Behavioral Empathy Prediction in Dyadic Conversations with a Robot *** Supplementary Material ***

Christian Arzate Cruz*[1], Edwin C. Montiel-Vazquez*[2], Ruishuang Liu[1], Jorge Adolfo Ramírez Uresti[2], and Randy Gomez[1]

I. Introduction

Our empathy definition: The definition of empathy is still heavily discussed in psychological literature [1], [2]. This has led to differing definitions for the same concept in affective computing applications [3], [4], [5], [6], [7]. The definition closer to the one we adopt in this work is based on conceptualizing empathy as a multi-dimensional [2], [1] and hierarchical [8], [9] construct.

In particular, we adopt the empathy definition based on a multi-disciplinary body of research in [2]. It presents 3 empathy dimensions, which are related to experiencing a similar emotional state in response to another person's (affective), understanding another person's affective state (cognitive), and acting in a manner that is consistent with those two previous dimensions (behavioral).

Affective empathy is considered the lowest form of empathy [9], [10], supplemented with cognitive elements to obtain the highest form of empathy.

Our empathy cues definition: Empathy cues have been used, but not explicitly named as a concept, in many affective computing applications. In the field of conversational artificial intelligence, empathy-related concepts like intent [11] or communication mechanisms [12] have been integrated into the generation pipeline to increase empathetic capabilities in such systems [13], [14], [15], [16], [17], [18], [19]. Likewise, empathy cues have found use in empathy prediction by incorporating elements like emotion recognition, emotional signals, and sentiment [20], [21].

Previous work by Lee et al. [22] has shown that empathy classifiers only consider surface-level details. Therefore, we wished to obtain insight by studying the influence of empathy cues in classification.

Empathy in HRI: Most works focus on studying the effects of having an empathetic robot in an HRI setting [23], [24], [25]. Besides, some works explore the impact of robot mediators that foster discussion in groups with empathetic communication [26].

Another study area is how empathetic humans are towards robots. Authors in [27] found that depending on the robot's behavior, and the rest of the humans in the group, children demonstrate different levels of empathy to the robot.

* Joint first authorship.
[1] Honda Research Institute Japan (HRI-JP), Wako City, Japan. christian.arzate@jp.honda-ri.com
[2] Tecnologico de Monterrey, Monterrey, Mexico. edwincmv@exatec.tec.mx

Designing EERobot: Our approach for a multi-modal dataset for empathy predictions in dyadic conversations is based on EmpatheticDialogues [28]. However, we were interested in the expression of empathy in the presence of a robot. We adapt the strategy of a *speaker* role presenting a situation in which they felt a given emotion. Then, a *listener* role responds to that utterance. The emotional contexts for each conversation are one of Plutchik's 8 main emotions [29]. Each speaker-listener pair explores 8 emotional contexts. This was to ensure diversity in responses. Likewise, we designed to have different levels of empathy by randomly assigning levels of familiarity between the participants. This is based on research that shows differences in empathy modulation due to familiarity [30]. This information was provided to the participants through an initial interjection by a social robot, ensuring that both participants were aware of the distinct factors to consider during the conversation.

There are significant limitations to our database. The main limiting factor is its size, as there is a significant difference between EmpatheticDialogues and EERobot. Likewise, there is a lack of diversity among participants in the speaker role. Therefore, there might be biases introduced due to their demographic characteristics. Similarly, our emotional contexts are limited to higher-level emotions rather than those more fine-grained in EmpatheticDialogues.

Another HRI setting to explore in the future is with a social robot that actively mediates the participant's interaction. Furthermore, we want to explore the use of our system in group settings with a robot.

II. Implementation Details

Why we chose PBC4cip: It is designed to address class imbalance by computing weights for each class and using them alongside pattern support to perform classification. Its main advantage is that it is explainable and has outperformed others in empathy classification [20], [21]. We were interested in using this characteristic to study empathy classification in the context of HRI. However, the explainable capabilities of PBC4cip have been tied to the training set, which was of a different domain [20]. Therefore, we modify the algorithm to provide explainable information per individual inference. We name our implementation the contrast **P**attern-**B**ased **C**lassifier **for emp**athy (PBC4emp) and train it using a novel representation of just empathy cues in dyadic exchanges.

Besides, this classifier has already been proven to outperform classical machine-learning approaches for empathy classification [20], [21]. Likewise, it is advantageous since it is explainable. This information is given in the form of the influence of each empathy cue per prediction, which would allow us to study their effects in the context of HRI. It receives as an input the empathy cues present in conversation exchanges.

Empathy annotation: We elected to use annotators over self-annotation to label the conversations since we focused on empathetic behavior displayed [2]. Similar research showed third-party annotation to be more reliable [20], [31]. We obtained empathy labels from two annotators with experience. They took the Empathy Quotient (EQ) questionnaire [32], [33] to ensure they did not present deficits in their empathetic abilities.

This questionnaire was developed to identify a lack of empathy as a feature of psychological conditions. As such, it is useful for measuring when a person is incapable of empathy using a single score [20]. The cut-off point for people with empathy deficits is 30 [32]. Our annotators presented a level of 52 and 63. Therefore, there was no reason to doubt their capabilities in identifying empathy. Annotators were then given our empathy definition alongside the levels of empathy and were familiarized with examples from a version of empathetic dialogues with verified empathy labels developed in [20]. They were instructed to label the display of empathy of the listener at the conversation level.

We present metrics of annotators against the final result consensus label in Table I.

TABLE I: Annotators original labeling against final consensus. CEM = closeness evaluation measure, and F1 = F1-score.

2-level					
Annotator	Accuracy	CEM	F1	Precision	Recall
First	0.831	-	0.832	0.837	0.831
Second	0.859	-	0.857	0.862	0.859
3-level					
Annotator	Accuracy	CEM	F1	Precision	Recall
First	0.677	0.788	0.665	0.657	0.677
Second	0.724	0.823	0.728	0.739	0.724

Thanks to this method, we obtained empathy labels for the EERobot dataset. Furthermore, we provide a baseline for classification by comparing the results to human labeling.

Changes in arousal and valence values: In Figure 1, we present violin plots of the arousal and valence values distribution from text and video sources, and after performing a weighted sum of both.

III. Prompts for GPT-4o

We include the prompts for GPT-4o used for benchmarking. The prompt presented is for classification using transcriptions, VA vectors, and pose mimicry.

Rate the behavioral empathy displayed by the listener per exchange. Our definition of empathy supports three dimensions in a hierarchical order: Affective, Cognitive, and Behavioral. Affective empathy means having emotional congruence with another person (same or similar emotion); Cognitive empathy is to understand the feelings and emotions of another and why they feel that way; Behavioral empathy is to communicate in a manner that demonstrates cognitive and affective empathy. Note that high empathy includes expressions of interest for the other person's emotional state or their situation.

Also consider the additional metrics we extracted from video: Arousal (range: -1 to 1): Reflects the level of emotional intensity of the speaker and listener, with -1 being very calm and 1 being highly aroused. Valence (range: -1 to 1): Indicates the emotional valence (positive or negative) of the speaker and listener, with -1 being very negative and 1 being very positive. Pose mimicry (binary: 1 or 0): Indicates whether the listener mimics the speaker's head movement (1 for yes, 0 for no).

Give your answer on this scale: 1 - little to no empathy, 2 - somewhat empathetic, and 3-empathetic.

An example of a pair of exchanges with empathy level 1: speaker: There was strange noise from the kitchen at night. listener: Did you have your alarm on speaker: Yes, but it didnt ring. I was spooked but it turned out I had left the window open and it was the wind. listener: Oh wow, I would have thought maybe an animal had gotten in.

An example of a pair of exchanges with empathy level 2: speaker: I was babysitting a few days ago and my cousin kept throwing her food on the floor. listener: That must have been frustrating. What did you do? speaker: I kept cleaning it up and put down a bunch of towels around her to protect the floor. I think it comes with the territory, she gets what she wants. listener: That made me laugh. And you are a great babysitter.

An example of a pair of exchanges with empathy level 3: speaker: My dad knew that I have recently hit a rough patch with my finances and gave me 2 grand. I cannot thank him enough. listener: How kind of him! I bet that is going to help you out a lot! speaker: It really is going to help a lot. I feel so blessed! listener: Dads are the best, what would we do without them?

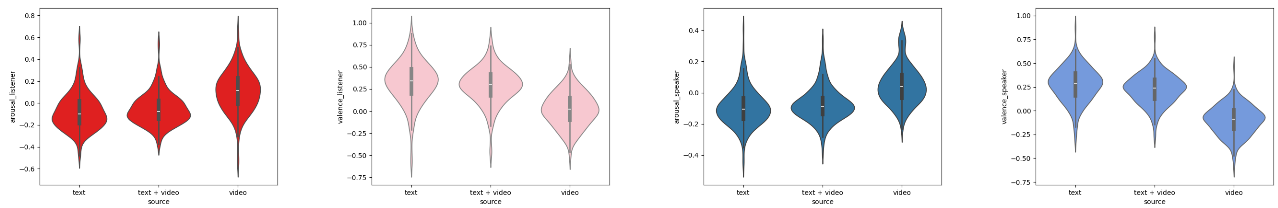
Answer format: Rate: ... Reason: ...

Other prompts used for this purpose had the following differences:

- Modified rating scale to be *Give your answer on this scale: 1 - no empathy, 2 - empathetic. .*
- Modified example with level 3 to have level 2 during binary classification.
- Eliminated the condition to consider additional metrics.

IV. Interpretability

Empathy cues ranking: In Figure 2, we present the ranking of cues. In Figure 3, we present the differences between the models that only use text against their counterparts that include video-based cues. The inclusion of video-based data changes the relevance of the cues for 2-level and 3-level classification. The biggest change is present in the listener's polarity response and the listener's emotional response. This shows that video-based cues alongside polarity and intent decrease the relevance of factors such as the explicit statement



(a) listener arousal (b) listener valence (c) speaker arousal (d) speaker valence

Fig. 1: valence and arousal values for text, text and video, and video alone.

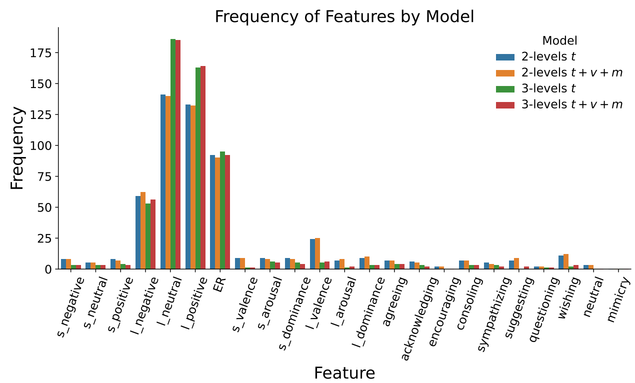


Fig. 2: Empathy cues ranking.

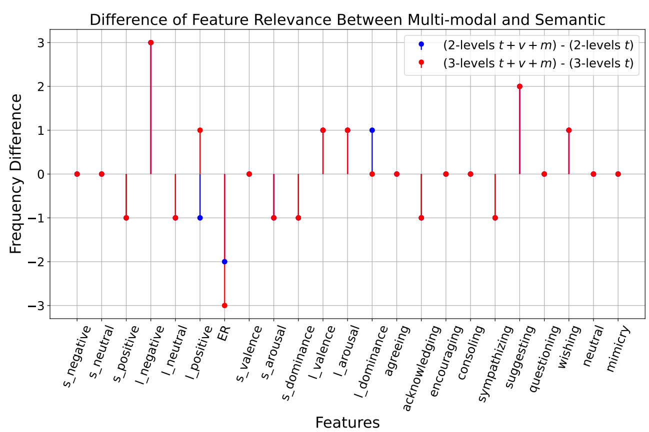


Fig. 3: Differences in the empathy cues rankings.

of an emotional reaction. Further work is necessary to study if this trend is present in larger datasets.

PBC4emp's explanations: In Table

GPT-4o's explanations:

References

[1] "Empathy: A review of the concept," *Emotion Review*, vol. 8, no. 2, pp. 144–153, 2016.

[2] M. A. Clark, M. M. Robertson, and S. Young, ""I feel your pain": A critical review of organizational research on empathy," *Journal of Organizational Behavior*, vol. 40, no. 2, pp. 166–192, Feb. 2019.

[3] C. D. Batson, J. Fultz, and P. A. Schoenrade, "Distress and Empathy: Two Qualitatively Distinct Vicarious Emotions with Different Motivational Consequences," *Journal of Personality*, vol. 55, no. 1, pp. 19–39, mar 1987. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-6494.1987.tb00426.x

[4] S. Buechel, A. Buffone, B. Slaff, L. Ungar, and J. Sedoc, "Modeling empathy and distress in reaction to news stories," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 4758–4765.

[5] D. Omitaomu, S. Tafreshi, T. Liu, S. Buechel, C. Callison-Burch, J. Eichstaedt, L. Ungar, and J. Sedoc, "Empathic conversations: A multi-level dataset of contextualized conversations."

[6] J. Shen, M. Sap, P. Colon-Hernandez, H. W. Park, and C. Breazeal, "Modeling empathic similarity in personal narratives." [Online]. Available: http://arxiv.org/abs/2305.14246

[7] S. A. Morelli, M. D. Lieberman, and J. Zaki, "The emerging study of positive empathy," *Social and Personality Psychology Compass*, vol. 9, no. 2, pp. 57–68, Feb. 2015.

[8] C. Cliffordson, "The hierarchical structure of empathy: Dimensional organization and relations to social functioning," vol. 43, no. 1, pp. 49–59.

[9] D. Waal and F. B.m, "The 'russian doll' model of empathy and imitation," in *On Being Moved: From mirror neurons to empathy*, ser. Advances in Consciousness Research, S. Bråten, Ed. John Benjamins Publishing Company, pp. 49–69.

[10] T. Fuchs, "Levels of empathy – primary, extended, and reiterated empathy," in *Empathy: Epistemic Problems and Cultural-Historical Perspectives of a Cross-Disciplinary Concept*, V. Lux and S. Weigel, Eds. Palgrave Macmillan UK, pp. 27–47.

[11] A. Welivita and P. Pu, "A Taxonomy of Empathetic Response Intents in Human Social Conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 4886–4899.

[12] A. Sharma, A. Miner, D. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 5263–5276.

[13] A. S. Raamkumar and Y. Yang, "Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2722–2739, Oct. 2023.

[14] Y. Zhang, F. Kong, P. Wang, S. Sun, L. Wang, S. Feng, D. Wang, Y. Zhang, and K. Song, "StickerConv: Generating multimodal empathetic responses from scratch." [Online]. Available: http://arxiv.org/abs/2402.01679

[15] Z. Wen, J. Cao, J. Shen, R. Yang, S. Liu, and M. Sun, "Personality-affected emotion generation in dialog systems," vol. 42, no. 5, pp. 1–27.

[16] O. Sotolar, "EmPO: Theory-driven dataset construction for empathetic response generation through preference optimization."

[17] B. Li, H. Fei, F. Su, F. Li, and D. Ji, "Integrating discourse features and response assessment for advancing empathetic dialogue," vol. 61, no. 5, p. 103803.

[18] O. Hamad, A. Hamdi, and K. Shaban, "ASEM: Enhancing empathy in chatbot through attention-based sentiment and emotion modeling."

[19] W. Li, Y. Yang, P. Tuerxun, X. Fan, and Y. Diao, "A response generation framework based on empathy factors, common sense, and persona," vol. 12, pp. 26 819–26 829.

[20] E. C. Montiel-Vázquez, J. A. Ramírez Uresti, and O. Loyola-González, "An Explainable Artificial Intelligence Approach for Detecting Empathy in Textual Communication," *Applied Sciences*, vol. 12, no. 19, p. 9407, Sep. 2022.

[21] M. R. Hasan, M. Z. Hossain, S. Ghosh, S. Soon, and T. Gedeon, "Empathy detection using machine learning on text, audiovisual, audio or physiological signals," *arXiv preprint arXiv:2311.00721*, 2023.

[22] A. Lee, J. K. Kummerfeld, L. An, and R. Mihalcea, "Empathy identification systems are not accurately accounting for context," A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1686–1695.

[23] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll, and L. Hall, "Caring for agents and agents that care: Building empathic relations with synthetic agents," in *Autonomous Agents and Multiagent Systems, International Joint Conference on*, vol. 2. IEEE Computer Society, 2004, pp. 194–201.

[24] B. Gonsior, S. Sosnowski, M. Buß, D. Wollherr, and K. Kühnlenz, "An emotional adaption approach to increase helpfulness towards a robot," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 2429–2436.

[25] H. Cramer, J. Goddijn, B. Wielinga, and V. Evers, "Effects of (in) accurate empathy and situational valence on attitudes towards robots," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 141–142.

[26] P. Alves-Oliveira, P. Sequeira, F. S. Melo, G. Castellano, and A. Paiva, "Empathic robot for group learning: A field study," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 8, no. 1, pp. 1–34, 2019.

[27] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction: evaluating social presence, engagement and perceived support in children," *International Journal of Social Robotics*, vol. 6, pp. 329–341, 2014.

[28] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 5370–5381.

[29] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[30] Y. "Motomura, A. Takeshita, Y. Egashira, T. Nishimura, Y.-K. Kim, and S. Watanuki, ""interaction between valence of empathy and familiarity: is it difficult to empathize with the positive events of a stranger?"," *J Physiol Anthropol*, vol. 34, no. 1, p. "13", mar 2015.

[31] S. Shi, Y. Sun, J. Zavala, J. Moore, and R. Girju, "Modeling clinical empathy in narrative essays," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. IEEE, pp. 215–220.

[32] S. Baron-Cohen and S. Wheelwright, "The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences," vol. 34, pp. 163–175.

[33] E. J. Lawrence, P. Shaw, D. Baker, S. Baron-Cohen, and A. S. David, "Measuring empathy: reliability and validity of the empathy quotient," vol. 34, no. 5, pp. 911–920.