Is Sentiment in Movies the Same as Sentiment in Psychotherapy? Comparisons Using a New Psychotherapy Sentiment Database

Michael Tanana[1], Aaron Dembe[1], Christina S. Soma[1], David Atkins[2], Zac Imel[1] and Vivek Srikumar[3]
[1] Department of Educational Psychology, University of Utah, Salt Lake City, UT
[2] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA
[3] School of Computing, University of Utah, Salt Lake City, UT
michael.tanana@utah.edu aaron.dembe@utah.edu CSoma@sa.utah.edu
zac.imel@utah.edu datkins@u.washington.edu svivek@cs.utah.edu

Abstract

The sharing of emotional material is central to the process of psychotherapy and emotional problems are a primary reason for seeking treatment. Surprisingly, very little systematic research has been done on patterns of emotional exchange during psychotherapy. It is likely that a major reason for this void in the research is the enormous cost of annotating sessions for affective content. In the field of NLP, there have been major strides in the creation of algorithms for sentiment analysis, but most of this work has focused on written reviews of movies and twitter feeds with little work on spoken dialogue. We have created a new database of 97,497 utterances from psychotherapy transcripts labeled by humans for sentiment. We describe this dataset and present initial results for models identifying sentiment. We also show that one of the best models from the literature, trained on movie reviews, performed below many of our baseline models that trained on the psychotherapy corpus.

1 Introduction

People often seek psychotherapy because they feel emotionally distressed (e.g. anxious, unable to sleep). For well over a century, researchers and practitioners have consistently acknowledged the central role emotions play in psychotherapy (Freud and Breuer, 1895; Lane et al., 2015). Emotion, or affect, is directly involved in key concepts of psychotherapeutic process and outcome, including the formation of the therapeutic alliance (Safran and Muran, 2000), an individual's process of decision making (Bar-On et al., 2004; Isen, 2008) , behavior change (Lang and Bradley, 2010), personality style (Mischel, 2013), and happiness (Gross and Levenson, 1997). Affect is implicated in human memory (Schacter, 1999), and is an essential building block of empathy (Elliott et al., 2011; Imel et al., 2014a). The particular role of affect in different psychotherapy theories varies from encouraging patients to access and release suppressed emotions (as in psychoanalysis; e.g. (Kohut, 2013)) to identifing the impact of cognition on emotion (as in rational-emotive behavior therapy; (Ellis, 1962)). Carl Rogers, a progenitor of humanistic / person-centered therapy, theorized that empathy involved a therapist experiencing a client's affect as if it were his or her own, and that empathy constituted a necessary ingredient for human growth and change (Rogers, 1975). Empathy and emotion continues to be a primary area of research in psychological science (Decety and Ickes, 2009).

In psychotherapy, there are many ways that clients and therapists communicate how they are feeling (e.g., facial expression (Haggard and Isaacs, 1966), body positioning (Beier and Young, 1998), vocal tone (Imel et al., 2014a), but clearly one is the words they use. For example, there is evidence that greater use of affect words predicts positive treatment outcome (Anderson et al., 1999; Stalikas, 1995). Similarly, Mergenthaler (1996) developed a theory on how the pattern of emotional expression should proceed between a client and therapist. However, this research has been limited to dictionary based methods (see also Mergenthaler (2008)). Until very re-

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 33–41,
San Diego, California, June 16, 2016.

cently, the exploration of emotion in psychotherapy has been limited by the lack of methodology for looking at sentiment in a more nuanced way.

2 Sentiment Analysis

There is a long tradition in the field of Natural Language Processing (NLP) for trying to correctly identify the sentiment of passages of text and as a result there are a large number of techniques that have been tested (for a review on the subject, see Pang and Lee (2008)). Some common methods involve using n-grams combined with classifier models (SVM, CRF, Naive Bayes) to identify the sentiment of sentences or passages (Pak and Paroubek, 2015). Another method involves using pre-compiled dictionaries of common terms with their polarity (positive or negative) (Baccianella et al., 2010). As with many NLP methods, researchers have attempted to go beyond the mere presentation of words and use sentence structure and contextual information to improve accuracy. Along these lines, more recently researchers have used deep learning techniques to improve accuracy on sentiment datasets, with some success (Maas et al., 2011; Socher et al., 2013).

2.1 Domain Adaptation: Why Create A New Sentiment Dataset

The purpose of this project is to create and evaluate a dataset for training machine learning sentiment analysis models that could then be applied to the domain of psychotherapy and mental health. Pang and Lee (2008) have pointed out that sentiment analysis is domain specific. Thus, creating a sentiment dataset specific to psychotherapy addresses the possibility that the words and ratings used to train models in other contexts may have very different connotations than those in spoken psychotherapy. For example, if one were reviewing a movie and wrote that 'the movie was very effective emotionally, deeply sad', this 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 words that might be extremely rare in other datasets, but are very common in psychotherapy. For example, the word 'Zoloft' (an anti-depression medication) may never occur in a movie review dataset, but it occurs 381 times in our collection of therapy transcripts. Moreover, psychotherapy text typically comes from transcribed dialogue - not written communication. Modeling strategies that work well on written text may perform poorly on spoken language. For example, methods that require parse trees (recursive neural nets) may have difficulty on the disfluencies, fillers and fragments that come from dialogue.

Databases used for sentiment analysis have come from a variety of written prose ranging from classic literature (Yussupova et al., 2012; Qiu et al., 2011; Liu and Zhang, 2012), news articles (see Pang and Lee (2008) for a list of databases), to social media text (for examples see Bohlouli et al. (2015), Gokulakrishnan et al. (2012) and Pak and Paroubek (2015)). Databases have been created from archived text via the Internet. Additionally, researchers have used a variety of techniques to harvest a live feed of tweets and posts from social media outlets as Twitter and Facebook, respectively, so as to access fresh data (Bohlouli et al., 2015). Virtually all of the databases for sentiment analysis are written and none (that we are aware of) come from a mental health domain.

3 Data Collection

Data were obtained from a large corpus of psychotherapy transcripts that are published by Alexander Street Press (http://alexanderstreet.com/). These transcripts come from a variety of different theoretical perspectives (Psychodynamic, Experiential/Humanistic, Cognitive Behavioral and Drug Therapy/Medication Management) (Imel et al., 2014b). Importantly, these transcripts are available through library subscription and can be downloaded from the web. As a result they can be shared more easily than a typical psychotherapy datasets. At the time of writing, there were 2,354 sessions, with 514,118 talk turns.

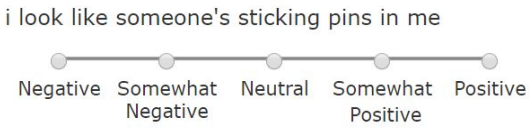

Figure 1: Example Mechanical Turk Rating

Before sampling from the dataset, we segmented

talk turns on sentence boundaries (based on periods, exclamation and question marks). 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 'mm-hmm', 'yeah', 'ok' that are neutral in nature). We left in non-verbal indicators that were transcribed like '(laugh)' or '(sigh)'. We randomly sampled from the entire dataset of utterances that met the criteria for length, without any stratification by session.

We used Amazon Mechanical Turk (MTurk) to code the dataset for sentiment. We limited the workers to individuals in the United States to reduce the variability in the ratings to only US English speakers. In addition, we required that workers were all 'master' certified by the system (which means that they had a track record of successfully performing other tasks). We packaged each utterance 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 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 (see figure 1 to see the exact presentation). Each utterance in the main dataset was rated by one person.

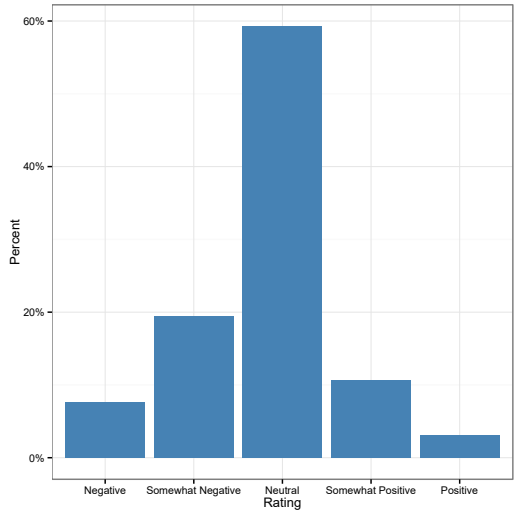

Figure 2: Distribution of Sentiment Ratings

3.1 Interrater Dataset

In addition to the main dataset, where one worker rated each utterance, we created another dataset where a random selection of 100 utterances were rated by 75 workers each. The purpose of this dataset was 1) to estimate the numeric interrater reliability of human coding of sentiment and 2) to be able to see the distribution of sentiment ratings for different utterances.

4 Data Description

4.1 Sentiment Dataset Description

The sentiment ratings were completed by 221 different workers on MTurk. The workers completed 97,497 ratings. The mean length of the utterances was 13.6 (SD = 11.1) and the median length was 10 words. The most frequent rating was neutral (59.2%) and the ratings generally skewed more negative than positive (see figure 2).

There was a similar trend to the one observed by Socher et al. (2013) that shorter sentences tended to be more neutral than longer ones. Though in contrast, even in longer phrases, our dataset skewed more negative and had a larger neutral percentage. This makes sense, given that the dataset comes from a collection of psychotherapy transcripts where participants are likely to be discussing the problems that brought the client to psychotherapy.

4.2 Data Splits

From the overall collection of 97,497 ratings we randomly split the data into a training, development and test set. We allocated 60% to the training set (58,496), 20% to the development (19,503) and 20% to the test set (19,498).

4.3 Interrater Dataset Description

The interrater dataset was used to determine the level of interrater agreement when rating sentiment in this dataset. We used the Intraclass Correlation Coefficient (ICC) to assess this agreement (Shrout and Fleiss, 1979). Using a two way random effects model for absolute agreement, treating the data as ordinal, the ICC was .54[1]. (95% CI [.47, .62])[2].

[1]This rated "fair" by the criterion of Cicchetti (1994)

[2]CI=Confidence Interval

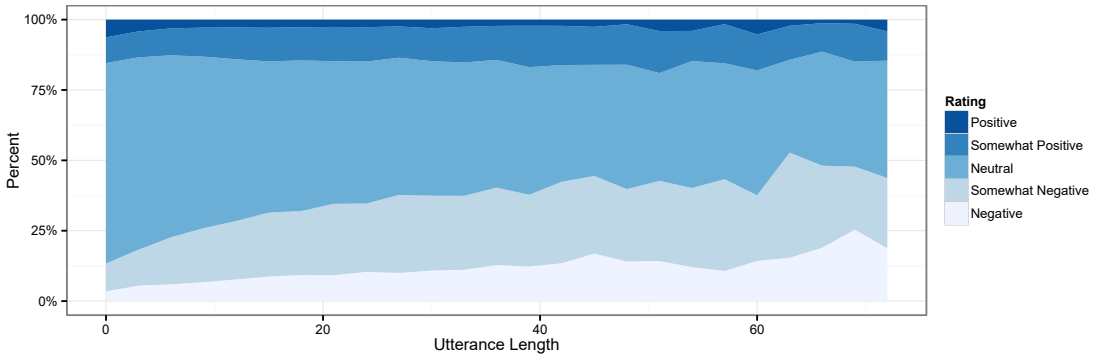

Figure 3: Distribution of Sentiment Ratings by Length of Utterance

The interrater set provides an illustration of how different types of utterances produce different responses in people. We present several examples in figure 4 that were chosen to illustrate different patterns in ratings. As can be seen in the examples, there are certain ratings where the vast majority of raters agree on the sentiment. For example 'and then left the house' was rated neutral by most of the raters. Another example, 'no, I don't even like him' was agreed to be some degree of negative by most raters. Another utterance that was generally rated positive was '(chuckling) but I know I didn't feel as good as I do now'. But even in examples where the vast majority of raters agreed on the direction of the rating (positive or negative) there was almost never complete agreement on the degree of sentiment. This finding lends support to the method of training models that predict the polarity of the sentiment, but not the degree (Socher et al., 2013). In many utterances a large proportion of raters agreed a phrase was not neutral, but there was low agreement on what direction of the sentiment was. The example 'see I don't need any therapy' illustrates this point. The modal rating was neutral, but the example had a wide distribution of ratings. Different raters had very different views on the sentiment of the sentence. It is possible that these different assesments could map onto ways in which therapists might view such a statement - some taking it at face value and an indicator a client was doing well, while others might view it as a failure to acknowledge problems that brought them to psychotherapy.

5 Models

We tested several common NLP models to predict the labels on the dataset from the text. The purpose of the modeling was to build baseline measures that could serve as comparisons for future studies.

5.1 Features

We tested the models with several n-gram combinations. 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.

5.2 Evaluation

All of the models were evaluated on how well they predicted the course sentiment labels, which were 'positive', 'negative' and 'neutral'. We used several metrics: 1) Overall accuracy predicting labels 2) F1-Score for each of the labels and 3) Cohen's Kappa (weighted). Because the base rate for neutral was high in our dataset, the Kappa metric probably gives the best overall measure of the performance of these models, correcting for chance agreement on neutral ratings. Although accuracy is reported in the table, we feel that Kappa is a better metric because in our dataset, an accuracy of .59 could be achieved by guessing the majority class.

All models were tuned against the development set. Once the final hyperparameters were selected for each model, they were trained on both the training and development set and run once against the test set.

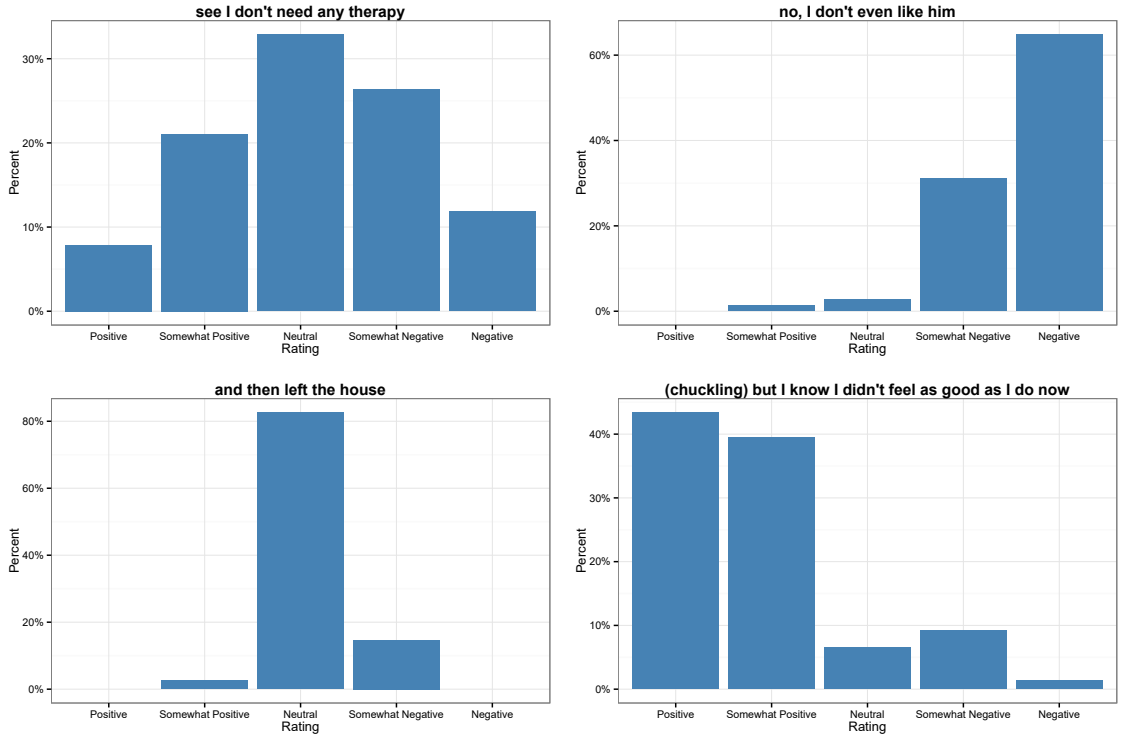

Figure 4: Examples Rating Distributions from Interrater Set

6 Classifier Model

We tested each of the feature sets with a Maximum Entropy Model. The L1 regularization on each of the models was tuned against the development set. The model functioned as local classifier (that is, they could only see each utterance in isolation, without any surrounding information from the session from which it was drawn).

In addition, we tested a pre-trained version of the Stanford sentiment model based on a Recursive Neural Network as a comparison for how a model that was trained on movie reviews would perform on psychotherapy dialogue (Socher et al., 2013). Because of the way that this model is set up to learn, it was not possible to train it on our data[1]. The RNN model from the previously mentioned paper requires parse trees of the training set, labeled at each node. Our training dataset only has the top level of the sentence labeled. In our tables, the specific model is identified as a Recursive Neural Tensor Network (RNTN).

[1]As a side note: this is not a limitation of Recursive Neural Networks (RNN) in general, but rather the way that Socher's implementation was designed to learn. The movie dataset that their group created labeled all of the sections of a parse tree and gave these labeled tree structures to the RNN. One could have designed an RNN to learn from just the top label of a tree, but then one would have to use a different implementation of an RNN. It may also surprise readers to learn that Socher's model in this paper relied on the Stanford parser to pre-parse the sentence trees, instead of letting the RNN parse the sentence.

7 Results

In general, we found that the n-gram models trained on this dataset had similar accuracy on the categorization of the course sentiment rating, but varied to a large degree in their F1 scores and Kappa statistics (see table 1). The maxent trigram model had the best overall accuracy, but by a relatively small margin.

The best F1 for positive statements was from the maxent unigram and bigram models and the best F1 for negative statements was from the maxent model as well. The maxent model had the highest Kappa score. Surprisingly, there was not a wide divergence in scores by the length of the grams in the models. The Kappa score for the maxent model did not change by more than .01 between a unigram and a tri-gram model.

The RNTN from Socher, et al. (2013) that was

Model	Accuracy	F1-Neutral	F1-Positive	F1-Negative	Kappa
Unigram Features					
Maxent	.601	.706	**.339**	**.451**	**.308**
Bigram Features					
Maxent	.603	.709	**.339**	.446	.306
Trigram Features					
Maxent	**.606**	**.714**	.337	.434	.300
RNN					
RNTN Trained on Movie Reviews	.484	.559	.319	.450	.227

Table 1: Results Test Set. Best scores for each category are bolded.

only pretrained on the movie review data had much lower accuracy than the other models (.484) and a lower Kappa score than the maxent models. The F1 scores for positive and negative statements were comparable to the best models, but the F1 score for negative was lower than any of the other models tested (.559).

In table 2 we present the best predictors of the positive and negative classes from the unigram maxent model. It is interesting to note that these words give some insight into why it is important to have a sentiment dataset that is specific to psychotherapy. We can see that 'scary' is one of the top ten negative words in the dataset. We should note that in a movie review, the word 'scary' might be a positive indicator. Additionally, psychologically relevant words are frequent on the list of good predictors like 'depressed' and 'relaxed' .

The confusion matrix for the maxent unigram model (see figure 5) shows that the basic model is generally accurate in the polarity of the statement (that is, there are very few errors of positive sentences coded as negative, or negative sentences coded as positive). The errors are generally classifying a positive utterance as neutral or a neutral utterance as positive.

8 Discussion

Psychotherapy is an often emotional process, and most theories of psychotherapy involve hypotheses about emotional expression, but very few researchers have systematically explored how affect works empirically in these situations. There are several databases of sentiment ratings in text but few of them involve dialogue and none are from a mental health setting. This dataset represents an initial step towards the study of sentiment in psychotherapy.

Most Positive Words
nice
thank, amazing
glad, good
proud, great
relaxed, helpful
fine, interesting
forward, helped
special, helps
cool, better
enjoyed, excited
Most Negative Words
sad, crap
hated, screwed
afraid, terrible
fear, can't
bothers, rejection
worst, death
hard, scary
horrible, worse
stupid, ugly
pissed, depressed

Table 2: Most Positive and Negative Words from Maxent Unigram Model

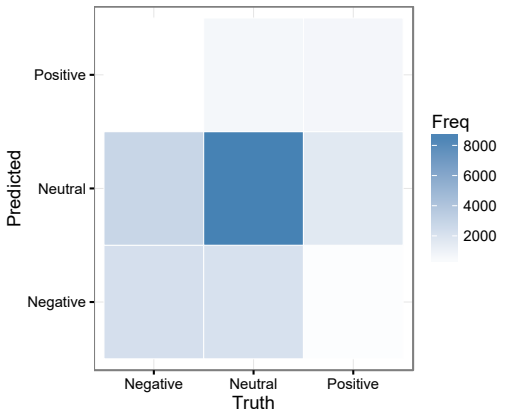

Figure 5: Confusion Matrix for Maxent Unigram Model (Test Set)

One of the novel contributions of this dataset to the area of sentiment analysis in general is the interrater reliability subset. In our literature search we were unable to find examples where researchers had estimated what human agreement was on different datasets used for sentiment analysis. This will allow us to compare our models to human-human agreement and also provide a qualitative sense for what kinds of utterances humans agree on and on which ones they disagree.

We hope that the creation of this dataset will improve researchers' ability to predict sentiment from dialogue and in psychotherapy settings. It is clear from the interrater reliability dataset that we should not expect models to perfectly rate sentiment because even humans do not completely agree on many types of utterances. However, it may be reasonable for machine rated reliability to approach the human range of reliability.

The models suggested by this paper are not intended to be a comprehensive list of models that may work well on the dataset, but are intended to be a baseline for other work to compare to. There is a long list of possible models that should be tried in the future, including LIWC counts, LDA models, word vectors and more comprehensive tests with Recursive Neural Networks. Testing all of these models and their variations is beyond the scope of this paper, but we hope that this dataset will give a baseline for different groups to test what works well in this type of data.

One of the interesting findings of this paper was the comparison of the RNTN model from Socher et al. (2013) that was only trained on the original movie review data. This dataset, it should be noted, is much larger than our own, but it is from a very different context. This is consistent with the conclusions of Pang and Lee (2008) that context is extremely important in identifying sentiment. Our work provides a test of the viability of domain adaption of models trained on very different datasets. It would appear accuracy will suffer if we use models that were trained on datasets like movie reviews and apply them directly to mental health contexts.

Finally, there were not substantial differences in accuracy between unigram models and the bigram, trigram models, suggesting that the more complex word patterns to not necessarily improve accuracy. This may be a side effect of the characteristics of dialogue, which are not always as gramatically clear as written text.

8.1 Psychotherapy Compared to Other Sentiment Domains

It may be surprising that the accuracy of some of our initial models are lower than other similar models used on other sentiment datasets. Part of this is may be a result of our decision to not use extremely short phrases (our dataset has a large number of neutral listening utterances like 'mm-hmm' and 'yeah' that we wanted to exclude). It should be noted that even in Socher et al. (2013) all of the models tested had an accuracy below .6 on anything that had 5 words or more (see figure 6 in their paper).

However, there may be a larger issue in the psychotherapy domain that makes labeling these utterances more difficult in general. For example, when rating movies, the typical subject of the sentence is going to be the movie and whether or not the reviewer enjoyed it. While you may have sentences that express both positive and negative attitudes, but there is some sense that the purpose is always going to be to evaluate the movie. In psychotherapy, an utterance like "(chuckling) but I know I didn't feel as good as now" has a complicated temporal aspect to it. The rater may be confused about whether this should be positive because the person feels good now, or negative because they were not feeling good prior to now. An utterance like "see I don't need any therapy" is complicated because some raters may see this as a person in recovery and others may see a person in denial.

Consequently, our models may not necessarily be evaluating how a person is *feeling* about another person or themselves in a given moment. Instead raters evaluated the emotional valence of a statement which could target the speaker, another person, or something unspecified. The psychotherapy domain clearly presents a more complicated task than answering the question "is this movie review a positive one or a negative one?" which is a better defined tasks. Future work may attempt more challenging classification tasks like asking a rater to guess how a client or therapist may be feeling from text - similar to how a human interacting with a client or therapist might attempt to understand their partners inter-

nal state. However, even if models could be trained to accurately capture this particular aspect of sentiment, we could not be sure that models were capturing an actual internal state. Instead they would be learning human perception of this state, which in and of itself can be error prone.

8.2 Future Directions

Beyond the practical question of whether we can accurately rate sentiment in psychotherapy, we hope that models trained on this dataset will eventually be able to code entire psychotherapy sessions so that we can ask larger questions about how sentiment expressed by clients and therapists influences outcomes. For example, would we expect to see the largest improvement in symptoms from positive client expression or negative client expression? Or should there be a pattern from negative expressed sentiment to positive? Another important question is whether we would see the most improvement from therapists who focus on positive aspects of a client's experience or more negative ones. To answer these questions, we need to be able to label more data than is practical to do with just human raters.

Acknowledgments

We would like to thank Padhraic Smyth for his helpful thoughts on early versions of this manuscript. Funding for this project was provided by NIH/NIDA R34 DA034860.

References

Timothy Anderson, Edward Bein, Brian Pinnell, and Hans Strupp. 1999. Linguistic analysis of affective speech in psychotherapy: A case grammar approach. *Psychotherapy research*, 9(1):88–99.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Reuven Bar-On, Daniel Tranel, Natalie L Denburg, and Antoine Bechara. 2004. Emotional and social intelligence. *Social neuroscience: key readings*, 223.

Ernst G Beier and David M Young. 1998. *The silent language of psychotherapy*. Aldine.

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

Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.*, 6(4):284–290.

J. Decety and W. Ickes. 2009. *The social neuroscience of empathy*. MIT Press, Cambridge, MA.

R. Elliott, A. C. Bohart, J. C. Watson, and L. S. Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43.

Albert Ellis. 1962. *Reason and emotion in psychotherapy.* Lyle Stuart.

Sigmund Freud and Josef Breuer. 1895. Studies on hysteria. se, 2. *London: Hogarth*.

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

James J Gross and Robert W Levenson. 1997. Hiding feelings: the acute effects of inhibiting negative and positive emotion. *Journal of abnormal psychology*, 106(1):95.

Ernest A Haggard and Kenneth S Isaacs. 1966. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*, pages 154–165. Springer.

Z. E. Imel, J. S. Barco, H. J. Brown, B. R. Baucom, J. S. Baer, J. C. Kircher, and D. C. Atkins. 2014a. The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of counseling psychology*, 61(1):146.

Zac E Imel, Mark Steyvers, and David C Atkins. 2014b. Psychotherapy Computational Psychotherapy Research : Scaling up the Evaluation of Patient Provider Interactions Computational. *Psychotherapy*.

Alice M Isen. 2008. Some ways in which positive affect influences decision making and problem solving. *Handbook of emotions*, 3:548–573.

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

Richard D Lane, Lee Ryan, Lynn Nadel, and Leslie Greenberg. 2015. Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, 38:e1.

Peter J Lang and Margaret M Bradley. 2010. Emotion and the motivational brain. *Biological psychology*, 84(3):437–450.

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

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics.

Erhard Mergenthaler. 1996. Emotion–abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of consulting and clinical psychology*, 64(6):1306.

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

Walter Mischel. 2013. *Personality and assessment*. Psychology Press.

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

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retreival*, 1(2):91–231.

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

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

Jeremy D Safran and J Christopher Muran. 2000. *Negotiating the therapeutic alliance: A relational treatment guide.* Guilford Press.

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

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.*, 86(2):420–428.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proc. Conf. Empir. Methods*.

Anastassios Stalikas. 1995. Client good moments: An intensive analysis of a single session anastassios stalikas marilyn fitzpatrick mcgill university. *Canadian Journal of Counselling*, 29:2.

N Yussupova, Diana Bogdanova, and M Boyko. 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*, pages 8–14.