Conversational Agents and Mental Health: Theory-Informed Assessment of Language and Affect

Adam Miner
Stanford University
adam.miner@gmail.com

Amanda Chow
Stanford University
amdchow@stanford.edu

Sarah Adler
Stanford University
sadler1@stanford.edu

Ilia Zaitsev
Existor Limited
ilia@existor.com

Paul Tero
Existor Limited
paul@existor.com

Alison Darcy
Stanford University
adarcy@stanford.edu

Andreas Paepcke
Stanford University
paepcke@cs.stanford.edu

ABSTRACT

A study deployed the mental health Relational Frame Theory as grounding for an analysis of sentiment dynamics in human-language dialogs. The work takes a step towards enabling use of conversational agents in mental health settings. Sentiment tendencies and mirroring behaviors in 11k human-human dialogs were compared with behaviors when humans interacted with conversational agents in a similar-sized collection. The study finds that human sentiment-related interaction norms persist in human-agent dialogs, but that humans are twice as likely to respond negatively when faced with a negative utterance by a robot than in a comparable situation with humans. Similarly, inhibition towards use of obscenity is greatly reduced. We introduce a new Affective Neural Net implementation that specializes in analyzing sentiment in real time.

ACM Classification Keywords

I.2.11 Distributed Artificial Intelligence: Intelligent Agents; J.4 Computer Applications: Social and Behavioral Sciences — Psychology

Author Keywords

psychotherapy; sentiment analysis; relational frame theory; conversational agents; neural sentiment model

INTRODUCTION

A longstanding goal of artificial intelligence has been the automation of interpersonal conversations between humans and artificial entities. This objective is particularly relevant in psychotherapy, as language is a primary tool to understand patients' experiences and express therapeutic interventions [31]. Although conversational agents have been developed for use in clinical settings [3], they are not deployed widely, and there is little research assessing the potential impact of this type of tool on mental health [28][1].

Conversational agents are software systems that receive human-language inputs via a medium such as the internet, and offer human-language responses to those inputs. The responses are generally delivered over the medium from which the input was received. This class of system offers unique benefits to mental health, such as always being available, responding to context of both the user and the user's language, and the capacity to provide responses based on clinically relevant theories of mental health.

Conversational agents are one example of *Behavioral Intervention Technologie*s (BITS). These are behavioral or psychological interventions that utilize and communicate health-relevant information to address mental health processes and outcomes, and are a focus of national research agendas in mental health [28]. The use of technology-augmented care for mental health problems such as depression and anxiety are promising, using both internet-, and text message based facilities (e.g. [9, 21]). Randomized controlled trials have shown these technologies to be effective in decreasing clinically meaningful symptoms such as depressed mood. Self-guided, non-interactive therapies are also effective [25] and address barriers to care such as geography, cost, and absence of skilled clinicians. However, the materials are often static and do not respond dynamically to consumers' needs or personal experiences. Conversational agents promise to introduce dynamic interaction while retaining these advantages of non-interactive therapies.

Unfortunately, to date, few explorations in natural language processing (NLP) have been driven by evidence based men-

[1]Conversational agents are also called relational agents, virtual agents, or chatbots.

tal health theoretical models. Two-way connections between these disciplines are difficult to establish, partly because the respective research communities are disjointed, and partly because NLP capabilities are still far behind the linguistic and semantic sophistication level at which mental health theory operates. However, as conversational agents become more embedded into daily activities through facilities such as Siri on the iPhone, the need for understanding their impact on our behavior is gaining urgency. Mental health issues are particularly relevant, as research has shown that commonly available conversational agents on smartphones respond inconsistently and incompletely to mental health crises [27].

We report here on first steps towards forging connections between NLP and mental health insights. Guided by an evidence based theory of psychological distress, *Relational Frame Theory* (RFT), we used NLP technologies to analyze sentiment dynamics in both human-human and human-agent dialogs. We also report on a novel conversational agent that emphasizes sentiment recognition, and is therefore highly suited for experiments with mergers of mental health theory and conversational agent technology. Affective language has in fact been a research focus of both clinical psychology [15] and computer science [34, 17].

An extant body of literature demonstrates how NLP can be deployed to inform clinical practice (e.g. [13, 19]). In 2014, Manavinakru et al. used NLP to create health interventions by selecting patient stories based in part on positive affect, demonstrating the utility of assessing the valence of language to design interventions [24]. By assessing large conversational datasets, NLP has been effective in differentiating clinically meaningful text patterns in mental health crisis-focused conversations [1]. Additionally, emotional contagion has been demonstrated in social networking sites, demonstrating the role language can play in how people interact with others [22].

In order to gain adoption in the psychiatric setting, it is imperative that new tools for assessing language are grounded in clinically relevant, evidence-based theories that have been supported in the literature to inform practice [7].

CONVERSATIONAL AGENTS

While the full levels of response sophistication required of conversational agents may lie in the future, such agents show surprising promise for mental health intervention. Here is a list of reasons.

First, users interact with software following social rules even when they are aware of their interaction partner being software [33], suggesting conversations between humans and software may be relevant to interpersonal interactions generally. Second, people form relationships with software, even knowing their non-human nature [2, 36]. Conversational agents have been shown to be acceptable in certain populations (e.g. [6]), even though use in clinically relevant populations is limited [16]. Thus it is plausible that important aspects of human-to-human relationships can be studied through the assessment of human-to-computer interactions. Lastly, as mentioned, conversational agents may address unique barriers endemic to mental health assessment and treatment such as geographic barriers to care; lack of quick, asynchronous communication; continuous availability; and stigma around mental health. Using accepted psychological theory to guide understanding of conversations is a significant first step towards further development and refinement of clinically useful tools.

We next very briefly introduce Relational Frame Theory as an approach we use to assess the role of affect in language. The theory calls for such affect-related assessment, but RFT additionally generates a number of important challenges for NLP that we do not have space to lay out here.

RELATIONAL FRAME THEORY

Clinical psychology uses learning paradigms where interactions between physiological sensation (affect), language as an internal process (cognition), and behavior underlie or maintain maladaptive symptoms such as depressed mood or excessive worry [15]. These theories are used to guide clinical conceptualizations and understand pathological symptoms, guiding towards treatment targets, usually in the form of talk therapy. Although there are many evidence based theories in the field, we focus on Relational Frame Theory (RFT) which posits that humans use linguistic frames to understand the world around them, and subsequently solve problems. RFT has been suggested as an approach to understanding natural language systems [14]. The theory lends itself well to assessment with NLP precisely because it relies on understanding interaction between sensation, affect, language, and behavior.

RFT articulates three *frames* that humans use for problem solving. First, *events and attributions*; second, *time and contingency*; and third, *comparison and evaluation*. This paper focuses on the first frame: events and attribution. Although each of these frames is a necessary part of understanding both human capacity for problem solving and distress, our initial analyses focus on identifying verbal representations associated with frame 1: events and their attributes. When someone uses language, they are labeling their experience. For example, someone might report "This carnival ride is scary!" indicating a fear based affective response. We are using NLP to assess this label, as the labeling of language to understand psychologically meaningful constructs such as affect have been previously demonstrated [30] and applied to better understand the role of affect in conversation.

Affect is physiological sensation that guides learning when language is used to create attributions or labels. For example, in anxiety disorders, interoceptive sensations such as autonomic nervous system responses (e.g. increased heart rate, rapid breathing, and gastrointestinal distress) are misinterpreted by the brain signaling danger despite lack of inherent threat. The cognitive attribution of the physiological symptoms can directly lead to avoidance of environments or conditions that lead to such sensations, creating an inner-directed feedback loop that maintains maladaptive processes. RFT is one explanation for why patterns of this type become overlearned in human beings. RFT posits a behavior-analytic framework describing these patterns as a special form of relational responding unique to humans, where language holds an abstract and symbolic role in governing human behavior: solving problems. This type of verbal problem solving has made human

beings into the dominant species, despite not being fastest, strongest, or most adaptable, but is also hypothesized to maintain psychological distress.

Simply labeling events and their attributes as 'positive' or 'negative' increases associated memories and emotional salience. This type of relational network can be evoked with any number of internal or external stimuli, triggering the aforementioned internal feedback loop, and leading to psychological distress. For example, describing a 'negative' event such as a trauma can evoke intense fear and sadness and subsequent sobbing. The very description of the event as 'negative' can be used as a probable proxy for psychological distress. If primary relational frames such as 'positive' and 'negative' utterances can be systematically identified and understood in conversations with machines, conversational agents become a viable tool in clinical and other settings. The following section examines how relational frames and NLP are connected. We will use NLP to extract from dialogs psychological insight that can begin to suggest useful conversational agent responses.

SENTIMENT CONNECTS RFT TO NLP

The NLP task known as *sentiment analysis* is closely related to the *events and attributions* frame. Sentiment analysis has been approached both based on observation of affective word occurrences, and machine learning approaches. For example, Ding et al. combined rules and word affect orientations to determine sentiment in product reviews [8]. Liu et al. also argue that multiple approaches to sentiment analysis are required [23]. However, word- and rule-based approaches are complex. The authors explain, for example, that an understanding of several special linguistic cases is needed for robust sentiment labeling: a comparative sentence, such as *The brand-X computer is much faster than brand-Y* implies positive sentiment towards brand-X, even though no typical sentiment-related word is present [12, 20].

Similarly, modal sentences such as *I wish my husband were the outdoorsy type,* and *I think you should have done the dishes* harbor sentiment, which [23] attempts to extract via a combination of rules and machine learning.

Note that the techniques in [12, 20] are relevant to both the RFT *events and attributions* and *comparison and evaluation* frames, although we do not have space to introduce the latter. The papers demonstrate that sentiment labeling and comparison are related.

We therefore turn now to an in-depth analysis of how sentiment dynamics manifest in conversations. This examination illustrates how NLP can be deployed to assess affect and psychological distress, and more generally how psychological theory can guide the development of clinically applicable tools. We will in particular focus on how humans mirror sentiment, or in psychological terms, reflective and validating language. Awareness of affect mirroring is a key construct in successful therapeutic interactions [29].

Table 1 organizes this portion of our contribution. For example, cell *human-human* refers to human mirroring behavior when a human conversation partner offers a positive-sentiment, or negative-sentiment utterance. (*mirror++/mirror–*).

Table 1. Our behavioral explorations of dialogs between humans and conversational agents. We focused on actions initiated by humans. The lower-left cell thus refers to human→robot interaction, not vice versa.

	Dialog Partner	
	Human	Robot
Human	mirror++ mirror–	Frequent obscenity
Robot	mirror++ mirror–	

We compare this human-human mirroring behavior with human behavior when communicating with a conversational agent (cell *robot-human*). In this context, we are not interested in how any particular conversational agent mirrors stimuli by humans (upper right cell). We will, however, offer observations about the content that humans feed to conversational agents. In an effort to ensure validity, we will examine mirroring behaviors with several datasets, and with two methods of measuring sentiment — a keyword-based approach, and a sentiment-trained recurrent neural net, which we call *Affective Neural Model*.

After the examination of mirroring behavior, we will describe our Affective Neural Model, which generates sentiment analysis in real time, while a stimulus is being typed. The system is in operation, and classifies sentiment into seven types of emotions.

Human-Human Communication

Understanding patterns of human-human interaction is a natural baseline to examining human-computer interactions. Reflective and validating communication is an almost universally accepted communication strategy in therapy [32]. With reflective listening, a therapist displays empathy, or a reflection of sentiment back to the patient to demonstrate understanding, which results in connection and engagement. This is an iterative process where the therapist must adapt to the sequential response and continue to adopt an empathic stance. However, what makes a therapeutically successful conversation may be dramatically different from a non-therapeutic conversation.

We analyzed the Fisher English Training transcript collection of 11,600 telephone conversations between human participants (corpus Fisher11k). The conversations lasted up to 10 minutes, and were each to cover one of 40 randomly assigned topics [5].

We used VADER [18] to classify each conversation turn in Fisher11k into *positive*, *negative*, and *neutral*. VADER uses five grammatical and syntactical rules, and a lexicon that extends the *Linguistic Inquiry and Word Count* (LIWC) lexicon to cover micro-blogs. For example, the lexicon includes emoticons, which are irrelevant for the Fisher collection telephone transcripts, but will be important in later sections.

For each pair of Fisher utterances we determined the valences of the first speech turn (i.e. a stimulus) and the second turn

(i.e. the conversation partner's response). From the resulting counts we constructed valence-response plots (Figure 1).

Figure 1 shows stimulus sentiment along the abscissas, and response sentiments along the ordinal axes. The left-most chart summarizes all sentiments in conversations around a cheerful topic: *An Anonymous Benefactor*. The middle chart is the result of a conversation about *Terrorism*, while the right-side graph corresponds to a conversation about *Foreign Relations*.

In computing these figures we only considered the first 80 utterances. Since conversations were not monitored, they tended to start with the assigned topic, but then drifted.

Points in the left half of each plot contain all negative stimuli; all positive stimuli are contained in the right half of the plots. Consequently, the upper-left quadrant holds points in which a positive-sentiment reply was offered to a negative stimulus. The upper-right quadrant shows the cases where a positive response was given to a positive stimulus, and so on. In all conversations, more positive-sentiment statements occurred on either side of the conversation than contributions with negative valence. This tendency was of course most notable for the upbeat topics.

We computed the probabilities that a conversation partner would mirror a negative or positive stimulus:

$$P_{++} = P(R_{pos}|S_{pos}) \quad (1)$$
$$P_{--} = P(R_{neg}|S_{neg}) \quad (2)$$
$$P_{-+} = 1 - P_{++} \quad (3)$$
$$P_{+-} = 1 - P_{--} \quad (4)$$

Equation 1 shows the definition of P_{++}. The quantity is the probability of a positive response (R_{pos}) given a positive stimulus (S_{pos}) by the conversation partner. The remaining equations are analogous.

On average across all conversations, the probability that a positive stimulus drew a positive response was $0.84(SD = 0.06)$. For negative stimuli the probability of mirroring, i.e. evoking a negative response was just $0.22(SD = 0.05)$.

The benefactor chart in Figure 1 shows a strong preponderance of positive-sentiment utterances in the conversations. All sentiment distributions of uplifting topics looked like this chart. The median of Pearson correlations between the sentiment of a stimulus, and the sentiment of the resulting response among the benefactor conversations was $r = 0.26, p < 0.05$. The maximum correlation among those conversations was $r = 0.50$. These numbers are typical for all conversations.

In summary, the Fisher conversations show a very strong tendency for participants to formulate positive-sentiment statements. Upon encountering negative statements, the participants showed a consistent tendency towards moving the conversations in a positive direction. Interestingly, this observation may identify a pattern that would not be clinically useful and could differentiate between non-therapeutic and therapeutic interactions.

Human-Agent Communication

We saw that among the Fisher participants humans tend to reply with positive sentiments when in a generally non-combative context, even when confronted with negative stimuli. This trend may, of course, be different in an already negatively charged situation. We did not examine those scenarios.

Our next question was the sentiment mirroring behavior that humans would bring to conversational agents. Would one conversation partner being non-human change behaviors? For validation of our findings we conducted two independent human-agent experiments, using very different sentiment analysis methods.

The first experiment used the VADER software package for determining sentiment, as per the human-human experiment described earlier. The dataset consisted of 5000 conversations with the Cleverbot [35] conversational agent(corpus CB5k).

Visitors to the Cleverbot site are free to conduct any conversation they like. No topic prompts are offered as in the Fisher collection, nor were conversation lengths recommended. Consequently, conversation durations varied. In many cases, visitors clearly experimented to explore the conversational agent's behavior. Many—sometimes long—conversations consisted of obscenities and insults patiently delivered to the conversational agent across what must have amounted to a considerable span of time. In fact, 97% of the visitor utterances in CB5k contained at least one swear word. Many expletives were obfuscated by the use of special characters in place of letters. VADER did not recognize these tokens, and classified them as neutral. For simplicity we converted these obfuscations into a single swear word that VADER already understood. Note that the presence of a swear word in a sentence does not necessary produce a negative-sentiment classification for the sentence as a whole.

Figure 2 shows the tendency of human Cleverbot site visitors to respond positively to the conversational agent. The chart shows conversational agent sentiment along the abscissa, and human-utterance sentiment along the ordinal. The pronounced horizontal and vertical data cluster lines are utterances that VADER classified as neutral. Note again the accumulation of positive-sentiment user statements in the upper-right quadrant.

The sentiment mirroring behaviors do carry over from what we observed in the human-human Fisher context. However, the effect sizes change significantly when humans correspond with Cleverbot. The probability of the human conversation partner mirroring a positive Cleverbot statement with an also positive response was 0.75. The probability of mirroring a negative statement was 0.41. Table 2 summarizes the comparison. Humans are twice as likely to respond with a negatively valenced reply when confronted with negative robot sentiment than when conversing with a human partner.

In a health care related context, this difference needs to be considered. We therefore explored sentiment behavior towards conversational agents further, using a larger dataset and a very different approach to measurement. The sentiment measurements above were obtained using standard methods. The

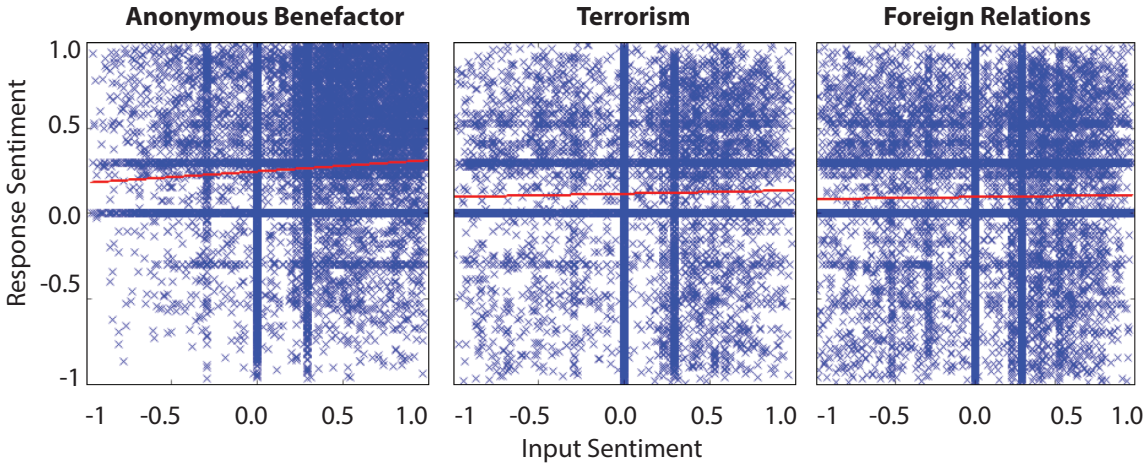

Figure 1. Tendency of sentiment for all conversations within three topics. A positive trend held for all 40 conversations.

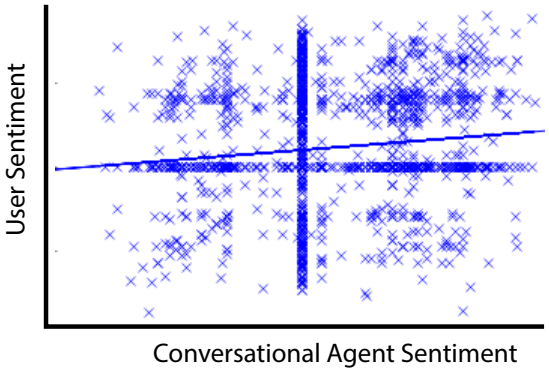

Figure 2. Site visitors tended to respond with positive sentiment to conversational agent stimuli (CB5k dataset). This tendency carries over from human-human behaviors.

following section describes a new version of the Cleverbot software that specializes in sentiment recognition performed in real time. For this **second human-robot experiment** we used 10,000 additional conversations conducted by site visitors to cleverbot.com and eviebot.com (corpus CB10k). These conversations were analyzed for sentiment using our Affective Neural Model (ANM). The model is a recurrent neural net (RNN). We discuss the details of this model and its implementation in Section 5. The model was trained using emoticons

Table 2. Summary of behavior differences when humans interact with another human as opposed to a robot.

Dialog Partner		
	Human	Robot
Human	mirror++: 0.84 mirror–: 0.22	Frequent obscenity
Robot	mirror++: 0.75 mirror–: 0.41	

that human partners embedded in their utterances to the system.

We mapped over 300 emoticons and other signals in CB10k to seven different affects. We chose these affects as an amalgamation of three sources. First, we included the manually observed emotions generally exhibited by Cleverbot users. Second, we included the emotion set recently suggested by Facebook, which includes *like*, *love*, *haha*, *wow*, *sad*, and *angry*. Finally, we included the six basic emotions suggested by Paul Ekman's studies in 1984: happiness, surprise, disgust, anger, fear and sadness [10].

Figure 3 shows two screen shots taken while typing a sentence to the ANM. The sentiment changes from predominantly sad to happy by the completion of the sentence. The change illustrates the complexity of sarcasm, with its mixture of multiple emotions. The complete sentence of course expresses anger, but the clever sarcasm also brings up mirth. The algorithm leaned towards the positive.

THE AFFECTIVE NEURAL MODEL

In total we gathered 2,778,737 data points (*angry*: 421847, *surprise*: 249859, *happy*: 573422, *love*: 579476, *sad*: 422172, *disgust*: 25701, *laughter*: 506260). The dataset's vocabulary size (number of unique words) is 177,000 words.

Our model operates as a predictive model for the affects given a pair of dialogue lines (something Cleverbot said followed by the user's reply) and their affect labels. Our RNN deploys two Gated Recurrent Unit (GRU) [4] layers and a Softmax activation function on the output layer. The RNN models a probability distribution across the 7 affect labels conditioned on the given sequence of words. The probability of a given affect (such as *happy*) following a sequence of n words $w_1, w_2, ...w_n$ is approximated as:

$$P(affect|w_1, w_2...w_n) \approx P(affect|w_1) * P(affect|w_2) * ... * P(affect|w_n)$$

Figure 3. Predominant emotion changes as "I never forget a face" is expanded to "but in your case I'll be glad to make an exception."

The training process is very similar to RNN language models, but instead of predicting next words given previous ones our model tries to predict one of the 7 affects at the end of a given sequence of words, approximating the equation above. The sequence length (number of input words) is fixed at 30 words. Any sequences shorter than this use a special <pad> word to bring them up to 30.

A standard RNN layer has two sets of weights connecting it to the previous layer and to itself. A GRU instead uses six sets of weights to enable the layer to remember previous values. We therefore extended the context of each word beyond the single-vector history of regular RNNs.

The equations for a GRU are as follows. The inputs from the previous layer are given as x_t and the recurrent inputs as h_{t-1}. The update gate u multiplies these two sets of inputs by two corresponding sets of weights to determine whether the GRU should keep its previous values or update to a new one. The reset gate performs a similar function to determine which of the previous values should be reset/ignored. The third equation computes a new potential value to remember as m. The final equation produces the unit's output h_t either from the new potential value m or the previous value:

$$
\begin{aligned}
&h_{t-1}: \\
&u = \sigma\left(x_t W_u + h^{t-1} U_u\right) \\
&r = \sigma\left(x_t W_r + h^{t-1} U_r\right) \\
&m = \tanh\left(x_t W_m + (r \odot h_{t-1}) U_m\right) \\
&h_t = u \odot m + (1-u) \odot h_{t-1}
\end{aligned}
$$

The weights between the input one-hot vectors and the first GRU layer in the network are commonly known as *word embeddings*. They can be treated as vector representations of each word. Words with similar affective value (e.g. *happy* and *wonderful*) are close to each other in the vector space. This is analogous to word embeddings produced by algorithms such as word2vec [26], where words are grouped together by semantic value. We also used dropout between the two GRU layers [11] as a form of regularization. This procedure randomly sets some values to zeros, adding artificial noise to the network to prevent overfitting. We separated 10% of the data points into a test set.

After training our model on 90% of the nearly 3 million data points, we then tested on the remaining 10% containing 277,874 data points. On this test set, the model shows 90% accuracy. In other words, our model correctly predicted the overriding affect for 90% of unseen lines of conversation. If we then type *I'm depressed* into our model, it guesses that we feel sad (with a probability of 67%).

Sentiment measurements of the human contributions again confirm the preponderance of positive sentiment (Figure 4).

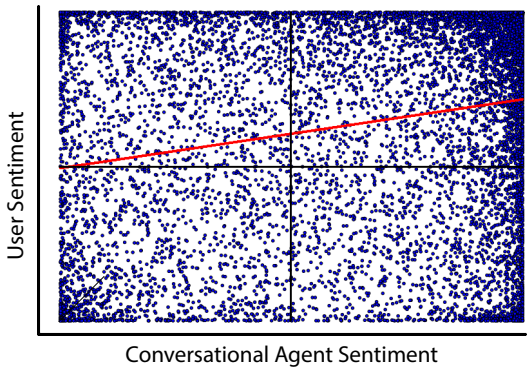

Figure 4. Site visitor vs. ANM sentiment (RNN10k dataset). Note again the upward tendency.

Limitations

Our study is limited along the following dimensions. First, we did not assess language that occurred in mental health focused conversations, but a mixture of conversations about upbeat and sad topics. Second, the use of positively or negatively labeled words does not necessarily map onto the user's emotional state [31], limiting the generalizability of this approach. Finally, the Fisher collection consists of phone conversation transcripts, while the CBxk collections were conducted over the Internet. Further study is needed to determine the extent to which this difference is a confound in the comparison between human-human and human-agent behavior. For example, it is possible that for younger generations, affective differences between typed (e.g. texted) and spoken communication have begun to fade.

CONCLUSION

Our work extends the role of NLP by contextualizing the assessment of sentiment within a clinically relevant theory of language and mental health (RFT). Through the creation of NLP tools grounded within accepted psychological theory, conversational agents may be developed to assess real time features of a user's conversation. These in turn inform mental health screening and treatment. Such information is useful for both assessment, and measurement of treatment impact on real world functioning. For example, RFT is a way to understand how cognitions interact with behavior to produce suffering. RFT articulates why people get stuck in maladaptive patterns, and NLP allows us to systematically identify these patterns in naturally occurring conversations.

Automating the evaluation of psychological symptoms could impact mental health by increasing access to services, and potentially decreasing costs associated with non-treatment or traditional clinical models. The development of frameworks for identifying affect-relevant labels may improve screening for maladaptive patterns of language that cause distress, and provide targets for intervention. Designing NLP-based approaches that assess clinically relevant language would allow for the design of conversational agents that are scalable, transparent in their labeling of language, and responsive to patient and language features.

Our study shows that human sentiment conventions carry over to human-agent interactions, but that those norms are weakened. Humans are twice as likely to respond with a negatively valenced reply when confronted with negative robot sentiment than when conversing with a human partner. This finding may have important implications for the design of agents' reactions to human utterances: the a priori lowered sense of obligation towards 'niceness' must be taken into account.

Future work should assess the role of language in the additional two RFT components (*time and contingency*, and *comparison and evaluation*). Our finding that users mirror more positive language than negative language is consistent with previous research showing asynchronous conversational responses to positive and negative language (e.g. [22]). This is significant as it yields greater understanding of the differences between human-to-human and human-to-computer interactions that can inform the design of clinically focused tools. Also, language is only one way humans express and understand emotions. Integrating NLP-based approaches with other areas of affective computing (e.g. gaze, facial expression) and additional behavioral and social interactions will benefit the design of mental health focused conversational agents.

REFERENCES

1. Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Natural Language Processing for Mental Health: Large Scale Discourse Analysis of Counseling Conversations. *Transactions of the Association for Computational Linguistics* (2016).

2. Timothy Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling* 59, 1 (2005), 21–30.

3. Timothy W Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient education and counseling* 92, 2 (2013), 142–148.

4. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arxiv.org* (12 2014). http://arxiv.org/abs/1412.3555

5. Christopher Cieri and et al. 2004/2005. Fisher English Training Speech Part 1/2 Transcripts LDC2004T19. Linguistic Data Consortium. (2004/2005). https://catalog.ldc.upenn.edu/LDC2004T19

6. Rik Crutzen, Gjalt-Jorn Y Peters, Sarah Dias Portugal, Erwin M Fisser, and Jorne J Grolleman. 2011. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study. *Journal of Adolescent Health* 48, 5 (2011), 514–519.

7. Alison M Darcy, Alan K Louie, and Laura Weiss Roberts. 2016. Machine Learning and the Profession of Medicine. *JAMA* 315, 6 (2016), 551–552.

8. Xiaowen Ding and Bing Liu. 2007. The Utility of Linguistic Rules in Opinion Mining. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 811–812.

9. David Daniel Ebert, Anna-Carlotta Zarski, Helen Christensen, Yvonne Stikkelbroek, Pim Cuijpers, Matthias Berking, and Heleen Riper. 2015. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. *PloS one* 10, 3 (2015), e0119895.

10. Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to emotion* 3 (1984), 19–344.

11. Yarin Gal. 2015. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *arxiv.org* (12 2015). http://arxiv.org/abs/1512.05287

12. Murthy Ganapathibhotla and Bing Liu. 2008. Mining Opinions in Comparative Sentences. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1 (COLING '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 241–248. http://dl.acm.org/citation.cfm?id=1599081.1599112

13. Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. 2013. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ quality & safety* (2013), bmjqs–2012.

14. David E Greenway, Emily K Sandoz, and David R Perkins. 2010. Potential applications of relational frame theory to natural language systems. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, Vol. 6. IEEE, 2955–2958.

15. Steven C Hayes, Dermot Barnes-Holmes, and Bryan Roche. 2001. *Relational frame theory: A post-Skinnerian account of human language and cognition*. Springer Science & Business Media.

16. Jing Huang, Qi Li, Yuanyuan Xue, Taoran Cheng, Shuangqing Xu, Jia Jia, and Ling Feng. 2015. Teenchat: a chatterbot system for sensing and releasing adolescentsâĂŹ stress. In *Health Information Science*. Springer, 133–145.

17. Eva Hudlicka. 2003. To feel or not to feel: The role of affect in human–computer interaction. *International journal of human-computer studies* 59, 1 (2003), 1–32.

18. C.J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. AAAI Publications, Ann Arbor, MI, 216–225.

19. Kristin N Javaras, Nan M Laird, Ted Reichborn-Kjennerud, Cynthia M Bulik, Harrison G Pope, and James I Hudson. 2008. Familiality and heritability of binge eating disorder: results of a case-control family study and a twin study. *International Journal of Eating Disorders* 41, 2 (2008), 174–179.

20. Nitin Jindal and Bing Liu. 2006. Identifying Comparative Sentences in Text Documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 244–251. DOI: http://dx.doi.org/10.1145/1148170.1148215

21. David Kessler, Glyn Lewis, Surinder Kaur, Nicola Wiles, Michael King, Scott Weich, Debbie J Sharp, Ricardo Araya, Sandra Hollinghurst, and Tim J Peters. 2009. Therapist-delivered internet psychotherapy for depression in primary care: a randomised controlled trial. *The Lancet* 374, 9690 (2009), 628–634.

22. Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

23. Yang Liu, Xiaohui Yu, Zhongshuai Chen, and Bing Liu. 2013. Sentiment Analysis of Sentences with Modalities. In *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing (UnstructureNLP '13)*. ACM, New York, NY, USA, 39–44. DOI: http://dx.doi.org/10.1145/2513549.2513556

24. Ramesh Manuvinakurike, Wayne F Velicer, and Timothy W Bickmore. 2014. Automated indexing of Internet stories for health behavior change: weight loss attitude pilot study. *Journal of medical Internet research* 16, 12 (2014).

25. Evan Mayo-Wilson and Paul Montgomery. 2013. Media-delivered cognitive behavioural therapy and behavioural therapy (self-help) for anxiety disorders in adults. *Cochrane Database Syst Rev* 9 (2013), CD005330.

26. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119. http://arxiv.org/abs/1310.4546

27. Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* (2016).

28. David C Mohr, Michelle Nicole Burns, Stephen M Schueller, Gregory Clarke, and Michael Klinkman. 2013. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. *General hospital psychiatry* 35, 4 (2013), 332–338.

29. Theresa B Moyers and William R Miller. 2013. Is low therapist empathy toxic? *Psychology of Addictive Behaviors* 27, 3 (2013), 878.

30. James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works* (2015).

31. James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

32. Erik Rautalinko, Hans-Olof Lisper, and Bo Ekehammar. 2007. Reflective listening in counseling: effects of training time and evaluator social skills. *American journal of psychotherapy* 61, 2 (2007), 191.

33. Byron Reeves and Cliff Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press, NY.

34. Rainer Reisenzein, Eva Hudlicka, Mehdi Dastani, Jonathan Gratch, Koen Hindriks, Emiliano Lorini, and John-Jules Ch Meyer. 2013. Computational modeling of emotion: Toward improving the inter-and intradisciplinary exchange. *Affective Computing, IEEE Transactions on* 4, 3 (2013), 246–266.

35. Paul Tero, Ilia Zaitsev, and Rollo Carpenter. 2016. Cleverbot Data for Machine Learning.

http://www.existor.com/en/ml-cleverbot-data-for-machine-learning.html. (January 2016).

36. Adam Waytz, John Cacioppo, and Nicholas Epley. 2010. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science* 5, 3 (2010), 219–232.