EmpathyQA - Discovering Emotions based on Interactive Question Answering

Yunlong Li
yunlongl@sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Zhuo Ning
zna3@sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Jia Yi Wang
jyw30@sfu.ca
Simon Fraser University
Burnaby, BC, Canada

ABSTRACT

Empathy is an active research area in artificial intelligence (AI) and a lack of empathy in AI causes issues such as trust problems in users. As AI or machine learning algorithms are based on mathematical calculations, they do not experience empathy or subjective feelings. In order to make AI technology more relatable and comforting towards humans, we need to augment machines' emotional sensing capabilities. In this paper, we have developed an "empathetic" robot that responds to questions and displays emotional gestures as a response to the question which matches the emotional sentiment of the question. To get an appropriate response and emotion, we used a BERT model to perform context analysis on questions and a finetuned T5 text-to-text transformer for emotion classification. A experiment was conducted where 14 participants assigned a rating to the robot's ability to convey empathy through gestures and the appropriateness of the gestures from a scale of 1 to 7 which achieved a mean score of 3.29 and 3.93, respectively.

KEYWORDS

Empathy, Robotic, Affective Computing, Human-Centered Artificial Intelligence

1 INTRODUCTION

We want to make artificial intelligence serve people better, so that the technology is no longer a cold concept. At the moment when Human Centered AI is starting to develop, we hope that AI can have a significant missing part – empathy.

As Elon Musk announced that Tesla will start to production of humanoid robots in January 2023, the robot will likely to change our lifestyle. The urgency of the trust between machines and humans arise, and the main problem would be that humans have emotion while machines do not. The technology needs to help people to build a better world, not to control humans. As a result, making a machine simulate a similar feeling with humans is an important task that needs to resolved. At present, most of the existing research is based on unilateral analysis, integrating sentiment analysis to the robotic project to realize the interaction, so that the results can be observed and evaluated more intuitively.

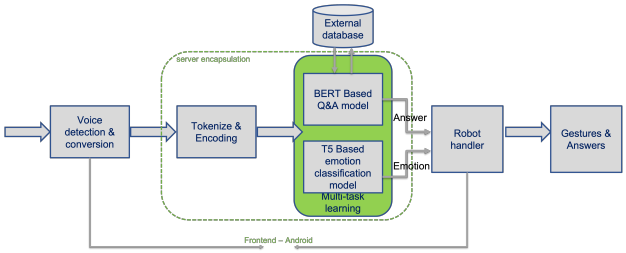
The definition of emotion is a subjective concept, which is difficult to measure, and sentiment analysis relies heavily on contextual information. When the Large Language Model was not popular in the past, the limitations of datasets and computing power greatly affected the practice of emotion in AI technology.

The goal of this project is to develop machine learning models trained on emotional text datasets that is able to detect human emotional expression in real world data and provide a response (answer in an emotional way). We present an approach where machine learning models will be first trained on emotional text with various emotions. We will construct a model that takes questions as an input and outputs an appropriate response and emotion classification based on the emotional sentiment of the question. A robot will be created and the model will be integrated with it, so that it can talk to humans and perform various gestures depending on the emotion class output. Then, a human evaluation will be done where participants evaluate the empathetic ability of the robot and how appropriate the gestures are as a response.

The project confirms the method of detecting emotion through language, and the scalability of the project broadens the application scenarios in different fields. The project has great potential applications in daily care such as accompanying and caring for patients through emotional conversations. It also has great prospects in the travel service industry (e.g. answering the questions of guests or travelers and giving them emotional comfort when they need help according to the situation of different problems).

2 APPROACH

2.1 System



We used two main systems in our approach, the robot agent and machine learning server. The robot is a humanoid robot called Pepper and has a build in Android device. The Android application built in the robot acts as a data collecting agent, which is responsible for receiving speech from the user via microphone and translating voice into text. The robot can capture views via front camera. The server acts as a data processing service, which is responsible for hosting our machine learning models, receiving requests from robot agents, processing data, and responding with model outputs. Our machine learning method will be discussed in the next section.

An isolated Android app is developed in Java and deployed into the robot using Android and the Pepper SDK. The app provides a simple UI for users to interact with. Similar to most of the chatbots on the market, the UI consists of a button, a text label, and a text box. The user needs to put the IP address of the server into the text box to use the application. When users press the button, the Android device microphone will be activated and record the voice.

After translating, text will be shown in the text label which reflects the speech. At the same time, the app will also trigger the robot to respond with gestures.

A simple Python server is published into a computer within the same local area network. The server initializes and loads two trained models into memory and provides several endpoints ready for receiving the robot agent's request. When the robot sends text or images to a dedicated endpoint, the server will transform them into an appropriate format and feed them into dedicated models. The output of our models will be responded synchronously to the robot.

The system is event trigger based, where both systems will be constantly listening to input. Users' interaction with the Android app will initiate the event. The system takes the users' questions as input. After the event is triggered, the robot will first collect the question from the users' and send it to the server. The server will then feed the data into models, and respond back to the robot with a response and an emotion class. Finally, the robot will speak out the response and perform gestures according to the emotion class.

2.2 Methods

Two Machine Learning models were used: A finetuned Bidirectional Encoder Representations from Transformers (BERT) Question Answering model and a Finetuned Text-to-Text-Transfer-Transformer (T5) emotion detection model.

Model 1 – BERT QA: A Finetuned BERT model with embedding layer to calculate the start index and end index in the context for the question, with the expectation–maximization algorithm.

Model 2 – T5 emotion classification: In this more comprehensive model, the downstream training task is to use transfer-learning techniques with emotional data set to obtain a most likely label form six predetermined classes (joy, sadness, love, angry, fear, surprised).

Since the models we use are in the deep learning field, the data required for downstream training comes from high-quality conference emotional datasets (impossible to self generated for Large Language Models).

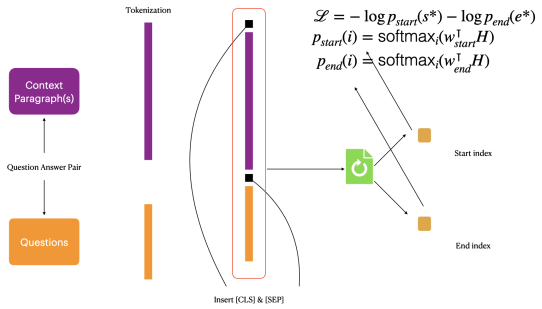


Figure 1: Finetuned BERT QA Model

For BERT QA model, we use SQuAD2.0 dataset[1] to perform the downstream training. For T5 model, the CARER-Processed text[5] data is used for the downstream training. Given that the BERT and T5 are Large Language Models (LLM), the size of data need for downstream training is enormous. The data used for the task is from the conference and benchmark dataset in QA area.

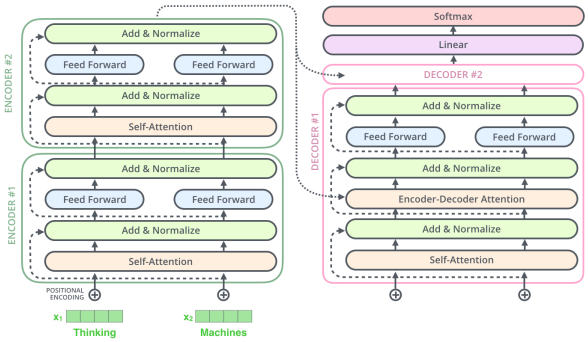


Figure 2: T5 Model illustration
http://jalammar.github.io/illustrated-transformer/

2.3 Approach

For the BERT QA model[1], we need to represent the questions and contexts as the sequences separately and calculate the probabilities by dot producting between each inputs embeddings and the final linear layer in BERT-based models (range of answer). After obtaining the quantitative results, we compare them to the baseline approaches and analyze the differences.

We put the embedding of each word in the text into the input layer of BERT in turn, then insert the "<SEP>" mark, and finally, insert the embeddings of each word in the question in turn to the BERT model.

The loss function we used:

$$\mathcal{L} = -\log p_{start}(s^*) - \log p_{end}(e^*) \tag{1}$$

$$p_{start}(i) = \text{softmax}_i(w_{start}^T H) \tag{2}$$

$$p_{end}(i) = \text{softmax}_i(w_{end}^T H) \tag{3}$$

$H = [h_1, h_2, ..., h_N]$ are the hidden vectors of the paragraph returned by BERT. The loss function of BERT model is the cross-entropy of the output and the target. The output is predicted by a softmax function with the following steps: Each layer produces a vector of word indices and softmax outputs. The word indices will be used to look up the output of the corresponding layer model in the next step. The output of the softmax function is a vector of probabilities. Each probability is the probability of the corresponding word. The softmax output is converted to a probability distribution. The output of the last layer is a probability distribution. For each word index in the distribution, there is a probability that the word is the current word.

Figure 1 is the illustration of the 'context-question' pair construction and how to calculate the loss function.

For the T5 Model[4], we treat the emotion labels as the 'text', and perform the text-to-text transfer transformer training. Mask out sequence of words in the original text contexts and make the model to predict the sequence.

The model was performed with original text layer, input layer, and target layer. As the T5 model achieved the state-of-the-art performance, the parameters are set as the default one as the original model from Hugging Face.

Figure 2 is the illustration of the T5 model structure overall, and the fine-tuning is to change the inputs.

3 EXPERIMENTS AND RESULTS

The robot provides emotion labels (corresponding action) based on the question asked and responds with accurate answers to it in various contexts.

The project is based on closed-domain QA. Contexts limit the scalability, which could be improved through hashing into a comprehensive database.

3.1 Human Evaluation

Participants were recruited online for a human evaluation study on the robot's performance. Each participant watched two videos of a human talking to a robot. In one of the videos, the robot was outputting an emotion class from the T5 classification model. In another video, the robot was outputting an emotion class from a baseline randomization algorithm. In both videos, the robot outputs a gesture that corresponds to the emotion class. Figure 3 shows the gestures that were mapped to each emotion class. The participants were asked two questions for each video: (1) Rate the empathetic ability of the robot (1-7) and (2) Rate the appropriateness of the robot's gestures as a response to the question (1-7) where a higher number means higher satisfaction.

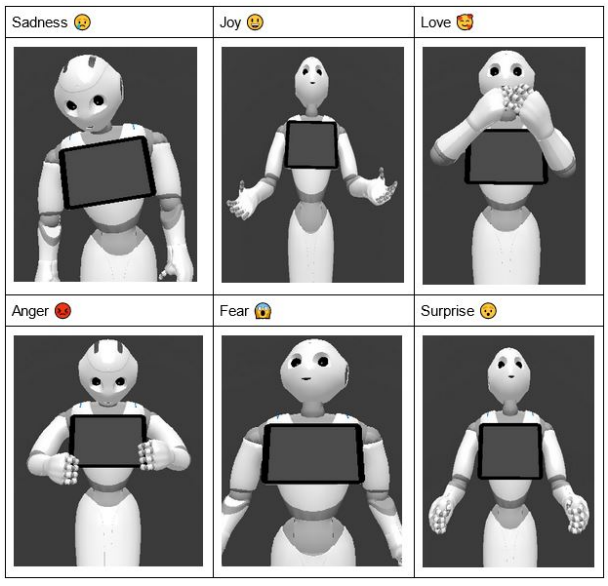
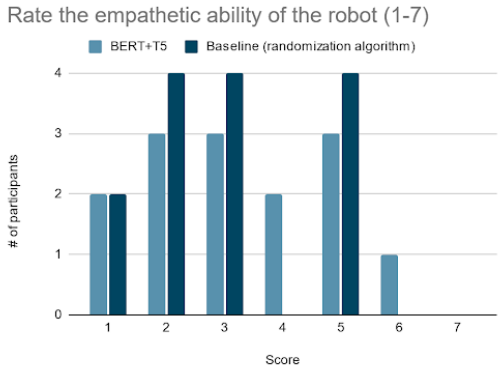
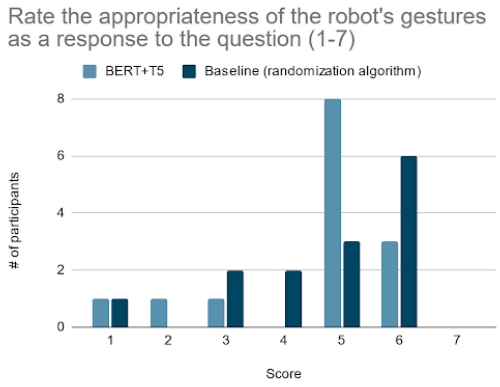


Figure 3: Six gestures that correspond to each emotion

We tested whether there is a significant difference in scores between the robot using the BERT+T5 model and the baseline algorithm. Results are shown in Figure 4. The null hypothesis is that there is no significant difference. An alternative hypothesis is that the scores for the BERT+T5 model will be greater than the baseline on the empathetic ability and appropriate gestures. The scores for the empathetic ability for the BERT+T5 model (M=3.29 , SD=1.53, n=14) and baseline (M=3.0, SD=1.41, n=14; t(0.51)=26.0, p>.05) are not statistically significant. The scores for the appropriate gestures for the BERT+T5 model (M=3.93, SD=1.33, n=14) and baseline (M=4.71, SD=1.48, n=14; t(-1.47)=26.0, p>.05) are also not statistically significant. The Student t-test was used to calculate the p-value. We accept the null hypothesis that there is no significant difference between the scores. The Fleiss' kappa score for all participants is 0.103, which means that there is a slight agreement between raters.



(a) empathetic ability



(b) appropriate gestures

Figure 4: Participants were asked to rate the empathetic ability and appropriateness of the robot's gestures (1-7)

3.2 Examples

Figure 5 shows examples of six questions asked with our model's corresponding response and emotion label. A file relating to the context of the question is passed to the QA model and the model outputs a response based on the context file. The context was gathered

from Wikipedia articles and other various articles on the Internet. The emotion class was given by the T5 model and decides the class by performing emotion sentiment analysis on the question.

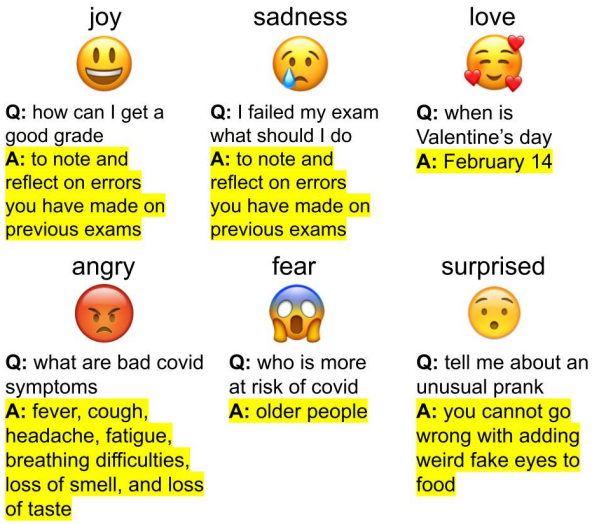


Figure 5: Six gestures that correspond to each emotion

4 DISCUSSION

4.1 Discussion on human evaluation

The reason why there may not be a significant difference between empathetic ability for the two approaches may be caused by the lack of facial expression of the robot and having no difference in tone of voice for different emotions. The robot also did not express sentences that convey empathy (e.g. "I'm sorry to hear that" for sadness). These missing features will be implemented in further work to make the robot appear more empathetic.

The limitations that come with making gestures that are deemed "appropriate" are the fact that different people may interpret body language differently, depending on factors such as culture. This may be why the inter-rater agreement score for Fleiss' kappa suggests a slight agreement instead of a substantial agreement.

There can also be multiple gestures and emotions that can be considered appropriate for a response. There are also many gestures that can convey one emotion and some gestures may be more suitable than others depending on the context or the position of the emotion on the Pleasure-Arousal-Dominance (PAD) space. In future works, we will create multiple gestures for each emotion class and choose these gestures based on context and the emotion's position in the PAD space to get the robot to select a more appropriate response.

4.2 Discussion on examples

From the examples shown in Figure 5, it appears that certain keywords in the question correlate to the emotion class that gets selected. For example, using the word 'good' gets associated with joy, using the word 'failed' gets associated with sadness, and using the word 'bad' gets associated with anger. The limitation of this method is that there may be cases where this approach may not be an accurate way of capturing emotion sentiment. For example, having the user express a question in a certain tone of voice or making certain facial/body expression may change the context of the expressed emotion. Further work will be done to incorporate audio or a spectrogram of a person talking and a picture of their face/body into an audio processing and/or computer vision model and this will be factored into choosing an emotion class.

5 CONCLUSION

Part of the purpose of emotional communication is achieved through the interaction between humans and robots. At the same time, making systems with human-like emotions is a very challenging problem. First, the vast majority of current models and algorithms are based on mathematical and statistical models, which sets up barriers to adding subjectivity factors. Second, the evaluation criteria of the system is very vague, because emotion itself is a subjective concept, and each individual has a different criteria for evaluating someone's emotions or determining whether a system is empathetic. This makes it difficult to improve the performance of the model in the human evaluation.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[3] Hugging Face 2021. Hugging Face. https://huggingface.co/ https://huggingface.co/.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[5] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3687–3697. https://doi.org/10.18653/v1/D18-1404

A ANSWER TO DATASHEETS FOR DATASETS

A.1 3.1 Motivation

A.1.1 For what purpose is the dataset created? Used for emotion classification

A.1.2 Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Saravia, Elvis and Liu, Hsien-Chi Toby and Huang, Yen-Hao and Wu, Junlin and Chen, Yi-Shin, they made the dataset for publishing the paper in ACL

A.1.3 Who funded the creation of the dataset? N/A

A.1.4 Any other comments? N/A

A.2 3.2 Composition

A.2.1 What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The text conversation with emotion labels.

A.2.2 How many instances are there in total (of each type, if appropriate)? train: 16000 validation: 2000 test: 2000

A.2.3 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? All possible values we intended to have

A.2.4 What data does each instance consist of? A text string and a emotion label

A.2.5 Is there a label or target associated with each instance? Yes. Text (string) and label (class label represented as an integer)

A.2.6 Is any information missing from individual instances? No

A.2.7 Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? No

A.2.8 Are there recommended data splits (e.g., training, development/validation, testing)? It has already been splitted where train=16000, validation=2000, and test=2000. No rationale was provided.

A.2.9 Are there any errors, sources of noise, or redundancies in the dataset? No

A.2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? It is self contained.

A.2.11 Does the dataset contain data that might be considered confiden- tial (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)? No

A.2.12 Does the dataset contain data that, if viewed directly, might be of- fensive, insulting, threatening, or might otherwise cause anxiety? There might be some swear words in some of the tweets collected.

A.3 3.3 Collection Process

A.3.1 How was the data associated with each instance acquired? The text data was acquired from Twitter. The annotated emotion class label was obtained using a CNN model called CARER.

A.3.2 What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? The twitter API was used to collect tweets for the data. The twitter posts can be shown in the dataset.

A.3.3 If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Manual human curation. It is unknown how it was validated.

A.3.4 Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Information not available.

A.3.5 Over what timeframe was the data collected? 2018. The tweets could be made any time earlier than the day it was collected.

A.3.6 Were any ethical review processes conducted (e.g., by an institu- tional review board)? No

A.4 3.4 Preprocessing/cleaning/labeling

A.4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of miss- ing values)? Yes. The Twitter data was tokenized by white spaces and preprocessed by adding a lower case and replacing user mentions, hashtags, and URLs with a placeholder.

A.4.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? No

A.4.3 Is the software that was used to preprocess/clean/label the data available? No

A.4.4 Any other comments? No

A.5 3.5 Uses

A.5.1 Has the dataset been used for any tasks already? Emotion classification.

A.5.2 Is there a repository that links to any or all papers or systems that use the dataset? https://github.com/dair-ai/emotion_dataset

A.5.3 What (other) tasks could the dataset be used for? Downstreaming training the NLP model

A.5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No

A.5.5 Are there tasks for which the dataset should not be used? On Github, it says that the dataset should only be used for educational and research purposes. So it shouldn't be used for any other purpose besides that.

A.5.6 Any other comments? No

B CONSENT FORM

Double click on the link to download the consent form.

C CONTRIBUTIONS

Yunlong Li:

- Training the BERT QA Model and T5 Emotional Model
- Processing the datasets for two machine learning models
- Outlining the structure for the integration of Model and Server
- Setup the project, poster and report write up

Zhuo Ning:

- Developing machine learning server in python
- Integrating machine learning models into the server
- Establishing connection between server and robot agents
- Training computer vision model

Susan Wang:

- Development of the Android application
- Development of the Pepper robot actions and gestures
- Drafting consent forms and conducting experiments
- Integration of server into the application

Shea Janke (No code contribution):

- Helped (follow the website link) integrate speech to text and resolve crashes in the android application
- Shea encountered issues with running the Pepper application locally, but assisted in the Android development with some ideas and suggestions.

D IMAGE SEGMENTATION AND CAPTION GENERATION APPROACH

In 2018, BERT: Pre-trained Bidirectional Transformers [2], an extension of Transformer encoder, will leap the performance to a new stage. Through stacking Transformer encoders and pre-training with general tasks using large amounts of data, BERT can generalize to most of the existing NLP problems and achieve better results in most of them. Moreover, BERT is also able to adapt visual tasks. Hence, in this final model, we want to see if we can task advantage on a BERT-like model to get better image captioning results.

After searching, we found VisualBERT, a pre-trained model that is designed to help solve many version-and-language problems such as visual question answering and visual reasoning. We can easily use it since it is integrated by huggingface [3]. In our final solution, we used it to transform the image features to more language-related embedding before we use another Transformer decoder to decode it.

Compared to using a typical Transformer encoder as the caption generator encoder, we believe the pre-trained VisualBERT model that is designed for cross CV and NLP tasks performs better. Note that VisualBERT is trained with features extracted from Mask R-CNN. Thus, the use of VisualBERT also gives another reason to use Mask R-CNN.

In this final model, we used the VisualBERT and a Transformer decoder as our caption generator. The Transformer decoder has 6 layers and 8 heads.