EasyChair Preprint

№ 11648

Towards Designing Empathetic and Trustworthy AI Chatbots: an Exploratory Study

Melik Ozolcer, Mohammad Rahul Islam, Abdullah Mohammed, Tongze Zhang, Sang Won Bae and Ting Liao

January 2, 2024

Towards Designing Empathetic and Trustworthy AI Chatbots: An Exploratory Study

***Abstract*—Intelligent agents, such as chatbots, have recently attracted enormous attention due to their advanced capability to augment humans in information gathering and task execution. While they are designed to be more understanding, the existing literature lacks deep knowledge of how chatbots should track and respond to users' emotions in real-time with empathy. In this exploratory study, we proposed an innovative emotion-detecting system that combines CNN-based facial expression recognition algorithms with text-based sentiment analysis to improve real-time interactions between users and an AI-powered chatbot by recognizing users' emotional expressions and delivering empathetic responses appropriately. We present preliminary results of a human-subject study with distinct versions of chatbots. We confirm that adding facial expression detection improves the predictability of the models of user-perceived trust and empathy.**

***Index Terms*—Facial Expression, Trust, Empathy, Chatbot, Emotional States, Human-AI Interaction**

I. Introduction

Chatbots empowered by artificial intelligence (AI) have recently attracted enormous attention from novice users and researchers due to their advanced capabilities and easy-to-use interface [1]. The rapidly evolving chatbots create new interaction dynamics, which involve social aspects inevitably. The Computers Are Social Actors (CASA) paradigm indicates that people may apply the social norms of human relationships when interacting with intelligent agents [2, 3]. To design intelligent agents that comply with social rules, researchers examine different design strategies that enable the agents to understand human users and mimic human behavior. Understanding another individual often involves understanding what it feels like to be that person – in short, it entails empathy [4]. While intelligent agents, such as chatbots, are designed to be more understanding or empathetic, the existing literature lacks deep knowledge of how chatbots should track and respond to users' emotions in real-time with empathy.

Human users express their feelings through conversations with chatbots, where text-based sentiment analysis has been commonly used to understand their emotional states. Yet, some researchers argue that people are unwilling or unable to express their true feelings in words [5]. Since the existing studies have confirmed the benefits of facial expressions on understanding complex mental states [6, 7], in this study, we proposed an innovative emotion-detecting system to track users' emotional states using a combination of facial expression recognition algorithms with text-based sentiment analysis, to improve real-time interactions between users and an AI-powered chatbot by delivering empathetic responses when recognizing users' emotional expressions.

II. Background

A. Empathy in Design of Intelligent Agents

Empathy refers to the "reactions of one individual to the observed experiences of another" [8, 9]. It plays a critical interpersonal and societal role, enabling sharing experiences, needs, and desires between individuals and promoting prosocial behavior. The current results show that the empathetic chatbot using first-person narratives and recognizing people's emotions leads to higher user satisfaction than the non-empathetic chatbot [10]. It confirms the benefit of incorporating emotional empathy in the design of AI systems.

B. Facial Expression Algorithms in Detecting Mental States

Facial expressions are a primary form of human communication essential in expressing emotions. These signals typically appear as sensations of anger, disgust, fear, happiness, sadness, surprise, and neutral, regardless of one's culture or background [11]. Researchers believe these emotional reactions are inherent in humans, having evolved to be universally understood. For facial expression recognition, we implemented a convolutional network (CNN) [12] and trained it with the FER 2013 dataset to infer a person's discrete facial expression in seven frequently appearing facial expressions that are a proxy for communicating feelings and complex mental states.

III. Method

To complement the existing text-based sentiment analysis, we proposed that adding facial expression detection could better understand people's feelings and enable timely empathetic responses of the chatbot. As an exploratory study, we implemented a chatbot interface using Flask and tested three conditions, each with a distinct version of the chatbot. The conditions include the control (**CR**), where the chatbot only acknowledges receiving users' inputs but not their emotional expressions; empathetic chatbot based on sentiment analysis (**SA**); empathetic chatbot based on *an integration of* the sentiment analysis and facial expression detection (**SAFE**). We hypothesize that an empathetic chatbot using either sentimental analysis alone (SA) or sentiment analysis and facial expression detection together (SAFE) can improve user-perceived trust, and empathy, compared to the non-empathetic one (CR). To test the hypotheses, we conducted a human-subject experiment approved by the university's Institutional Review Board (IRB) in the lab.

The experiment consists of three steps. Participants first filled out a survey about their general beliefs about automated agents. Then, participants were randomly assigned to interact with a chatbot using free text. The chatbot introduced itself and asked participants about coursework experience, career plans,

etc. During the interaction, the empathetic chatbot (SA and SAFE) responded to participants' feelings when the chatbot recognized emotional expressions. The non-empathetic (CR) chatbot delivered neutral responses such as "Thank you for answering," regardless. The interaction ended when participants answered all eight pre-determined questions. Then, participants completed a post-interaction survey about their trust and perceptions of the chatbot.

IV. Results and Analysis

A. Data Collection and Preparation

We recruited 39 participants, including 59% male, 38% female, and 3% non-binary individuals. Fifty-one percent were between 18 and 25 years old, and the rest were between 25 and 35.

The dataset consists of three parts – (i) self-reported survey responses, (ii) interaction transcripts, and (iii) facial detection results. Five-point Likert scale (1-5) surveys were administered before and after the interaction to capture participants' perceived effort, empathy, emotional acknowledgment, and trust in the chatbot, among other relevant dimensions.

To prepare for the analysis, the survey items were tested for reliability using Cronbach's alpha. Due to the high reliability ($\alpha > 0.70$), the scores of multiple-item questionnaires for emotional acknowledgment, trust, and empathy are aggregated into the mean values of each participant.

The participants' textual inputs while interacting with the chatbot were tracked and analyzed using the Valence Aware Dictionary and sEntiment Reasoner (VADER) model, producing a score for positive, negative, and neutral emotions [13]. In addition, participants' facial expressions were analyzed during the interaction to detect and classify facial expression features associated with different emotional states, such as happiness, surprise, anger, disgust, fear, and sadness, as well as neutral [14].

B. Comparison Across Conditions

Survey data were first examined across conditions using the Kruskal-Wallis test. Among all the survey metrics, there is a significant difference in emotional acknowledgment across three conditions ($p = 0.00$). This shows the participants perceived the empathetic chatbots to better acknowledge their emotions (Table I), and the chatbot's empathetic behavior was well-received. Yet, there is no significant difference in trust and empathy across conditions. Thus, the hypotheses are not supported at the aggregated level.

TABLE I
Statistical Results of Emotional Acknowledgement

Condition	Mean	Std
CR	2.12	0.67
SA	3.40	0.81
SAFE	3.29	1.06

TABLE II
Model Comparison of Different Feature Combinations

Model	Features	Accuracy	F1-score	Precision	Recall
Trust	FE	0.877 (0.002)	0.769 (0.003)	0.781 (0.003)	0.758 (0.004)
Trust	SA	0.498 (0.003)	0.441 (0.006)	0.362 (0.004)	0.564 (0.01)
Trust	SA + FE	0.870 (0.002)	0.758 (0.004)	0.763 (0.005)	0.753 (0.006)
Empathy	FE	0.854 (0.004)	0.800 (0.006)	0.800 (0.008)	0.800 (0.004)
Empathy	SA	0.576 (0.003)	0.441 (0.002)	0.343 (0.002)	0.618 (0.004)
Empathy	SA + FE	0.851 (0.001)	0.788 (0.002)	0.788 (0.003)	0.788 (0.002)

C. Predictive Models using Machine Learning

To further investigate how the integration of facial expression detection with sentiment analysis can enhance the predictability of self-reported trust and empathy, respectively, in the context of user-chatbot interaction, we built the machine learning models with the EXtreme Gradient Boosting Machine classifier (XGBoost) [15], *regardless of the conditions*. We leveraged facial behavior features when participants interacted with the chatbot and self-reported trust and empathy across three feature group combinations – facial expression features only (FE), sentiment analysis features only (SA)[1], and facial expression and sentiment analysis features combined (SA + FE). The findings for this analysis are reported in Table II.

1) Machine Learning Model Development: To build the model, we first binarized scores as low-neutral trust/empathy vs. high trust/empathy. Then, to address the class imbalance, we employed Synthetic Minority Oversampling Technique (SMOTE) [16] and evaluated our models using 10-fold cross-validation (CV), to maximize the utilization of available samples. SMOTE was applied only to the training set in each of the ten iterations in the 10-fold CV. Even with the 10-fold CV, the performance of the models was unstable and seemed to be influenced by randomness in the data splitting. To address this issue, we conducted 100 simulations for each model and reported average results and standard deviations in parentheses (in Table II). This approach enabled a more robust and reliable assessment of the performance of our models and the impact of feature groups.

As shown in Table II, we observed that the text-based sentiment analysis (SA) models exhibited poor performance for both trust (Accuracy = 0.498) and empathy (Accuracy= 0.576) models. However, incorporating facial expression features (SA + FE) into the models significantly improved, with trust and empathy predictive models achieving performance boosts of 37.2% and 27.5%, respectively. The FE-only models demonstrated similar accuracy to the SA + FE model yet

[1]SA here refers to results of the sentiment analysis (scores of positive, negative and neutral emotions), rather than an experiment condition.

outperformed both models by striking a balance between precision (Trust = 0.781, Empathy = 0.800) and recall (Trust = 0.758, Empathy = 0.788), resulting in F1-scores of 0.769 for the trust model, and 0.788 for the empathy model. For both models, only incorporating facial expression features (FE) yielded the best performance, with accuracy scores of 0.877 and 0.854, F1-scores of 0.769 and 0.800 for trust and empathy models respectively. These outcomes confirm the efficacy of facial expression features in enhancing both the trust and empathy models, respectively.

V. Discussion

In summary, user-perceived trust and empathy are not significantly different across conditions with a simple comparison. However, the features of facial expressions show a strong impact on identifying empathy and trust, respectively, across participants. Detecting facial behavior markers during interactions with AI chatbots can improve the predictability of building models of empathy and trust based on emotional states.

VI. Limitation and Future Work

Due to the limited number of samples, it is challenging to determine whether the facial expression features capture participants' reactions to the chatbot responses or their mental state while contemplating how to respond to the generated question by the chatbot. To gain more insights, we intend to distinguish and analyze each phase of the chatbot interaction while extending its duration.

It is important to acknowledge that the model's performance may decline when employing unseen participants' data as a test set in real-world case scenarios. To enhance the generalizability of the model, we further plan to gather data on a large scale at our institution.

We also intend to conduct more advanced statistical analysis beyond the simple group comparison for hypothesis testing. We will also look into ways of better integrating sentiment analysis and facial expression detection in the SAFE condition.

VII. Conclusion

We developed and tested empathetic chatbots using text-based sentiment analysis or a combination of sentiment analysis and facial expression detection to recognize and respond to participants' emotional expressions in real time. Both empathetic chatbots lead to a higher level of emotional acknowledgment perceived by the participants than a non-empathetic chatbot in the control condition. Yet, self-reported trust and empathy are not significantly different across conditions *at the aggregated level*. Besides group comparison, we found that integrating facial expression detection into sentiment analysis enhances the predictability of the trust and empathy models, achieving an accuracy of 92% for trust and 93% for empathy in user-chatbot interaction.

Our results show that understanding human emotional states is critical for AI-empowered chatbots to appropriately correspond to individuals experiencing feelings at the moment, allowing for more personalized and empathetic interactions.

While pre-trained learning and in-context learning models have been applied, the current Large Language Models (LLMs), such as ChatGPT4, encounters a limitation regarding uncontrollability regarding topics. Users' prompts may steer the conversation off-topic, which can potentially cause distrust in the system. Therefore, it is crucial to not only comprehend the meaning of users' text inputs (prompts) but also understand human reactions and infer mental states expressed during interactions with chatbots. Moving forward, researchers and engineers can gain insights into designing AI chatbots capable of generating real-time, personalized responses that cater to individuals' needs and preferences, going beyond mere user prompts.

References

[1] K. Hu, "ChatGPT Hits 100 Million Users, Google Invests In AI Bot And CatGPT Goes Viral," 2 2023.

[2] T. Liao and E. F. MacDonald, "Manipulating Users' Trust of Autonomous Products With Affective Priming," *Journal of Mechanical Design*, vol. 143, no. 5, pp. 1–12, 2021.

[3] C. S. U. Nass and Y. H. U. Moon, "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues*, vol. 1, no. 56, pp. 81–103, 2000.

[4] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[5] L. Click, "31 Empathetic Statements for When You Don't Know What to Say," 11 2017.

[6] P. Ekman, E. Rosenberg, and J. Hager, "Facial Action Coding System Affect Interpretation Dictionary (FACSAID)," 1998.

[7] S. Roper and S. W. Bae, "Exploring Students' Flow States Using Facial Behavior Markers in an Online At-Home Learning Environment," in *2022 CHI Conference on Human Factors in Computing Systems*. The ACM CHI Conference on Human Factors in Computing Systems workshop on the Future of Emotion in Human-Computer Interaction, 2022, pp. 1–5.

[8] J. L. Hess and N. D. Fila, "The development and growth of empathy among engineering students," *ASEE Annual Conference and Exposition, Conference Proceedings*, vol. 2016-June, 2016.

[9] S. R. Daly, S. Yilmaz, J. L. Christian, C. M. Seifert, and R. Gonzalez, "Design heuristics in engineering concept generation," *Journal of Engineering Education*, vol. 101, no. 4, pp. 601–629, 10 2012.

[10] T. Liao and B. Yan, "Are You Feeling Happy? The Effect of Emotions on People's Interaction Experience Toward Empathetic Chatbots," in *International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, 2022.

[11] M. Frank, "Facial Expressions," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J.

Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 5230–5234.
[12] N. Sharma, "Facial Emotion Recognition on FER2013 Dataset Using a Convolutional Neural Network," 2018.
[13] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, 5 2014.
[14] M. Zhao, F. Yu, and Y. Dai, "Deep Facial Expression Recognition Using ResNet34," *World Scientific Research Journal*, vol. 6, pp. 380–386, 2020.
[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.