a better metric because in our dataset, an accuracy of .59 could be achieved by guessing the neutral class for every utterance.

Results

As noted above, models were initially trained on the training and development subsets of data. Results are reported for a single, final fit to the test subset of data. Table 1 shows the relative performance of the different models on the test set, which none of the models were allowed to train on. All of the MaxEnt models significantly outperformed the LIWC model. The RNN model as well as the LIWC model have lower accuracy compared to all of the MaxEnt models. BERT performed significantly better than any of the other models tested with an overall kappa of .48.

In order to compare our model performance with human reliability, we computed the average pairwise Cohen's Kappa for course (positive, negative, neutral) human sentiment ratings. The average Kappa for this set was .42. Based on this computation, our best model exceeded human performance on a test set by 14%. This may be surprising that a model can exceed human performance on this task, but we should note that we are comparing to the *average* human-human agreement (some rater-pairs had agreement as high as .54). Moreover, the model we tested has exceeded the performance of individual humans on a number of NLP tasks (Devlin et al, 2018).

Discussion

This study compared existing sentiment models (Tanana et a, 2016) with the LIWC coding system, as well as an innovative deep learning technique BERT (Devlin et al, 2018). We found that the newer NLP method (BERT), which can leverage large existing language datasets, outperformed the prior n-gram and RNN models from Tanana et al, as well as the commonly used dictionary model LIWC. In practice, statistical NLP methods have been shown to be superior to lexical-based dictionary methods such as LIWC (Gonçalves et al, 2013). At present, psychotherapy researchers have been restricted to dictionary based attempts to model emotion with linguistic data. Our results suggest that these dictionary-based methods (e.g., LIWC) are largely similar to other models trained out of other domains (e.g., RNN), and that the linguistic detection of emotion in psychotherapy could be improved by utilizing newer methods that can leverage both prior knowledge about language in general, and training from an in-domain dataset (BERT). BERT was the superior model to the prior N-gram models, suggesting that context, and potentially the variety of training text available, provides a superior model for rating emotion. For example, one major difficulty for sentiment analysis methods is contrastive conjunctions (Socher et al, 2013). These are passages that contain two different clauses with the opposite sentiment. For example, "I sometimes like my boyfriend, but I've had it with this relationship." Dictionary based methods and n-gram models may have difficulties with these types of passages and may over or underestimate the sentiment present.

Psychotherapy is often an emotional process, and many theories of psychotherapy involve hypotheses about emotional expression as a potential catalyst for change. However, the methodologies available to explore these processes have been limited. One important reason for this gap in the literature is that it is time consuming and expensive for human coders to rate every utterance in a session for emotional expression. Additionally, some emotion coding systems, typically used in psychotherapy science (e.g., LIWC) are expensive programs and may not be widely utilized due to financial restrictions. In general, our results indicate that the modern machine learning methods perform better at predicting sentiment in psychotherapy dialogue than dictionary-based methods like LIWC that have been utilized to evaluate psychotherapy, as well as models trained on other domains. The methods presented have the possibility of being free, open source, solutions for emotion coding in psychotherapy. Thus, a competitive alternative to traditional methods. These results extend on current sentiment analysis research within the psychotherapy speech domain (e.g., Tanana et al, 2016), and provide methods for continued innovation in the field.

Limitations

Our paper relies on a broad conceptualization of emotion wherein raters were simply asked to rate the positive or negative sentiment of a set of text. However, it is unclear how the sentiment perceived by raters from reading the text maps onto the internal emotional state of clients or therapists in psychotherapy. It does not appear that this is well defined in the field of computer science. While sentiment appears to be the attitudes being expressed by an individual, it is unclear if that maps onto an internal emotional state. For example, rating the attitude of an individual that states 'this movie was terrible!' is relatively straightforward. The individual feels something negative towards the movie that they were assessing. However, in psychotherapy a patient may state that 'I don't need therapy anymore,' which may reflect a positive internal emotional state, if true. The same statement, however, may be indicative of a negative internal emotional state if the client is expressing resistance to treatment. As a result, assessing sentiment in psychotherapy may need clearer definitions and instructions than methods used for domains like restaurant or movie reviews. The lower rating of human reliability in this study may also suggest that for psychotherapy, researchers may need to create more specific rating systems than those used for movies or Twitter messages. The instructions 'rate the sentiment of the following phrase' may be clear when applied to movie reviews, but may be unclear for psychotherapy. Future research might experiment with several different rating systems and compare the interrater reliability of each type. For example, researchers may test the rating system used in this paper against others like 'please rate how you think the person who said these words feels.' Alternatively, it may be that in some cases, the emotional expression of the speaker is truly ambiguous, and thus a distribution of ratings may be a better representation of the emotional state of the target as compared to a specific single point on a scale. Future work should explore the differences between rating sentiment and rating emotional expression.

Future Directions in Psychotherapy Research

Our results suggest that NLP based models can be useful tools for more nuanced examinations of psychotherapy processes, given that the NLP models we used were able to predict sentiment of therapist and client utterances more accurately than traditional dictionary based methods of sentiment analysis (Merganthaler, 1996). While constructs such as the working alliance (Horvath & Symonds, 1991), empathy (Elliot et al., 2011), and cultural awareness (Tao, Owen, Pace & Imel, 2015) have been identified as important to client symptom improvement, there is still limited understanding of how these processes unfold in sessions. More specifically, despite the importance of emotion in psychotherapy there is little focus on how immediate changes in sentiment are associated with important therapeutic constructs. Applying sentiment analysis to psychotherapy transcripts could also allow researchers to better understand patterns of emotional interactions (e.g., countertransference; Dahl, Røssberg, Bøgwald, Gabbard, & Høglend, 2012) that contribute to positive or negative perceptions of important therapeutic processes by examining changes in sentiment in each therapist-client conversation exchange. Prior research has demonstrated that reviewing sessions with computer-based recording systems can increase personal reflection, and aid in supervision and collaboration during psychotherapy training (Slovák et al, 2015; Murphey et al, 2019). Future studies could investigate clinician experiences of utilizing BERT during supervision, whereby trainees could report on their perceptions

of client sentiment, and review BERT outputs with their supervisor to reflect on their experiences, and potentially on broader psychotherapy processes happening. Integrating NLP based models with both self-report measures of therapeutic processes and physiological measures of arousal could lend further understanding of how certain emotions, arousal, and perceptions of therapeutic processes interact together, using larger datasets and fewer resources.

In addition to better understanding psychotherapy processes, future research can also focus on exploring how changes in emotional expression in therapist-client interactions are associated with symptom improvement and outcomes. While researchers have started examining how in-session emotional expressions and interactions are associated with symptom improvement (e.g. McCarthy et al., 2011; Walter et al., 2009), the small sample sizes in these studies limit their generalizability. Examining changes in sentiment using large psychotherapy datasets that encompass different treatments could help researchers identify patterns of therapist and client emotional expression that are associated with improved clinical outcomes across treatments. Similarly, future research could also examine therapist variability in eliciting emotional change in clients and its relationship clinical outcomes. Although therapists are significant sources of variability in client outcomes (Del Re et al., 2012; Laska et al., 2013; Wampold & Brown, 2005), to date there are no large-scale studies examining in-session therapist behaviors that account for these differences. Given the importance of communication and expression of emotions in therapy, examining how therapists vary in eliciting emotions in their clients, and how these differences relate to symptom improvement, could shed insight into therapist behaviors that account for clinical change. Future research could examine how therapists differ in eliciting emotions in clients, how these differences are associated with symptom improvement, and other factors (e.g. client and therapist demographics, client presenting concerns) that may interact with emotional expression and symptom improvement using NLP models.

Applications to Clinical Practice

In addition to research implications, NLP models predicting sentiment can also facilitate supervision of clinicians-in-training. Combined with systems that can automatically transcribe entire psychotherapy sessions (Georgiou, Black, Lammert, Baucom, & Narayanan, 2011), methods employed in this paper could help supervisors and supervisees better focus on specific parts of therapy sessions. Currently, supervision is based mainly on verbal reports from supervisees, without any quantitative indicators of important moments during the hours of psychotherapy that may have taken place over a week (Amerikaner & Rose, 2012; Goodyear & Nelson, 1997). Visualizations of coded therapy sessions for sentiment (like the ones presented in this paper) could allow supervisors to select session recordings that appear emotionally salient (e.g. very positive, negative, or emotionally variable) to review with their supervisees. For instance, it may be particularly important for therapists to identify negative emotions to better understand client avoidance of negative emotions (e.g., Acceptance and Commitment Therapy; Hayes et al., 2006). While NLP models of sentiment may not accurately label all therapist and client statements, these models could still allow supervisors to focus on specific moments in therapy that *may* be of interest, see how their supervisee perform, and process with their supervisees what occurred.

Using a methodology like this may help supervisors use their time more efficiently and listen to portions of psychotherapy sessions that go beyond those that were selected by their supervisees.

Automatically transcribing a psychotherapy session, identifying speakers, and automatically coding the sentiment of every statement may seem unrealistic and impractical on the surface, but have been well studied and generally available. Methods like the ones used in Georgiou et al (2011) to transcribe and identify speakers in long recordings are now commercially available from companies like amazon (https://aws.amazon.com/transcribe/). The methods presented in this paper to identify sentiment from text are relatively standard methods in the field of natural language processing (Jurafsky & Martin, 2008). With the hope of continuing to advance psychotherapy science and clinical practice (e.g., Barnard, 2004), NLP represents one innovative interdisciplinary channel in which therapists may find additional resources.

Conclusion

Psychotherapy often revolves around the discussion of emotionally charged topics, and most theories of psychotherapy involve some idea of how emotions influence future behavior. However, it has been extremely difficult to study these processes in an empirical way because manually coding sessions for emotional content is expensive and time consuming. In psychotherapy, researchers have typically relied on LIWC in an attempt to automate this laborious coding, but this method has serious limitations. More modern NLP methods exist, but have been trained on out of domain datasets that do not perform well on psychotherapy data. This study proposes a method that addresses these problems by training on a large dataset of psychotherapy based data, which outperformed both LIWC and a modern publicly available NLP method trained on out of domain data. However, much more improvement is needed to reach the same performance as human raters. This approach represents an important first step towards allowing researchers to begin a more rigorous study of emotion in psychotherapy.

References

Aafjes-van Doorn K, Kamsteeg C, Bate J, & Aafjes M (2020). A scoping review of machine learning in psychotherapy research. Psychotherapy Research, 1–25.

Albright L, Kenny DA, & Malloy TE (1988). Consensus in personality judgments at zero acquaintance. Journal of Personality and Social Psychology, 55(3), 387. [PubMed: 3171912]

Ambady N, & Rosenthal R (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of Personality and Social Psychology, 64(3), 431.

Amerikaner M, & Rose T (2012). Direct observation of psychology supervisees' clinical work: A snapshot of current practice. The Clinical Supervisor, 31, 61–80. 10.1080/07325223.2012.671721

Anderson T, Bein E, Pinnell B, & Strupp H (1999). Linguistic analysis of affective speech in psychotherapy: A case grammar approach. Psychotherapy research, 9(1), 88–99.

Baccianella S, Esuli A, & Sebastiani F (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, pp. 2200–2204).

Bar-On R, Tranel D, Denburg NL, & Bechara A (2004). Emotional and social intelligence. Social neuroscience: key readings, 223.

Barnard PJ (2004). Bridging between basic theory and clinical practice. Behaviour Research and Therapy, 42(9), 977–1000. [PubMed: 15325897]

Bohlouli M, Dalter J, Dronhofer M, Zenkert J, & Fathi M (2015). Knowledge discovery from social media using big data-provided sentiment analysis (somabit). Journal of Information Science, 41(6), 779–798.

Bottou L (2012). Stochastic gradient descent tricks. Neural Networks: Tricks of the Trade, 1(1), 1–16.

Brockmeyer T, Zimmermann J, Kulessa D, Hautzinger M, Bents H, Friederich HC, … & Backenstrass M (2015). Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. Frontiers in Psychology, 6, 1564. [PubMed: 26500601]

Buhrmester M, Kwang T, & Gosling SD (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 6(1), 3–5. [PubMed: 26162106]

Choi BH, Pos AE, & Magnusson MS (2016). Emotional change process in resolving self-criticism during experiential treatment of depression. Psychotherapy Research : Journal of the Society for Psychotherapy Research, 26(4), 484–499. [PubMed: 26067352]

Chui H, Hill CE, Kline K, Kuo P, & Mohr JJ (2016). Are you in the mood? Therapist affect and psychotherapy process. Journal Of Counseling Psychology, 63(4), 405–418. doi:10.1037/cou0000155 [PubMed: 27177026]

Collingwood L, Jurka T, Boydstun AE, Grossman E, & van Atteveldt WH (2013). RTextTools: A supervised learning package for text classification.

Crawford John R.; Henry Julie D. (2004). "The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample" (PDF). British Journal of Clinical Psychology.

Dahl HSJ, Røssberg JI, Bøgwald KP, Gabbard GO, & Høglend PA (2012). Countertransference feelings in one year of individual therapy: An evaluation of the factor structure in the feeling word checklist-58. Psychotherapy Research, 22(1), 12–25. [PubMed: 22040366]

Del Re AC, Flückiger C, Horvath AO, Symonds D, & Wampold BE (2012). Therapist effects in the therapeutic alliance–outcome relationship: A restricted-maximum likelihood meta-analysis. Clinical Psychology Review, 32(7), 642–649. [PubMed: 22922705]

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

Diener E, Scollon CN, & Lucas RE (2009). The evolving concept of subjective well-being: The multifaceted nature of happiness. In Assessing well-being (pp. 67–100).

Dredze M, Blitzer J, Talukdar PP, Ganchev K, Graca J, & Pereira FC (2007, 6). Frustratingly Hard Domain Adaptation for Dependency Parsing. In EMNLP-CoNLL (pp. 1051–1055).

Ellis A (1962). Reason and emotion in psychotherapy. Stuart Lyle.

Elliott R, Bohart AC, Watson JC, & Greenberg LS (2011). Empathy. Psychotherapy, 48(1), 43–49. doi:10.1037/a0022187 [PubMed: 21401273]

Freud S, & Breuer J (1895). Studies on hysteria. se, 2. London: Hogarth.

Gokulakrishnan B, Priyanthan P, Ragavan T, Prasath N, & Perera A (2012). Opinion mining and sentiment analysis on a twitter data stream. in advances in ict for emerging regions (icter), 2012 international conference. IEEE, 182–188.

Gonçalves P, Araújo M, Benevenuto F, & Cha M (2013, 10). Comparing and combining sentiment analysis methods. In Proceedings of the first ACM conference on Online social networks (pp. 27–38). ACM.

Goodyear RK, & Nelson ML (1997). The major formats of psychotherapy supervision. In Watkins CC (Ed.), Handbook of psychotherapy supervision (pp. 328–344). Hoboken, NJ: Wiley.

Georgiou PG, Black MP, Lammert AC, Baucom BR, & Narayanan SS (2011, 10). "That's aggravating, very aggravating": is it possible to classify behaviors in couple interactions using automatically derived lexical features?. In International Conference on Affective Computing and Intelligent Interaction (pp. 87–96). Springer, Berlin, Heidelberg.

Hastie T, Tibshirani R, & Friedman J (2009). Unsupervised learning. In The elements of statistical learning (pp. 485–585). Springer, New York, NY.

Hayes SC, Luoma JB, Bond FW, Masuda A, & Lillis J (2006). Acceptance and commitment therapy: Model, processes and outcomes. Behaviour research and therapy, 44(1), 1–25. [PubMed: 16300724]

Horvath AO, & Symonds BD (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. Journal of Counseling Psychology, 38(2), 139.

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

Imel ZE, Steyvers M, & Atkins DC (2015). Psychotherapy Computational Psychotherapy Research : Scaling up the Evaluation of Patient – Provider Interactions Computational. Psychotherapy.

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

Isen AM (2008). Some ways in which positive affect influences decision making and problem solving. Handbook of emotions, 3, 548–573.

Jaynes ET, 1990. Notes on present status and future prospects. In: Grandy WT Jr., Schick LH (Eds.), Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht, The Netherlands, 1–13.

Joormann J, & Stanton CH (2016). Examining emotion regulation in depression: A review and future directions. Behaviour Research and Therapy, 86, 35–49. [PubMed: 27492851]

Jurafsky D, & Martin JH (2014). Speech and language processing. Pearson.

Jurka TP, Tsuruoka Y, Jurka MTP, Rcpp I, Rcpp L, & Tsuruoka Y (2013). Package 'maxent'.

Kahn JH, Tobin RM, Massey AE, & Anderson JA (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. The American journal of psychology, 263–286. [PubMed: 17650921]

Kingma DP, & Ba J (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kohut H (2013). The analysis of the self: A systematic approach to the psychoanalytic treatment of narcissistic personality disorders. University of Chicago Press.

Kramer U, Pascual-Leone A, Rohde KB, & Sachse R (2016). Emotional processing, interaction process, and outcome in clarification-oriented psychotherapy for personality disorders: A process-outcome analysis. Journal of Personality Disorders, 30(3), 373–394. 10.1521/pedi_2015_29_204 [PubMed: 26111248]

Kramer U, Pascual-Leone A, Despland J, & de Roten Y (2015). One minute of grief: Emotional processing in short-term dynamic psychotherapy for adjustment disorder. Journal of Consulting And Clinical Psychology,83(1), 187–198. doi:10.1037/a0037979 [PubMed: 25244391]

Lane RD, Ryan L, Nadel L, & Greenberg L (2015). Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science. Behavioral and Brain Sciences, 38, e1.

Lang PJ, & Bradley MM (2010). Emotion and the motivational brain. Biological psychology, 84(3), 437–450. [PubMed: 19879918]

Laska KM, Smith TL, Wislocki AP, Minami T, & Wampold BE (2013). Uniformity of evidence-based treatments in practice? Therapist effects in the delivery of cognitive processing therapy for PTSD. Journal of Counseling Psychology, 60(1), 31. [PubMed: 23356465]

Liu B, & Zhang L (2012). A survey of opinion mining and sentiment analysis. Mining Text Data, 415–463.

Lorimer B, Delgadillo J, Kellett S, & Brown G (2019). Exploring relapse through a network analysis of residual depression and anxiety symptoms after cognitive behavioural therapy: A proof-of-concept study. Psychotherapy Research.

McCarthy KL, Mergenthaler E, Schneider S, & Grenyer BF (2011). Psychodynamic change in psychotherapy: Cycles of patient–therapist linguistic interactions and interventions. Psychotherapy Research, 21(6), 722–731. [PubMed: 21955173]

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. [PubMed: 8991317]

Mergenthaler E (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. Psychotherapy Research, 18(2), 109–126. [PubMed: 18815969]

Messina I, Palmieri A, Sambin M, Kleinbub JR, Voci A, & Calvo V (2013). Somatic underpinnings of perceived empathy: The importance of psychotherapy training. Psychotherapy Research, 23(2), 169–177. doi:10.1080/10503307.2012.748940 [PubMed: 23234457]

Mikolov T, Sutskever I, Chen K, Corrado G, & Dean J (2013). Distributed representations of words and phrases and their compositionality. Proceedings of the 26th conference on Neural Information Processing Systems, 2, 3111–3119.

Mischel W (2013). Personality and assessment. Psychology Press.

Missirlian TM, Nasukawa T, & Yi J (2003). Sentiment analysis. In Proceedings of the international conference on Knowledge capture - K-CAP '03 (p. 70). New York, New York, USA: ACM Press.

Murphy D, Slovak P, Thieme A, Jackson D, Olivier P, & Fitzpatrick G (2019). Developing technology to enhance learning interpersonal skills in counsellor education. British Journal of Guidance & Counselling, 47(3), 328–341.

Pak A, & Paroubek P (2015). Twitter as a corpus for sentiment analysis and opinion mining. In LREc, 10, 1320–1326.

Pang B, & Lee L (2008). Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retreival, 1(2), 91–231.

Pennebaker JW, Booth RJ, Boyd RL, & Francis ME (2015). Linguistic Inquiry and Word Count: LIWC2015 Operator's Manual. Retrieved April 28, 2016.

Pennington J, Socher R, & Manning C (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

Qiu G, Liu B, Bu J, & Chen C (2011). Opinion word expansion and target extraction through double propagation. Computational Linguistics, 37(1), 9–27.

Read J (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. Proceedings of ACL, 43–48.

Rogers CR (1975). Empathic: An unappreciated way of being. Couns. Psychol, 5(2), 2–10.

Russell JA (2003). Core affect and the psychological construction of emotion. Psychological review, 110(1), 145. [PubMed: 12529060]

Safran JD, & Muran JC (2000). Negotiating the therapeutic alliance: A relational treatment guide. Guilford Press.

Samuel A (1962). Artificial Intelligence: A Frontier of Automation. The Annals of the American Academy of Political and Social Science, 340, 10–20.

Schacter DL (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. American psychologist, 54(3), 182.

Shiner B, Westgate CL, Simiola V, Thompson R, Schnurr PP, & Cook JM (2018). Measuring use of evidence-based psychotherapy for PTSD in VA residential treatment settings with clinician survey and electronic medical record templates. Military Medicine, 183(9–10), e539–e546. [PubMed: 29547909]

Shrout PE, & Fleiss JL (1979). Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull, 86(2), 420–428. [PubMed: 18839484]

Sloan DM, & Kring AM (2007). Measuring changes in emotion during psychotherapy: Conceptual and methodological issues. Clinical Psychology: Science and Practice, 14(4), 307–322.

Sloan DM, Marx BP, & Epstein EM (2005). Further examination of the exposure model underlying the efficacy of written emotional disclosure. Journal of Consulting and Clinical Psychology, 73(3), 549. [PubMed: 15982152]

Slovák P, Thieme A, Murphy D, Tennent P, Olivier P, & Fitzpatrick G (2015, 2). On becoming a counsellor: Challenges and opportunities to support interpersonal skills training. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 1336–1347).

Socher R, Pennington J, Huang E, Ng AY, & Manning CD (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. Proceedings of the EMNLP, (ii), 151–161.

Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, & Potts C (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proc. Conf. Empir. Methods

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, & Salakhutdinov R (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Eesearch, 15(1), 1929–1958.

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

Tausczik YR, & Pennebaker JW (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24–54.

Tibshirani R (1996). Regression Selection and Shrinkage via the Lasso. Journal of the Royal Statistical Society. 58(1), 267–288.

Waldinger RJ, Schulz MS, Hauser ST, Allen JP, & Crowell JA (2004). Reading Others' Emotions: The Role of Intuitive Judgments in Predicting Marital Satisfaction, Quality, and Stability. Journal of Family Psychology, 18(1), 58. [PubMed: 14992610]

Walter H, von Kalckreuth A, Schardt D, Stephan A, Goschke T, & Erk S (2009). The temporal dynamics of voluntary emotion regulation. PLoS one, 4(8), e6726. [PubMed: 21949675]

Wampold BE, & Brown GSJ (2005). Estimating variability in outcomes attributable to therapists: a naturalistic study of outcomes in managed care. Journal of Consulting and Clinical Psychology, 73(5), 914. [PubMed: 16287391]

Yussupova N, Bogdanova D, & Boyko M (2012). Applying of sentiment analysis for texts in russian based on machine learning approach. In Proceedings of second international conference on advances in information mining and management (pp. 8–14).

Xiao B, Imel ZE, Georgiou PG, Atkins DC, & Narayanan SS (2015). " Rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. PloS one, 10(12), e0143055. [PubMed: 26630392]

Xiao B, Imel ZE, Atkins DC, Georgiou PG, & Narayanan SS (2015). Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In Sixteenth Annual Conference of the International Speech Communication Association.

Zimmermann J, Brockmeyer T, Hunn M, Schauenburg H, & Wolf M (2017). Firstperson Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. Clinical psychology & psychotherapy, 24(2), 384–391. [PubMed: 26818665]

Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, & Fidler S (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision (pp. 19–27).

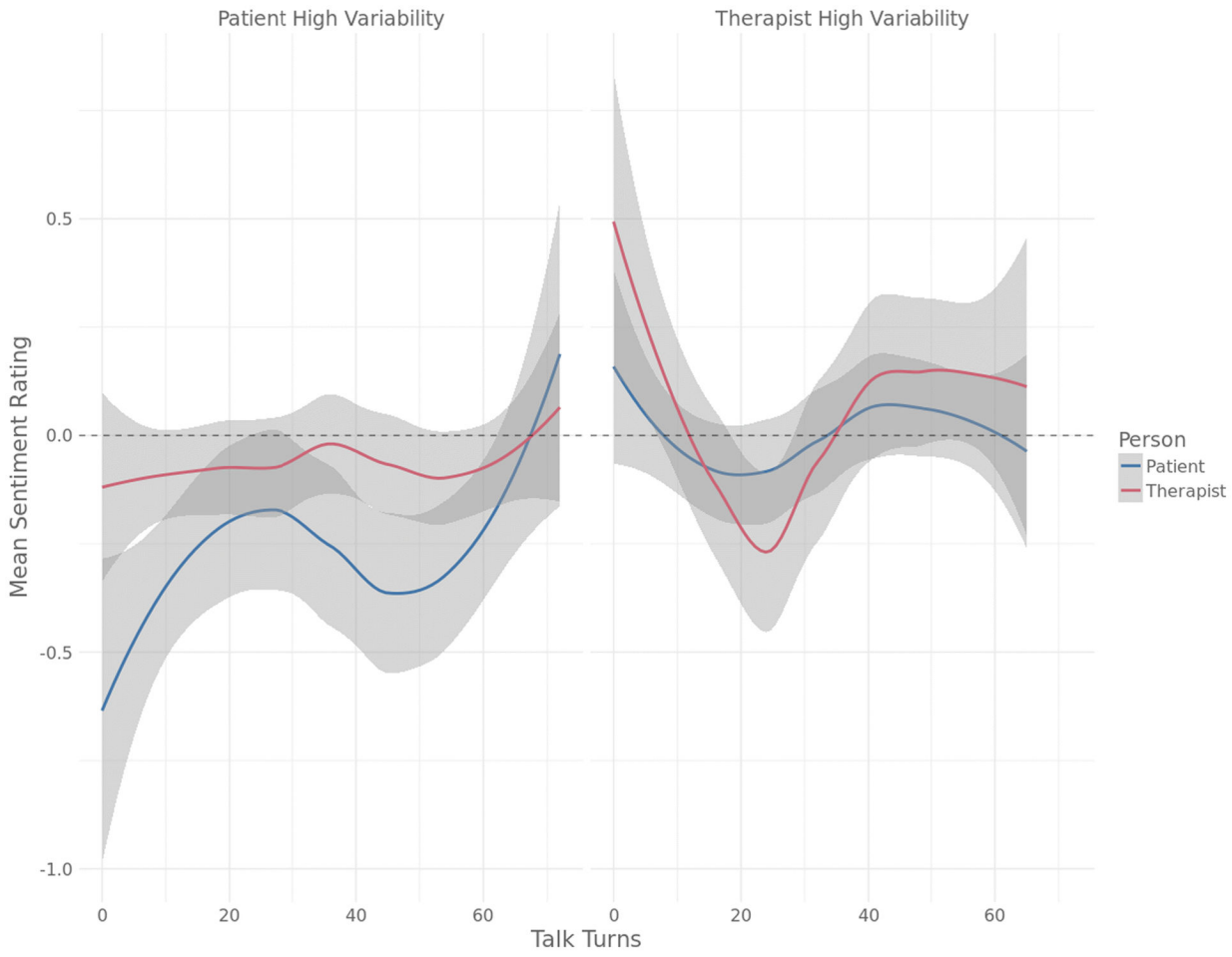

Figure 1:
Example Therapy Session Annotated for Sentiment

Table 1.

Model Performance on Test Set

			F1			Kappa	
	Model	**Accuracy**	**Neutral**	**Positive**	**Negative**	**Estimate**	**95% CI**
BERT MaxEnt		0.66	0.73	0.47	0.59	0.48	[.46–.50]
	Unigram	0.60	0.71	0.34	0.45	0.31	[.29–.32]
	Bigram	0.60	0.71	0.34	0.45	0.31	[.29–.31]
	Trigram	0.61	0.71	0.34	0.43	0.30	[.28–.31]
Comparison							
	RNN (Trained on Movie Reviews)	0.49	0.56	0.32	0.45	0.23	[.21–.24]
	LIWC	0.55	0.67	0.36	0.38	0.25	[.23–.26]