Empathetic Robot With Transformer-Based Dialogue Agent

Baijun Xie[1] and Chung Hyuk Park[1*]

***Abstract*— Natural Human-Robot interaction (HRI) attracts considerable interest in letting robots understand the users' emotional state. This paper demonstrates a method to introduce the affection model to the robotic system's conversational agent to provide natural and empathetic HRI. We use a large-scale pre-trained language model and fine-tune it on a dialogue dataset with empathetic characteristics. Based on existing studies' progress, we extend the current method and enable the agent to perform advanced sentiment analysis using the affection model. This dialogue agent will allow the robot to provide natural response along with emotion classification and the estimations of arousal and valence level. We evaluate our model using different metrics, comparing it with the recent studies and showing its emotion detection capacity.**

***Index Terms*— Human-Robot interaction (HRI), empathetic dialogue agent, Transformer, affection model**

I. Introduction

Empathy is the ability to feel another person's emotional states by placing oneself in another person's position. Letting users feel empathy and robot understand the user's affective states are important during Human-Robot interaction (HRI) [1]. Embodying an interactive dialog system is considered to be a crucial factor for building an empathetic robotic system. Recent developments in this dialogue system have led to enable the robotic system to detect user's emotional states and respond accordingly [2], [3]. However, most previous studies relied on modularized dialogue systems, such as natural language processing (NLP) and dialogue generation modules, making the robotic system complicated. There is an unmet need to build an end-to-end natural dialogue agent that can also understand users' feelings.

This paper presents an empathetic dialogue agent that can generate fluent and natural responses with empathy. Besides that, our agent can also provide the detection of 32 different emotion labels and the estimation of continuous levels of arousal and valence. In the second section, we provide the recent studies and related work. Our agent extends the previous method and incorporates a affection model for sentiment analysis. The third section gives a brief overview of our robotic interactive system. We showed the extended model architecture and the way we incorporate the affection model in the fourth section. Finally, we present and analyze the results of the model performance and give conclusions.

[1]Baijun Xie (bdxie@email.gwu.edu) and Dr. Chung Hyuk Park (chpark@gwu.edu) are with the Department of Biomedical Engineering, School of Engineering and Applied Science, The George Washington University, Washington DC, 20052, U.S.A.

* indicates the corresponding author.

II. Related Works

Recently, deep learning language models have gained plenty of attention and have shown a significant advantage over different NLP methods and over different NLP tasks [4], [5]. Rather than deploying the complex system and task-specific modules, deep learning language models are often be trained in end-to-end and data-driven manners. Detecting and responding to human emotions in empathetic ways is crucial in building natural and human-like dialogue agents for a robotic system. In the recent study, Rashkin et al. [6] proposed a new dataset for the empathetic dialogue agent and trained with Transformer-based [7] deep neural network. The experimental results showed that the dialogue models were observed to be more empathetic by human experts. Nevertheless, a significant challenge for empathetic dialogue agents is the dataset is relatively small, and the model is hard to learn enough language representation to generate natural and fluent responses.

More recently, using a large-scale pre-trained model and performing transfer learning which fine-tuning on target task, has become a feasible and effective way for dialogue agent [8]. This progress mitigates the problem of the shortage of task-specific datasets. This study proposes an empathetic dialogue agent with the affective model for our robotic platform. Comparing with Lin et al. [9], we also extend the idea of transfer learning method [8] on the empathetic dialogue dataset [6] with a large-scale pre-trained language model [5]. Besides an additional dialogue emotion classification task, we introduce Russel's circumplex model [10] of affect to provide the labels of the sentiment dimensions values of valence and arousal [11], and perform optimization jointly. This affective model enables us to achieve fine-grained sentiment analysis and better emotion recognizing performance from the empathetic dialogue dataset.

III. System Overview

Figure 1 shows the overview of our empathetic HRI system. In this study, we choose Pepper robotic platform manufactured by Softbank robotics [12]. Pepper is optimized for social interaction with the human through conversation, gestures, and touch screen capabilities. This system aims to recognize the user's emotional states during the interaction and generate empathy responses. Depending on the dialogues' content and acoustic features, the dialogue agent will predict the emotion labels and estimate valence and arousal measures. By leveraging the speech recognition module on Pepper, our empathetic dialogue agent can achieve natural human-robot interactions and give predictions and estimations of emotion in real-time.

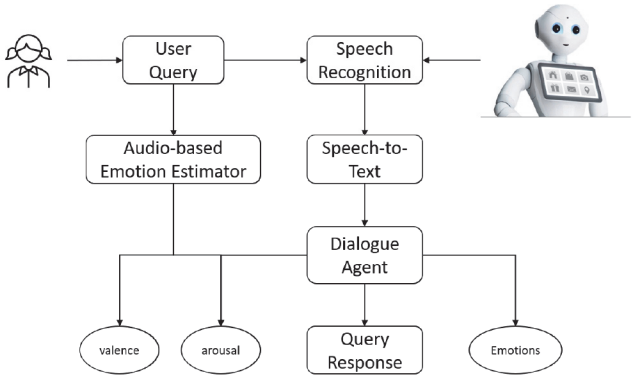

Fig. 1. The overview of our robotic interaction system

Label: excited Situation: My grand ma bought me a gold watch as a birthday gift, that really made me happy
Speaker: I have really wished to have one ever since I lost the one my mum bought during my convocation Listener: What are you talking about? Speaker: I am talking about a gold watch my grand ma bought for me Listener: Oh ok. That's awesome. It's nice to have something of value.
Label: sad Situation: I felt sad yesterday. At work our computer systems went down and customers were mad
Speaker: Well it started off I was unable to check people in. The IT people are very slow at fixing the issue. Listener: Oh man, where customers upset with you? Speaker: How customers blame employees like it's their fault Listener: Exactly and they are getting free healthcare. Some people are so entitled.

TABLE I
THE EXAMPLES FROM THE EAPATHETICDIALOGUES DATASET

An audio-based emotion estimator from our previous study [13] is also incorporated into the system. This emotion estimator can use acoustic speech features to get the levels of valence and arousal in real-time. The outputs from both the estimator and the dialogue agent will be combined to get the final estimates in the future user study.

IV. METHODOLOGY

A. Dataset

Rashkin et al. [6] proposed a novel dataset, EmpatheticDialogues, of 25k personal conversations with situations, which labeled by 32 emotions classes with the balanced distribution. Table I shows two examples taken from the EapatheticDialogues training set. As shown, every dialogue sample is given an emotion label and a sentence describing the situation the speaker felt. The listener tried to understand the speaker's feelings and respond with empathy accordingly. The emotion labels will be used for optimizing the emotional classification purpose while training our model.

B. Affection Model

The representation of the affection model is taken and adapted from the study of [11]. As illustrated in Figure 2, the two-dimensional circumplex space model is based on

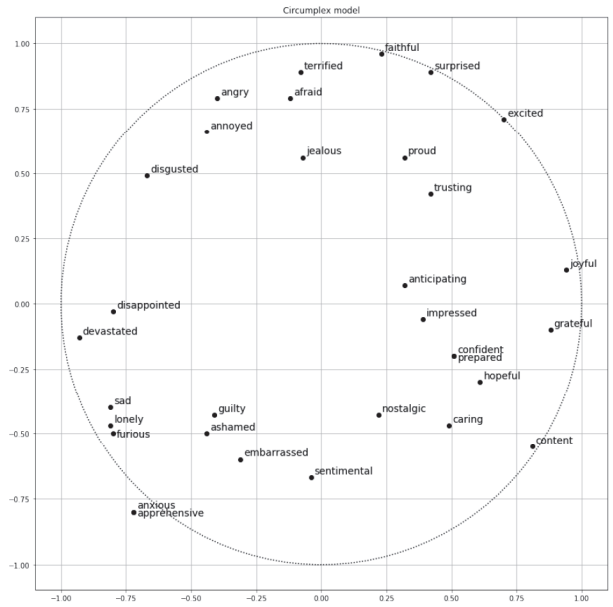

Fig. 2. The two-dimensional circumplex space model adapted from [11], all the emotional labels come from the EmpatheticDialogues dataset. These emotional labels are either exist or near-meaning in the affection model [11]

Russel's circumplex psychological model of affect [10]. The prediction of the arousal and valence value is estimated using a corpus of blog posts from LivcJournal [14]. Figure 2 shows the extracted moods from the LiveJournal dataset and is annotated with the approximated values. Fifteen emotion labels are common between the EapatheticDialogues dataset and the affection model. The synonyms of the rest of the emotion labels can also be found in Figure 2. Table II summarizes all the emotion labels listed by the EapatheticDialogues dataset, where the emotions in the parentheses are the synonyms that come from the model in Figure 2. The estimated values of valence and arousal are also given in Table II. It should be noted that we assign the *prepared* emotional label with the same values as the *confident* emotion because there is no *prepared* from the model of Figure 2, same for the *apprehensive* label. During the training, these estimated values will be used for the regression optimization.

C. Language Model

We use the Generative Pre-training Transformer (GPT) [5] as our pre-trained model. GPT is a Transformer-based model, where the transformer is only based on attention mechanism [7], and it does not rely on any recurrent structure. GPT is a multi-layer transformer with the multi-headed self-attention operation, which is pre-trained on the large BooksCorpus dataset [15]. Furthermore, to enrich more common-sense knowledge for our model, we use the pre-trained model weights from the study of [8]. That model was trained over the Persona-Chat dataset [16] and showed effectiveness on dialog tasks.

To empower our dialogue agent with empathy intelligence, we then fine-tune the model on the EapatheticDialogue

Emotion	Valence	Arousal	Emotion	Valence	Arousal
surprised	0.42	0.89	anxious	-0.72	-0.80
excited	0.70	0.71	anticipating (expectant)	0.32	0.07
angry	-0.40	0.79	joyful (joyous)	0.94	0.13
proud (feeling superior)	0.32	0.56	nostalgic (longing)	0.22	-0.43
sad	-0.81	-0.40	disappointed	-0.80	-0.03
annoyed	-0.44	0.66	prepared (confident)	0.51	-0.20
grateful (pleased)	0.88	-0.10	jealous	-0.07	0.56
lonely (depressed)	-0.81	-0.47	content	0.81	-0.55
afraid	-0.12	0.79	devastated (miserable)	-0.93	-0.13
terrified (alarmed)	-0.08	0.89	embarrassed	-0.31	-0.60
guilty (feel guilt)	-0.41	-0.43	caring (attentive)	0.49	-0.47
impressed	0.39	-0.06	sentimental (melancholic)	-0.04	-0.67
disgusted	-0.67	0.49	trusting (convinced)	0.42	0.42
hopeful	0.61	-0.30	ashamed	-0.44	-0.50
confident	0.51	-0.20	apprehensive (anxious)	-0.72	-0.80
furious (desperate)	-0.80	-0.50	faithful (reverent)	0.23	0.96

TABLE II

THE SUMMARY OF THE LEVEL OF VALENCE AND AROUSAL FOR THE EMOTION LABELS FROM THE EMPATHETICDIALOGUES DATASET [6]

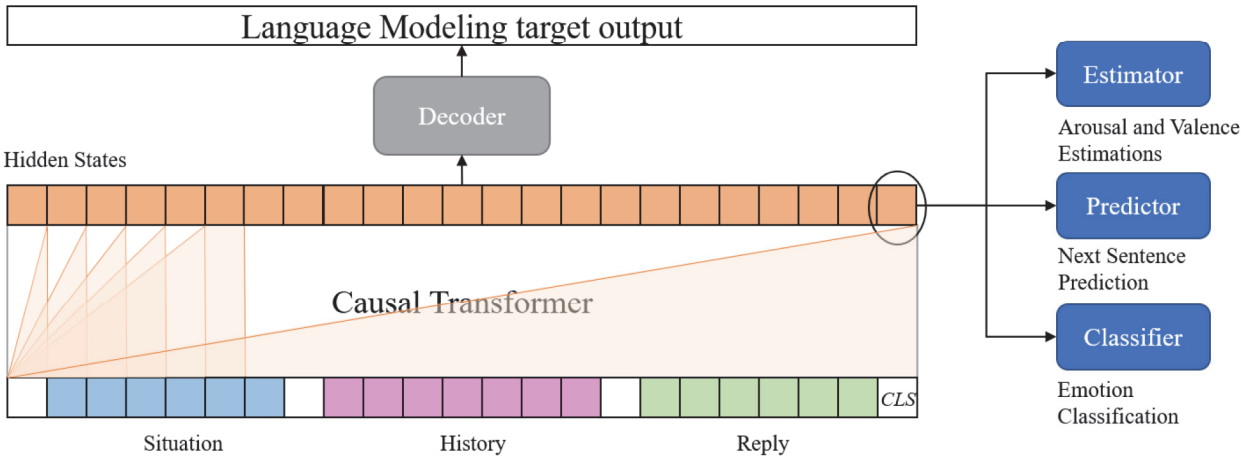

Fig. 3. The language model with emotion classifier and estimator

dataset. As can be seen in Figure 3, we pursue the transfer learning scheme proposed by Wolf et al. [8]. First, the situation, history of the dialogue, the reply, and the special tokes are concatenated to generate word embeddings. Then the segment embeddings and position embeddings are introduced to build parallel inputs by summing them into a single input sequence for the model. The input sequence is then fed into the casual transformer to generate the output sequence tokens. Also, the last hidden state of the model is extracted for the multi-tasks learning purpose. The optimization is performed using the multi-task objectives that contain language modeling, next-sentence prediction, emotion classification, and sentimental estimations.

D. Multi-Task Learning

Our fine-tuning is performed by optimizing a combination of four loss functions: (1) *next-sentence prediction loss*, (2) *language modeling loss*, (3) *emotion classification loss*, and (4) *valence and arousal regression loss*.

For the *next-sentence prediction loss*, a special token [*CLS*] is added to the end of the sequence. The goal is training a linear classifier to differentiate the gold reply from the randomly sampled distractor. The corresponding last hidden state is passed to the linear classifier, and the cross-entropy loss, $L(s)$, is computed to optimize the language model.

For the *language modeling loss*, the final hidden states of the model are used to predict the next reply tokens, and the cross-entropy loss, $L(l)$, is computed to optimize the language model.

For the *emotion classification loss*, a multi-layers classifier is introduced for classifying 32 emotion labels from the EapatheticDialogues dataset. During training, the last hidden state from the final hidden layer is passed to the classifier to predict the emotions, and the cross-entropy loss, $L(e)$, is computed to optimize the emotion model.

For the *valence and arousal regression loss*, a linear layer is used for estimating the arousal and valence values of the dialogues. We then used ReLu to activate the outputs. The last hidden state from the final hidden layer is fed into the linear layer to get the estimations, and the mean-squared

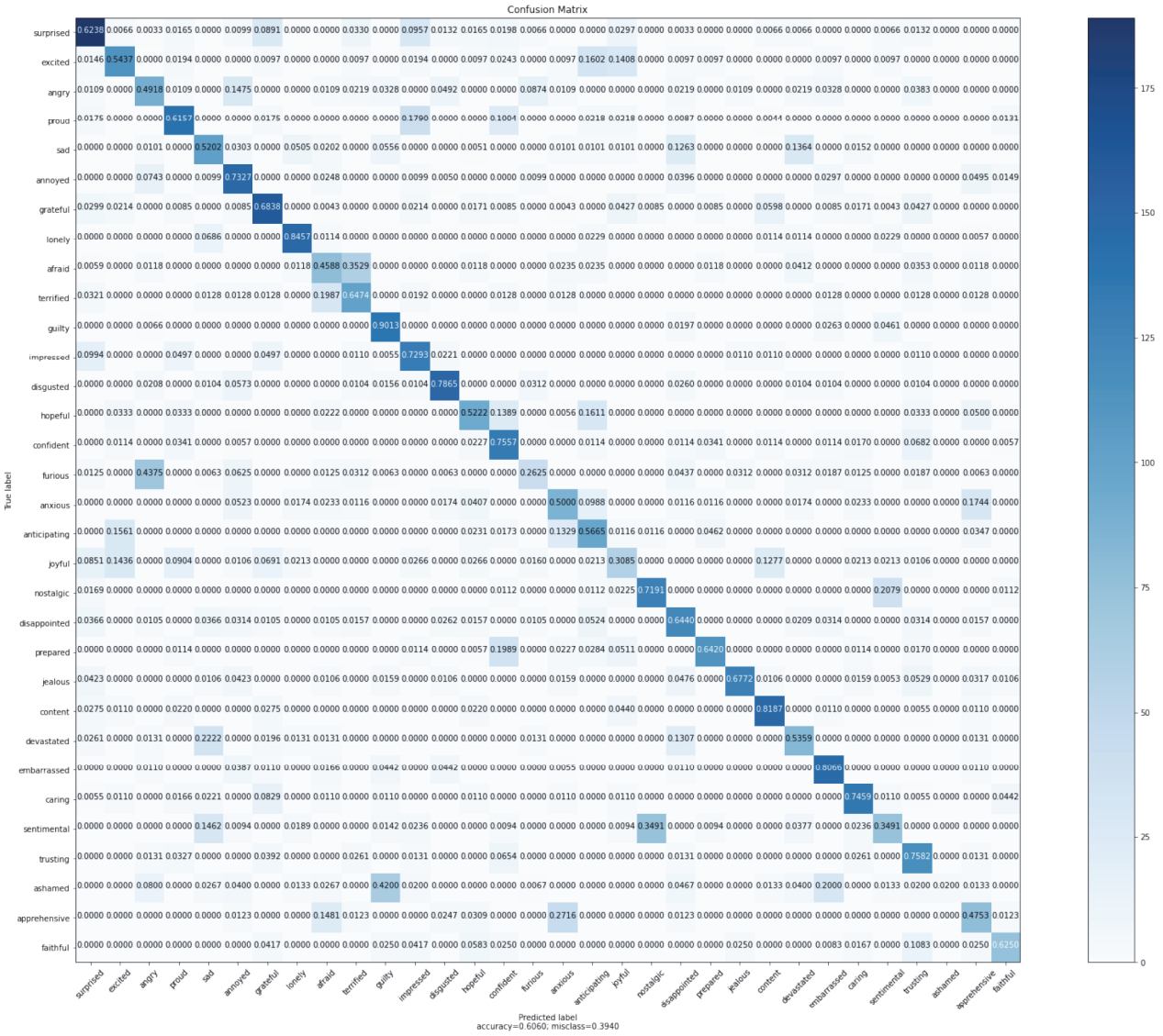

Fig. 4. The confusion matrix of the predicted emotion labels

error loss, $L(va)$, is computed to optimize the emotion model.

Finally, the total loss function is the weighted sum of these four losses:

$$L(total) = \alpha L(s) + \beta L(l) + L(e) + L(va)$$

where α and β are the weights for the *next-sentence prediction loss* and *language modeling loss*, accordingly.

Moreover, during the training, considering the language model was being fine-tuned, but the emotion classifier and estimator we introduced were trained from scratch, so we defined a higher learning rate for the emotion classifier and estimator than that for the language model. We used the Adam optimizer with a learning rate of 6.25e-5 for the language model and a learning rate of 0.01 for the emotional classifier and estimator. Ten epochs of training took about 12h on two V100 GPUs.

V. Results

Models	PPL	EMO ACC	A. MSE	V. MSE
Pretrained [6]	27.96	-	-	-
Fine-Tuned [6]	21.24	-	-	-
Multitask [6]	24.07	-	-	-
EmoPrepend-1 [6]	24.30	-	-	-
Ensem-DM [6]	19.05	-	-	-
CAiRE [9]	**13,32**	0.516	-	-
OUR	17.22	**0.606**	**0.175**	**0.129**

TABLE III

THE COMPARISON OF LANGUAGE MODEL AUTOMATIC METRIC AND EMOTIONAL METRICS BETWEEN MODELS ON THE TEST SET

To evaluate our model, we use the models' results from [6] as the baselines and compare the performance between our model and the model proposed in [9].

- **Pretrained** model is pretrained with the Transformer network on a dump of 1.7 billion REDDIT conversations
- **Fine-tuned** model is fine-tuned over the EapatheticDialogues dataset [6].
- **Multitask** model is trained with multi-task loss by adding a classifier on top of the Transformer to predict the emotion labels
- **EmoPrepend-1** model add a supervised classifier that prepends the top-1 predicted label to the beginning of the token sequence as encode input.
- **Ensem-DM** model concatenates the representation from the external classifiers with the fine-tuned transformer representation as to the input.
- **CAiRE** model deploy the GPT as the pretrained model and fine-tune with multi-task learning objectives. An emotion classifier is added on top of the model to predict the emotions.

We use perplexity (PPL), emotion classification accuracy (EMO ACC), arousal mean-squared error (A. MSE), and valence mean-squared error (V. MSE) as the evaluation metrics, where PPL is a measurement of the uncertainty of the language model.

As can be seen in Table III, our model outperforms the baselines models' performance about metric PPL, but worse than the performance of the CAiRE model. This result is not significant since we introduce more learning tasks by combining more losses. Nevertheless, our model shows better emotion classification accuracy of 60.6%, and the PPL of 17.22 has already given the model the satisfying capacity to generate empathetic responses. Besides that, our model can also output the valence and arousal level for further sentiment analysis.

The confusion matrix of the predicted emotion labels can be found in Figure 4 , where the x-axis and y-axis represent the true label and the predicted label accordingly. From the confusion matrix, it can be noted that most of the emotional classes show dominant correct predictions. However, as can be seen in Figure 4, the emotion *afraid* and *terrified* introduce a considerable amount of incorrect predictions between these two classes, but this is not particularly surprising given the fact that expressing *afraid* and *terrified* will have similar characteristics from linguistics standpoint. For the other emotion labels, most of the misclassification happen in their synonyms. For example, for the *ashamed* emotion, the model inaccurately predicts a majority of the samples to the *guilty* and *embarrassed*, but these three emotions are actually related in various contexts. Because this is only a relatively small size of the training dataset and the challenge of the 32-classes emotion classification problem, this current study has gone some way towards enhancing the dialogue agent's sentimental analysis ability in HRI.

Figure 5 shows the predicted emotions position (red) compared with the annotated ground-truth emotion (black) position from previous study [11]. From the graph, we can see most of the positions of predicted emotions locate near their corresponding ground-truth labels. However, the

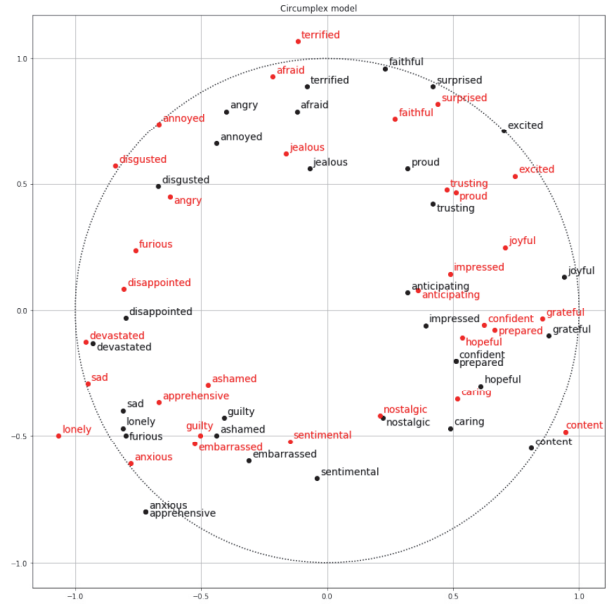

Fig. 5. The two-dimensional circumplex space model adapted from [11] with the predicted emotions labels. The emotions with black color indicate the original annotated positions, and the emotions with red color indicates the predicted positions.

position of the emotion *angry* and *furious* stay too far away from the annotated positions. The confusion matrix can likely explain the reason for this. It is apparent from Figure 4 that most of the samples with *furious* are falsely predicted to *angry*, and the *angry* also has similar characteristics with other emotion labels such as *annoyed* and *disgusted* which introduce the misclassification. These factors could well be responsible for this result.

VI. CONCLUSIONS

This paper proposes an empathetic dialogue agent with the state-of-the-art language model for our robotic system—the agent incorporating a 2D circumplex model for a new way to undertake the sentiment analysis. Compared with previous studies, our empathetic dialogue agent can generate natural responses and detect emotions and estimate the continuous values of arousal and valence in an end-to-end manner. The dialogue agent model is fine-tuned based on multi-task learning over a dataset with empathetic characteristics using a large-scale pre-trained language model. The outputting emotional information empowers the robot to perform sentiment analysis and react with optimal behaviors. By leveraging the speech recognition module and the robot platform we use, we can achieve real-time HRI with the dialogue agent as part of our robotic system.

As states in the results, our multi-task learning objectives enhance the language modeling and emotion detection task performance compared with previous methods. The training with the auxiliary task, the estimations of valence and arousal, significantly improve the emotion classifier's performance to 60.6% accuracy compared with the previous study.

It is worthwhile noting that the emotion model for classifying emotions and predicting the level of arousal and valence should be improved in future studies, and one practical way may increase the performance by backpropagating the losses only to their corresponding layers. For example, the cross-entropy loss from emotion classification will only optimize the emotion classifier.

We will apply our dialogue agent to evaluate child-robot interactions and the emotions classification and estimation in our future work. Furthermore, the dialogue agent will be optimized by the collected data from the experiments for building a more natural and personalized empathetic robotic system.

ACKNOWLEDGMENT

REFERENCES

[1] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Affect recognition for interactive companions: challenges and design in real world scenarios," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 89–98, 2010.

[2] G. Castellano, A. Paiva, A. Kappas, R. Aylett, H. Hastie, W. Barendregt, F. Nabais, and S. Bull, "Towards empathic virtual and robotic tutors," in *International conference on artificial intelligence in education*. Springer, 2013, pp. 733–736.

[3] P. Fung, D. Bertero, Y. Wan, A. Dey, R. H. Y. Chan, F. B. Siddique, Y. Yang, C.-S. Wu, and R. Lin, "Towards empathetic human-robot interactions," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2016, pp. 173–193.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[6] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[8] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *arXiv preprint arXiv:1901.08149*, 2019.

[9] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, "Caire: An end-to-end empathetic chatbot," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, 2020, pp. 13 622–13 623.

[10] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[11] G. Paltoglou and M. Thelwall, "Seeing stars of valence and arousal in blog posts," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 116–123, 2012.

[12] "pepper the humanoid and programmable robot: Softbank robotics," accessed 28 march 2021. [Online]. Available: "https://www.softbankrobotics.com/emea/en/pepper"

[13] J. C. Kim, P. Azzi, M. Jeon, A. M. Howard, and C. H. Park, "Audio-based emotion estimation for interactive robotic therapy for children with autism spectrum disorder," in *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, 2017, pp. 39–44.

[14] G. Mishne *et al.*, "Experiments with mood classification in blog posts," in *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, vol. 19. Citeseer, 2005, pp. 321–327.

[15] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[16] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" *arXiv preprint arXiv:1801.07243*, 2018.