Crisis counselor language and perceived genuine concern in crisis conversations

Greg Buda
Crisis Text Line

Ignacio J. Tripodi
Crisis Text Line

Margaret Meagher
Crisis Text Line

Elizabeth A. Olson
Crisis Text Line

Abstract

Although clients' perceptions of therapist empathy are known to correlate with therapy effectiveness, the specific ways that the therapist's language use contributes to perceived empathy remain less understood. Natural Language Processing techniques, such as transformer models, permit the quantitative, automated, and scalable analysis of therapists' verbal behaviors. Here, we present a novel approach to extract linguistic features from text-based crisis intervention transcripts to analyze associations between specific crisis counselor verbal behaviors and perceived genuine concern. Linguistic features associated with higher perceived genuine concern included positive emotional language and affirmations; features associated with lower perceived genuine concern included self-oriented talk and overuse of templates. These findings provide preliminary evidence toward pathways for automating real-time feedback to crisis counselors about clients' perception of the therapeutic relationship.

1 Introduction

In the context of mental health interventions, an extensive body of research has found significant associations between therapists' behavioral traits and clinical effectiveness. In particular, clients' perceptions of therapist empathy correlate with positive therapy outcome, as documented in various contexts, including addiction treatment, suicide intervention, and cognitive behavioral therapy for anxiety (Elliott et al., 2018; Moyers and Miller, 2013; Bryan et al., 2018; Hara et al., 2017).

Therapist empathy and genuineness are closely connected and are frequently evaluated together in relation to therapy outcomes (Shapiro, 1969). In an analysis of survey data from Crisis Text Line's text-based crisis interventions, Gould et al. (2022) found significant associations between texters' ratings of the overall effectiveness of the intervention and their perceptions of the counselor's genuine concern. Higher genuine concern ratings also were associated with greater texter reports of reduced suicidality over the course of the conversation.

Despite its relevance for therapy effectiveness, developing methods to recognize and improve perceived empathy remains technically demanding due to its complex, context-dependent nature (Cunningham et al., 2023). A few studies have explored how empathy manifests in therapists' language. However, many of these rely on qualitative conversation analyses (Wu, 2021;Wynn and Wynn, 2006), time-intensive manual annotation of data (Aafjes-van Doorn et al., 2020), or observer ratings rather than client feedback to gauge empathy (Lord et al., 2015). Recent advancements in natural language processing (NLP) offer a new avenue for quantitative, automated, and scalable analysis of therapists' verbal behaviors in relation to perceived empathy. Yet, academic investigation with computational approaches remains sparse and often depends on proprietary linguistic analysis tools like Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022), which do not offer the flexibility to extend beyond predefined word categories.

Leveraging a large pool of text-based crisis conversations, we employed NLP techniques to extract over 100 linguistic features and examine their relationship with texter-rated genuine concern in conversations with trained volunteer crisis counselors.

2 Data

We used data from Crisis Text Line, a non-profit organization that provides free, confidential mental health support and crisis intervention through text-based communication. All conversations are de-identified before analysis. Data are automatically scrubbed of identifying information prior to access by this research team. This research was evaluated by Sterling IRB, which issued an exempt determination.

Findings of the Association for Computational Linguistics: EACL 2024, pages 7149–7160
November 12-16, 2024 ©2024 Association for Computational Linguistics

After each conversation, texters complete an optional post-conversation survey where they evaluate crisis counselors' genuine concern. The survey item asks respondents to rate their agreement with the statement: *"In the conversation, I believe my Crisis Counselor was genuinely concerned for my well-being"*. Response options include *'Strongly agree', 'Somewhat agree', 'Neither agree nor disagree', 'Somewhat disagree'* and *'Strongly disagree'*. We combined multiple message-level and conversation-level data sources from Crisis Text Line's dataset to identify features associated with these ratings. The analysis was restricted to English-language conversations, excluding cases where the counselor indicated that the texter's intent was to prank or test the system. Additionally, because some texters have multiple conversations, we included the texters' first conversation, starting from 2017. We considered only conversations featuring at least five counselor *turns* (defined as concatenated consecutive messages sent by one conversational participant). 404,017 conversations met these criteria and had post-conversation genuine concern data between 2017 and 2023.

3 Methods

The primary goal of this analysis was to identify the relationship between crisis counselor (CC) linguistic cues and the perception of genuine concern by individuals in crisis. Features were selected a priori on the basis of a literature review regarding language efficacy in counseling relationships and existing psychotherapy-related NLP work (Aafjes-van Doorn et al., 2020; Atkins et al., 2014; Cunningham et al., 2023; Decker et al., 2014; Finset and Ørnes, 2017; Hasan et al., 2019; Lord et al., 2015; Miner et al., 2022; Wu, 2021; Wynn and Wynn, 2006; Xiao et al., 2016b,a), in consultation with clinical experts (EAO). We included many features related to alignment between the texter's and CC's language use, at the turn-by-turn conversational level (for a full list of features, see the Appendix). We used natural language processing (NLP) methods to extract linguistic features, including:

- Overall token counts and token counts by part-of-speech (POS)[1];
- Text complexity, measured via the Flesch-Kincaid reading level (Kincaid et al., 1975).;
- Formal/informal tone measured with the use frequency of specific parts-of-speech, following Heylighen and Dewaele (2002);
- Emotional valence (positivity or negativity of emotions), calculated using a pre-trained BERT model[2] (*bert-base-uncased* by Devlin et al. (2019)) fine-tuned on annotated datasets such as social media posts from Facebook[3] and Twitter[4]. To augment these datasets, we used the AFINN lexicon[5] and OpenAI's API of GPT-3.5 to generate example text messages including AFINN terms, using prompts like *"Generate 4 sentences using the word 'failure' that a 25 year-old in crisis would write in a text message"*. We measured valence as a continuous metric from 0 (extremely negative) to 1 (extremely positive);
- Motivational Interview (MI) response types used by the CC, including affirmations, reflections, and self-disclosure, among others. We used the Motivational Interviewing Treatment Integrity (MITI) code labels used by Welivita and Pu (2022). Leveraging their publicly available Motivational Interviewing Dataset, which contains approximately 2,000 hand-annotated conversations, we fine-tuned a pre-trained RoBERTa model[6] (*roberta-base* by Liu et al. (2019)) for sentence-level multi-class classification. Subsequently, we deployed this model to predict the appropriate label for each CC message;
- Use of past, present and future-oriented vocabulary from a self-compiled list of expressions;
- Use of expressions related to clinical communication strategies (e.g. hedging, giving advice etc.) using a self-compiled list reviewed by an expert clinician, as well as lists of examples for good contact techniques (e.g. tentafiers, strong feeling words) used in in-house counselor training;

[1] We used SpaCy (v3.7.4) to extract POS tags.

[2] We used the Transformers (v4.37.0), NumPy (v1.23.1), and Torch (v2.0.1) libraries. The last BERT hidden state was extended with four layers of a fully-connected neural network (512, 256, 128, and 1). We used a batch size of 32, a learning rate of 1×10^{-5}, and a dropout probability of 0.2.

[3] http://wwbp.org/downloads/public_data/dataset-fb-valence-arousal-anon.csv, downloaded on August 4, 2023.

[4] https://github.com/felipebravom/EmoInt, downloaded on August 4, 2023.

[5] http://www2.imm.dtu.dk/pubdb/pubs/6010-full.html

[6] The libraries used and the neural network attached to the final hidden states were identical to those used for emotional valence detection, except that the final layer consisted of 10 neurons corresponding to each MITI label.

- Count of templated utterances. CCs have access to a bank of commonly-used phrases; we identified these by calculating the Dice-Sørensen similarity coefficient (Dice, 1945; Sørensen et al., 1948) of 3-grams between each message and the most common 500 counselor messages;
- Timing-related features using message timestamps, including latency to reply to the first message and mean turn-by-turn response time.

Features potentially related to conversation word count were normalized by total word count and by the number of CC messages. For certain linguistic features, we created additional variables that compared the texter's statistics with those of the CC, either through ratios or differences, to represent mirroring of linguistic styles.

We converted the dependent variable, perceived genuine concern, into a binary format by labeling *"Strongly agree"* and *"Somewhat agree"* responses as positive, and all other response options as negative, similar to the approach used by Gould et al. (2022). The data showed a significant imbalance with 86% of responses being positive.

To address this, we employed stratified random sampling to select 50,000 conversations for the training set, balanced across the five original Likert-scaled answer options. For the test set, our sampling mirrored the overall base rates of each response option, resulting in an imbalanced set of 10,000 conversations.

We trained logistic regression with LASSO penalty (Tibshirani, 1996), Random Forest (Ho, 1995) and XGBoost (Chen and Guestrin, 2016) models with Bayesian optimization over hyper parameters in a 5-fold cross-validation setup[7]. The core goal of the project was to identify which linguistic features of CC utterances most strongly contribute to perceptions of genuine concern. We chose these models specifically because they allow us to readily identify how much each feature contributes to the prediction. Given the imbalanced nature of the test set and the importance of correctly classifying both classes, we assessed model performance using balanced accuracy. This metric represents the arithmetic mean of sensitivity and specificity.

Feature extraction required approximately 42 hours and was performed in a parallelized Spark environment using an NVIDIA Tesla T4 GPU, 4-core Intel Xeon 2.5GHz CPU, and 16GB of memory. Training each of the models (emotional valence, MITI labels, LASSO, Random Forest, and XGBoost) took approximately 30 minutes each.

4 Results

Table 1 compares the performance of the three models based on their predictions on the test set. As a baseline, we randomly assigned labels with prior probabilities corresponding to the base rates in the training set. The logistic regression with LASSO penalty performed best.

We used the 80 out of 140 features that the LASSO model did not shrink to zero and refitted the logistic model using these features. The variables were standardized. Table 2 provides the coefficient and odds ratio estimates for the 15 variables with the largest absolute coefficient magnitudes.

Conversations with a more positive emotional tone from the CC (mean and maximum valence) were associated with greater perceived genuine concern. At the same time, maintaining alignment with the texter's valence throughout the conversation is crucial: a substantial difference in turn-by-turn valence between CC and texter correlates with poorer perceptions of genuine concern.

Conversations with a higher CC word count tend to receive better ratings. Moreover, high mismatch between CC and texter in overall and turn-by-turn word count are associated with lower genuine concern. CCs maintaining a level of formality equal to or exceeding that of texters tend to be rated more positively. However, using significantly more complex language than the texter may lead to perceptions of lower genuineness.

The clinical content of CC messages reveals significant associations. Affirmations, including messages such as *"You should be proud of yourself for [...]"*, are associated with higher ratings of genuine concern. Conversely, Self-Disclosures, where CCs disclose personal information like *"I've also been in situations where [...]"*, tend to decrease the likelihood of a positive rating. Similarly, the use of first-person singular (e.g., 'I,' 'me') and first-person plural (e.g., 'we,' 'us') pronouns was associated with lower genuine concern, indicating that ratings tend to decrease when CCs talk about themselves.

[7] We used the scikit-learn (v1.1.1) library for statistical models and the scikit-optimize (v0.10.1) library for parameter tuning. Parameter tuning was conducted using the BayesSearchCV module with uniform priors. The optimal penalty for the LASSO model was found to be C=0.018.

Model	Balanced Accuracy
Baseline (random label assignment with prior)	50.8%
Logistic Regression with LASSO	**66.2%**
Random Forest	64.5%
XGBoost	65.1%

Table 1: Model performance comparison.

Variables	Coefficient	Odds Ratio	OR CI
Average CC emotional valence	0.428	1.533	[1.484, 1.584]
Max. (i.e., most positive) CC emotional valence	0.202	1.224	[1.183, 1.265]
Total number of words by CC	0.159	1.173	[1.140, 1.206]
Ratio of CC's and texter's writing formality	0.135	1.145	[1.122, 1.168]
% of CC turns including affirmations (MITI)	0.095	1.100	[1.039, 1.164]
Standard deviation in CC emotional valence	0.083	1.087	[1.048, 1.127]
Overall ratio of CC word count to texter word count	-0.039	0.962	[0.927, 0.998]
% of templated CC messages	-0.060	0.942	[0.901, 0.985]
Average num. of first-person singular pronouns per turn	-0.065	0.937	[0.911, 0.963]
Ratio of CC's and texter's mean answer latency	-0.066	0.936	[0.906, 0.967]
Average num. of first-person plural pronouns per turn	-0.067	0.935	[0.917, 0.955]
% of CC turns including self-disclosure (MITI)	-0.071	0.932	[0.912, 0.951]
Difference between CC and texter reading level	-0.105	0.901	[0.880, 0.922]
Average turn-by-turn ratio of CC word count to texter word count	-0.263	0.769	[0.722, 0.819]
Average difference between CC's and texter's emotional valence	-0.422	0.656	[0.635, 0.678]

Table 2: Coefficients, odds ratios (OR), and OR confidence intervals for the features with highest absolute coefficients in the refitted logistic regression. Features had been standardized; therefore, coefficients represent the change in log-odds for a one standard deviation increase in the independent variable. CC: Crisis Counselor. MITI: Motivational Interviewing Treatment Integrity Code.

CCs who use templated answers - evaluated at a 0.9 threshold of Dice-Sørensen similarity to the most common 500 counselor messages - are rated lower for genuineness. Finally, message timing matters: CCs who fall behind the texter's texting speed tend to receive worse ratings.

5 Discussion

Our results indicate that the relationship between perceived genuine concern and counselors' language is often context-dependent, with significant features reflecting differences between counselor and texter language. This conclusion is consistent with literature showing significant associations between perceived empathy and and therapist-client synchrony in specific aspects of language style (Lord et al., 2015). Additionally, our results demonstrate that maintaining conversational boundaries between counselor and texter can help perceived genuineness. Language that is less formal than the texter's corresponds with lower perceptions of empathy, in line with the findings of Lee et al. (2022). The literature shows mixed findings on the helpfulness of counselor self-disclosure (Henretty et al., 2014). In this unique text-based counseling setting, the use of self-disclosing messages and first-person pronouns were negatively associated with perceived genuine concern.

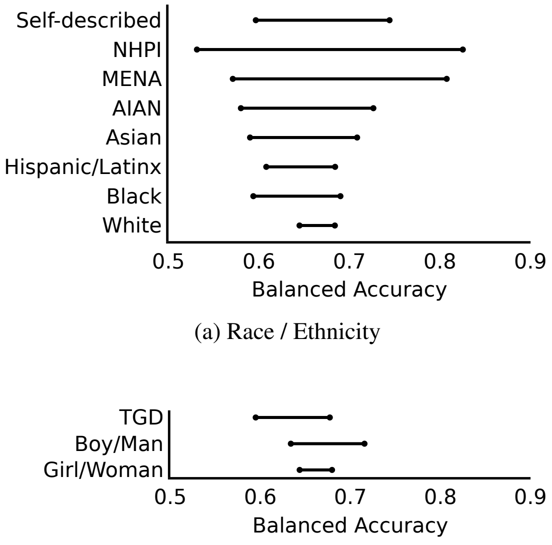
This work demonstrates the feasibility of using NLP techniques to extract clinically relevant linguistic features from text-based mental health interventions. Specifically, we fine-tuned complex transformer models for regression and classification tasks to extract features. Many of these features showed significant associations with perceived genuine concern. Our findings align with those of Imel et al. (2024), who documented the positive effect of using affirmations. Moreover, our results on the positive associations between counselor-texter synchrony in emotional valence and genuineness complement similar findings on synchrony in vocally encoded arousal and empathy (Imel et al., 2014). This foundational work identifies which linguistic cues are most strongly associated with genuine concern; future work could evaluate whether training CCs in how to increase these cues drives perceptions of higher genuine concern.

6 Limitations

First, while our study found significant associations between linguistic features and perceived genuine concern, establishing causality requires further investigation. For example, a higher mean counselor valence might occur in conversations where texters are already feeling better, leading counselors to make more positive statements. Second, this analysis incorporated results from an accessory model that evaluates emotional valence. Formal quantitative evaluation of that accessory model's performance is still pending; this requires extensive hand coding. The emotional valence model was trained using a synthetic dataset produced by GPT 3.5; therefore, the synthetic data generation process may not be fully reproducible. Third, regarding data sharing: while the dataset is de-identified, due to the highly sensitive content in messages, it is not publicly shareable, limiting reproducibility. Crisis Text Line has a research collaboration program that allows research teams to work with the dataset through an application process. Fourth: the features are interrelated (correlation matrix: in Appendix). We evaluated multicollinearity and used a LASSO approach for the logistic regression. However, we did not examine potential moderation effects, which could be an important direction for future work. Fifth: because the purpose of the current analysis was to identify features that relate to genuine concern, we did not compare our model to other approaches such as zero-shot and few-shot LLM-based models. Finally, we binarized the outcome variable (genuine concern). We initially planned to approach this using the fully granular (5-point) outcome. However, lack of sufficient data populating the lower ranges of the outcome limited our ability to take this approach.

7 Ethical Considerations

Despite the lack of clarity regarding causality, our model offers a foundation for developing tools to predict genuine concern in real time. However, there are potential risks associated with using the model to guide counselors. Inaccurate predictions could mislead counselors into taking actions that may not effectively improve perceived genuine concern. Additionally, if counselors are incentivized to achieve high scores, they may focus on enhancing specific linguistic features while neglecting other important aspects not addressed in this study that contribute to perceived empathy.



(a) Race / Ethnicity

(b) Gender identity

Figure 1: Model performance across demographic groups. Bars represent bootstrapped 95% confidence intervals. AIAN: American Indian or Alaska Native; MENA: Middle Eastern, North African, Arab; NHPI: Native Hawaiian or Pacific Islander; TGD: transgender and gender diverse.

Our sample of conversations includes texters from a wide range of ages, races, ethnicities, genders, and sexual orientations. We measured racial and gender bias by evaluating model performance across different demographic groups, using self-reported data from survey items where multiple selections were allowed. We drew 2,000 bootstrap samples from the test set and constructed a 95% confidence interval for balanced accuracy. Figure 1 shows that no significant difference in model performance was observed across these groups.

We acknowledge the potential for bias in the datasets used to train our feature-generating models, such as those for emotional valence detection and MITI technique labeling. A formal assessment of these biases has yet to be conducted. For further discussion of the replicability of our findings across race/ethnicity and gender, please see the Appendix.

Acknowledgments

This work was made possible by generous support to Crisis Text Line from the Jensen and Lori Huang Foundation.

References

Katie Aafjes-van Doorn, John Porcerelli, and Lena Christine Müller-Frommeyer. 2020. Language style matching in psychotherapy: An implicit aspect of alliance. *Journal of Counseling Psychology*, 67(4):509–522. Publisher: American Psychological Association.

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):49.

R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker. 2022. The development and psychometric properties of liwc-22.

Craig J. Bryan, Brian R. Baucom, Alex O. Crenshaw, Zac Imel, David C. Atkins, Tracy A. Clemans, Bruce Leeson, T. Scott Burch, Jim Mintz, and M. David Rudd. 2018. Associations of patient-rated emotional bond and vocally encoded emotional arousal among clinicians and acutely suicidal military personnel. *Journal of Consulting and Clinical Psychology*, 86(4):372–383.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ArXiv:1603.02754 [cs].

Phillippe B. Cunningham, Jordon Gilmore, Sylvie Naar, Stephanie D. Preston, Catherine F. Eubanks, Nina Christina Hubig, Jerome McClendon, Samiran Ghosh, and Stacy Ryan-Pettes. 2023. Opening the Black Box of Family-Based Treatments: An Artificial Intelligence Framework to Examine Therapeutic Alliance and Therapist Empathy. *Clinical Child and Family Psychology Review*, 26(4):975–993.

Suzanne E. Decker, Charla Nich, Kathleen M. Carroll, and Steve Martino. 2014. Development of the Therapist Empathy Scale. *Behavioural and Cognitive Psychotherapy*, 42(3):339–354.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy (Chicago, Ill.)*, 55(4):399–410.

Arnstein Finset and Knut Ørnes. 2017. Empathy in the Clinician–Patient Relationship. *Journal of patient experience*, 4(2):64–68.

Madelyn S. Gould, Anthony Pisani, Carlos Gallo, Ashkan Ertefaie, Donald Harrington, Caroline Kelberman, and Shannon Green. 2022. Crisis text-line interventions: Evaluation of texters' perceptions of effectiveness. *Suicide and Life-Threatening Behavior*, 52(3):583–595.

Kimberley M. Hara, Adi Aviram, Michael J. Constantino, Henny A. Westra, and Martin M. Antony. 2017. Therapist empathy, homework compliance, and outcome in cognitive behavioral therapy for generalized anxiety disorder: partitioning within- and between-therapist effects. *Cognitive Behaviour Therapy*, 46(5):375–390.

Mehedi Hasan, April Idalski Carcone, Sylvie Naar, Susan Eggly, Gwen L. Alexander, Kathryn E. Brogan Hartlieb, and Alexander Kotov. 2019. Identifying Effective Motivational Interviewing Communication Sequences Using Automated Pattern Analysis. *Journal of Healthcare Informatics Research*, 3(1):86–106.

Jennifer R. Henretty, Joseph M. Currier, Jeffrey S. Berman, and Heidi M. Levitt. 2014. The impact of counselor self-disclosure on clients: a meta-analytic review of experimental and quasi-experimental research. *Journal of Counseling Psychology*, 61(2):191–207.

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science*, 7(3):293–340.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Zac E. Imel, Jacqueline S. Barco, Halley J. Brown, Brian R. Baucom, John S. Baer, John C. Kircher, and David C. Atkins. 2014. The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology*, 61(1):146–153. Publisher: American Psychological Association.

Zac E. Imel, Michael J. Tanana, Christina S. Soma, Thomas D. Hull, Brian T. Pace, Sarah C. Stanco, Torrey A. Creed, Theresa B. Moyers, and David C. Atkins. 2024. Mental Health Counseling From Conversational Content With Transformer-Based Machine Learning. *JAMA Network Open*, 7(1):e2352590.

J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*.

Jonathan Him Nok Lee, Harold Chui, Tan Lee, Sarah Luk, Dehua Tao, and Nicolette Wing Tung Lee. 2022. Formality in psychotherapy: How are therapists' and clients' use of discourse particles related to therapist empathy? *Frontiers in Psychiatry*, 13:1018170.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sarah Peregrine Lord, Elisa Sheng, Zac E. Imel, John Baer, and David C. Atkins. 2015. More Than Reflections: Empathy in Motivational Interviewing Includes Language Style Synchrony Between Therapist and Client. *Behavior Therapy*, 46(3):296–303.

Adam S. Miner, Scott L. Fleming, Albert Haque, Jason A. Fries, Tim Althoff, Denise E. Wilfley, W. Stewart Agras, Arnold Milstein, Jeff Hancock, Steven M. Asch, Shannon Wiltsey Stirman, Bruce A. Arnow, and Nigam H. Shah. 2022. A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency. *npj Mental Health Research*, 1(1):1–12.

Theresa B. Moyers and William R. Miller. 2013. Is low therapist empathy toxic? *Psychology of Addictive Behaviors*, 27(3):878–884.

D. A. Shapiro. 1969. Empathy, warmth and genuineness in psychotherapy. *The British Journal of Social and Clinical Psychology*, 8(1):350–361.

T. Sørensen, T. Sørensen, T. Biering-Sørensen, Tia Sørensen, and J. T. Sorensen. 1948. A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on Danish commons.

Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Anuradha Welivita and Pearl Pu. 2022. Curating a Large-Scale Motivational Interviewing Dataset Using Peer Support Forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yijin Wu. 2021. Empathy in nurse-patient interaction: a conversation analysis. *BMC Nursing*, 20(1):18.

Rolf Wynn and Michael Wynn. 2006. Empathy as an interactionally achieved phenomenon in psychotherapy. *Journal of Pragmatics*, 38(9):1385–1397.

Bo Xiao, Chewei Huang, Zac E. Imel, David C. Atkins, Panayiotis Georgiou, and Shrikanth S. Narayanan. 2016a. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science*, 2:e59.

Bo Xiao, Zac E. Imel, Panayiotis Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2016b. Computational Analysis and Simulation of Empathic Behaviors: a Survey of Empathy Modeling with Behavioral Signal Processing Framework. *Current Psychiatry Reports*, 18(5):49.

Appendix

Class	**Precision**	**Recall**	**F1-score**
Advise	0.70	0.58	0.63
Affirm	0.46	0.42	0.44
Closed Question	0.62	0.95	0.75
Direct	0.50	0.70	0.58
Give Information	0.67	0.79	0.73
Open Question	0.00	0.00	0.00
Other	0.60	0.43	0.50
Reflection	0.47	0.46	0.46
Self-Disclosure	0.90	0.78	0.83
Support	0.65	0.68	0.67
Macro Averaging	0.56	0.58	0.56

Table 3: Motivational Interviewing model performance for each predicted class.

Feature	Type: Reciprocity
Timing	
Latency to the first message from the CC after the texter enters their crisis	
Number of times that the texter sent a message, waited five minutes, and then texted again with no response from the CC	
Number of times that the texter sent a message, waited three minutes, and then texted again with no response from the CC	
Mean time to respond to turn-by-turn texter messages from the CC	✓
Max time to respond to turn-by-turn texter messages from the CC	✓
Mean time to respond to turn-by-turn texter messages from the CC (normed by the parallel times from the texter)	✓
Max time to respond to turn-by-turn texter messages from the CC (normed by the parallel times from the texter)	✓
Writing mechanics	
Total number of words by CC	
Crisis counselor's word count per turn	
Overall ratio of CC word count to texter word count	✓
Average turn-by-turn ratio of CC word count to texter word count	✓
Ratio of CC word count to texter word count (turn-by-turn): standard deviation	✓
CC's writing formality	
Ratio of CC's and texter's writing formality	✓
CC reading level (whole conversation)	
Difference between CC and texter reading level	✓
Number of one-word responses from the CC	
Number of responses from the CC consisting of three words or fewer	
Punctuation: periods, normed for word count	
Punctuation: exclamations, normed by word count	
Punctuation: question marks, normed by word count	
Punctuation: commas, normed by word count	
Punctuation: semicolons, normed by word count	
Punctuation: dashes, normed by word count	
Punctuation: colons, normed by word count	
First person singular pronoun usage: mean [1], normed by word count [2], normed by count of pronouns [3]	
Second person singular pronoun usage: mean [1], normed by word count [2], normed by count of pronouns [3]	
First person plural pronoun usage: mean [1], normed by word count [2], normed by count of pronouns [3]	
Third person plural pronoun usage: : mean [1], normed by word count [2], normed by count of pronouns [3]	
Any pronoun usage: mean [1], normed by word count [2]	
Time orientation words: past: mean [1], normed by word count [2], normed by count of time oriented language uses [3]	
Time orientation words: present: mean [1], normed by word count [2], normed by count of time oriented language uses [3]	
Time orientation words: future: mean [1], normed by word count [2], normed by count of time oriented language uses [3]	
Questions starting with Who: mean [1], normed by word count [2]	
Questions starting with What: mean [1], normed by word count [2]	
Questions starting with Where: mean [1], normed by word count [2]	
Questions starting with When: mean [1], normed by word count [2]	
Questions starting with Why: mean [1], normed by word count [2]	
Questions starting with How: mean [1], normed by word count [2]	
Questions starting with Which: mean [1], normed by word count [2]	
Questions starting with Whose: mean [1], normed by word count [2]	
Happy emojis: mean [1], normed by word count [2]	
Heart emojis: mean [1], normed by word count [2]	
Happy emoticons: mean [1], normed by word count [2]	
Sad emoticons: mean [1], normed by word count [2]	
Clinical content	
MITI advise statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	
MITI affirm statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	
MITI question statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	
MITI direct statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	
MITI give information statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	
MITI reflection statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	

MITI self disclosure statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	
MITI support statements: mean [1], binary yes/no for present in the conversation [2], mean number that follow low valence statements from texter [3]	
Clinical content: checking understanding: mean [1], normed by word count [2]	
Clinical content: demonstrating understanding: mean [1], normed by word count [2]	
Clinical content: hedging: mean [1], normed by word count [2]	
Clinical content: absolutism: mean [1], normed by word count [2]	
Clinical content: giving advice: mean [1], normed by word count [2]	
Clinical content: giving opinions: mean [1], normed by word count [2]	
Clinical content: personal sharing / self-disclosure: mean [1], normed by word count [2]	
Good contact techniques: strong feeling words: mean [1], normed by word count [2]	
Good contact techniques: tentafiers: mean [1], normed by word count [2]	
Good contact techniques: validations: mean [1], normed by word count [2]	
Good contact techniques: strength identifications: mean [1], normed by word count [2]	
Good contact techniques: open ended questions: mean [1], normed by word count [2]	
Good contact techniques: clarifying questions: mean [1], normed by word count [2]	
Open questions: mean [1], normed by word count [2]	
Yes / No questions: mean [1], normed by word count [2]	
Emotional valence	
Minimum CC emotional valence	
Max. (i.e., most positive) CC emotional valence	
Average CC emotional valence	
Standard deviation CC emotional valence	
Minimum difference between CC emotional valence and texter emotional valence (turn-by-turn)	✓
Maximum difference between CC emotional valence and texter emotional valence (turn-by-turn)	✓
Mean difference between CC emotional valence and texter emotional valence (turn-by-turn)	✓
Standard deviation of difference between CC emotional valence and texter emotional valence (turn-by-turn)	✓
Other	
Number of simultaneous conversations the CC was handling during this conversation	
Templatedness: .6 similarity to templated statements	
Templatedness: .7 similarity to templated statements	
Templatedness: .8 similarity to templated statements	
Templatedness: .9 similarity to templated statements	
Templatedness: 1.00 similarity to templated statements	

Table 4: All variables included in the logistic regression model. CC: crisis counselor. Type: reciprocity indicates that the feature incorporates the interplay between the CC and corresponding behavior from the texter.

Supplemental analyses: top 15 features for demographic subgroups. We repeated the logistic regression analysis separately for different racial/ethnic categories, and for different gender categories. For many groups, fewer than 15 features were retained in the model. Note: for race/ethnicity, texters could select multiple options; texters are included in the analysis for each option they selected (i.e., if a texter selected 'Black or African American' and 'Asian', they are included in both analyses). Some texters answered the genuine concern survey item but elected not to report demographics.

Race/ethnicity:

- Asian, Asian American (N = 2838)
  - CC's formality, normed by texter's formality
  - Average turn-by-turn ratio of CC word count to texter word count
  - Open questions, normed by word count
  - Questions starting with What, normed by word count
  - Punctuation: periods, normed by word count
  - MITI reflection statements, mean number that follow low valence statements from texter
  - MITI question statements, mean number that follow low valence statements from texter
- Black or African American (N = 5627)
  - CC reading level (whole conversation)
  - Ratio of CC word count to texter word count (turn-by-turn): standard deviation
  - First person singular pronoun usage, normed by count of pronouns
  - Second person singular pronoun usage, normed by count of pronouns
  - Questions starting with Why, normed by word count
  - Questions starting with Why, mean
  - Questions starting with Which, normed by word count
  - Punctuation: dashes, normed by word count
  - Time orientation words: present, normed by word count
  - Happy emoticons, normed by word count
  - MITI direct statements, mean
  - MITI self disclosure statements, mean
  - MITI reflection statements, binary yes/no
  - Max. (i.e., most positive) CC emotional valence
- Latino / Latina / Latinx / Latine or Hispanic (N = 7503)
  - Latency to the first message from the CC after the texter enters their crisis
  - Max time to respond to turn-by-turn texter messages from the CC (normed by the parallel times from the texter)
  - Overall ratio of CC word count to texter word count
  - Questions starting with What, normed by word count
  - Questions starting with Where, mean
  - Punctuation: periods, normed by word count
  - Punctuation: semicolons, normed by word count
  - Time orientation words: future, normed by word count
  - Clinical content: absolutism, mean
  - Good contact techniques: open ended questions, mean
  - Good contact techniques: open ended questions, normed by word count
  - MITI direct statements, mean number that follow low valence statements from texter
  - MITI question statements, binary yes/no
  - Minimum CC emotional valence
  - Average CC emotional valence
- Middle Eastern, North African or Arab (N = 621): no features were retained
- Native American, Native Alaskan or Indigenous (N = 1777)
  - Any pronoun usage: mean
  - MITI question statements, binary yes/no
  - Minimum difference between CC emotional valence and texter emotional valence (turn-by-turn)
  - Standard deviation of difference between CC emotional valence and texter emotional valence (turn-by-turn)
- Native Hawaiian or Pacific Islander (N = 466): no features were retained
- White (N = 27,366)
  - Mean time to respond to turn-by-turn texter messages from the CC
  - Total number of words by CC
  - CC reading level (whole conversation)
  - Second person singular pronoun usage, normed by count of pronouns
  - Third person plural pronoun usage, normed by word count
  - Questions starting with Where, mean
  - Punctuation: colons, normed by word count
  - Punctuation: dashes, normed by word count
  - Time orientation words: future, normed by count of time oriented language uses
  - Time orientation words: past, normed by count of time oriented language uses
  - MITI affirmation statements, binary yes/no
  - MITI direct statements, mean
  - MITI reflection statements, mean number that follow low valence statements from texter
  - MITI question statements, mean number that follow low valence statements from texter
  - Good contact techniques: clarifying questions, normed by word count
- Prefer to self-describe (N = 1775): no features were retained

And for gender:

- Boy/Man Only (N = 6421)
  - Number of times that the texter sent a message, waited three minutes, and then texted again with no response from the CC
  - Total number of words by CC
  - Third person plural pronoun usage, mean
  - Questions starting with When, normed by word count

  - Questions starting with Where, mean
  - Punctuation: semicolons, normed by word count
  - Heart emojis, mean
  - Happy emoticons, normed by word count
  - Sad emoticons, normed by word count
  - Time orientation words: past, normed by count of time oriented language uses
  - Clinical content: personal sharing / self-disclosure, mean
  - MITI direct statements, mean
  - MITI reflection statements, mean
  - Number of simultaneous conversations the CC was handling during this conversation
  - Standard deviation CC emotional valence

- Girl/Woman Only (N = 31,662)
  - Latency to the first message from the CC after the texter enters their crisis
  - Mean time to respond to turn-by-turn texter messages from the CC (normed by the parallel times from the texter)
  - First person singular pronoun usage, normed by count of pronouns
  - Second person singular pronoun usage, normed by count of pronouns
  - Questions starting with Whose, mean
  - Happy emoticons, normed by word count
  - MITI affirm statements, mean number that follow low valence statements from texter
  - MITI question statements, mean number that follow low valence statements from texter
  - MITI self disclosure statements, mean number that follow low valence statements from texter
  - MITI reflection statements, binary yes/no
  - Good contact techniques: tentafiers, mean
  - Clinical content: hedging, mean
  - Yes / No questions, mean
  - Minimum CC emotional valence
  - Templatedness: .6 similarity to templated statements

- Transgender and Gender Diverse (N = 5810)
  - Max time to respond to turn-by-turn texter messages from the CC (normed by the parallel times from the texter)
  - First person plural pronoun usage, normed by count of pronouns
  - Questions starting with How, mean
  - Time orientation words: future, normed by word count
  - Happy emoticons, mean
  - Happy emojis, mean
  - MITI self disclosure statements, mean number that follow low valence statements from texter
  - MITI give information statements, mean
  - Good contact techniques: validations, normed by word count
  - Average CC emotional valence
  - Maximum difference between CC emotional valence and texter emotional valence (turn-by-turn)
  - Standard deviation of difference between CC emotional valence and texter emotional valence (turn-by-turn)

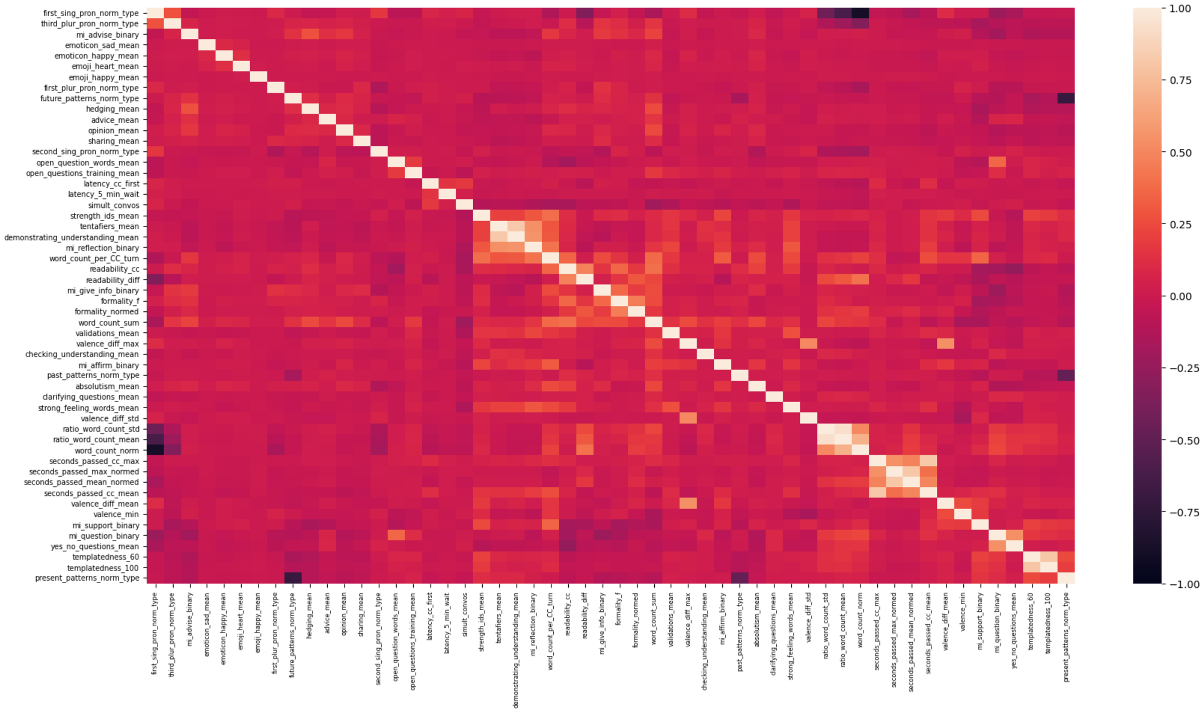


Figure 2: Feature correlation matrix (Spearman's correlation)