

A Chatbot That Actually Cares

Anjali Poornima Karri

Data Science and AI

Chalmers University of Technology

anjaliip@student.chalmers.se

Kuvalaya Datta Jukanti

Data Science and AI

Chalmers University of Technology

kuvalaya@student.chalmers.se

Abstract

As chatbots are becoming increasingly popular, with the recent progress in machine learning and natural language processing techniques, people have found new ways to utilize them. Having said that, in the recent years, the aim for providing more deeper and emotionally stocked conversations to users has become popular. Therefore, the motivation for this project is to build an empathetic chatbot having the competence to listen to people, understand the user's state and respond empathetically. This is done by considering a pre-trained GPT2 model and fine tune it on an empathetic dataset and check if the model improves its empathetic capability. One of the major challenges for this project is the evaluation metrics. Since there is no specific metric for evaluating the empathetic capability, the evaluation is done by letting people subjectively rate their experience with the chatbot.

1 Introduction

Chatbots, also known as conversational agents is a computer program that simulates a conversation with the users in natural language through text messages or voice commands. Fundamentally, there are two types of chatbots, command-based and intelligent/AI based bots. While command-based chatbots work on a predefined set of rules and hence can only handle a limited set of questions or use cases, the AI/intelligent chatbots are developed using machine learning and natural language processing techniques to provide intelligent answers by understanding the human language. Today, there is a surge of interest for businesses to

design and develop chatbots and to use it for various purposes. But, many of these chatbot interactions are transactional in nature such as "booking a flight ticket", "ordering food" etc. With the rise of technology from basic logic to complex AI together with chatbots becoming more smarter with developments in NLP, chatbots are finding new uses and applications. Having said that, recent attempts to design personalised and more human-like conversations have become popular. One such application is the development of emotional chatbots to support people with their issues or problems.

Mental-health chatbots act as a helping hand and provide a supportive presence to people helping with their anxiety issues or depressive symptoms or even cheer them in happy situations. In other words, it can be considered as a platform where people can express all their sadness, concerns, issues, worries and even happy moments. The chatbot can act as a friend, analyse the mood of the user and respond with empathy. Additionally, the chatbot can even possess the ability of providing advices to help users tackle with their negative thoughts. This way, it helps users to take support and talk to a bot resembling like a therapist and deal with their issues.

2 Related Work

The previous work on the chatbots are either task-oriented or non-task oriented with different approaches such as corpus-based and rule-based methods. Most of the current day chatbots are task-oriented with a corpus based approach. However, these corpuses are very difficult to build for each use case and hence make use of pretrained models which are trained by large datasets. There has been wide range of work in distributional representation for many emotion classification tasks that is built of deep networks pretrained on large

corpus of labelled data collected from social media platform published on Twitter. [1] Several other works were also attempted to make dialogue model more captivating but were only restricted to personal contexts (Eg. I am from Sweden). [5] As mental-health chatbots are gaining immense popularity, the fundamental work involving the building of empathetic dialogue dataset and the creation of empathetic chatbots which marked as a standard for empathetic dialogue generation used a full transformer model. [4] The next major work is the expansion of above mentioned work by using GPT pretrained on a books dataset [2] and then fine tuned on the empathetic dataset. The major difference with the performance is that they have used GPT (Generative Pre-trained Transformer) which is a transformer-decoder model instead of full transformer model.

3 Methodology

GPT2 [3] is a language model structured by stacking transformer decoder blocks. This is a generative model that is used to predict and generate text based on the previous words. This helps us to use GPT2 language models because the main aim is to generate text (generate responses based on the user input). These models generate a probability distribution for the next word based on the given input sequence and are generally trained using large corpus of text with the objective to predict successive word based on previous sequences for longer time and using powerful computational resources. In this project, the pre-trained GPT2 model from huggingface is used and then fine tuned on large-scale multi-turn empathetic dialogue dataset.

3.1 Data

The text corpus carried out for this work is a large-set multi-turn empathetic dialogue dataset gathered by using the ParlAI platform to interact with the Amazon Mechanical Turk. It contains the text in a one-to-one conversational setting which involves one speaker and one listener. The data is from the discussions happening between those two people talking about a situation that happened in their lives. In other words, the speaker talks about a feeling and the listener understands the emotion and responds empathetically. Altogether, the resulting dataset consists of around 25 thousand conversations gathered from 810 participants with 32 evenly distributed emotion labels. The dataset

has a split of 35k conversations for train and 3.5k conversations for test and validation respectively.

3.2 Preprocessing

The preprocessing is a crucial step as our data should be processed accordingly to fit the input to the GPT2 model in order to use the pre-trained language model to adapt to the problem use-case. Initially, the csv files from the datasets are processed. This is because the dataset from the empathetic dataset homepage has mixed conversations where two or more conversations are appended into same rows and additionally, the data has inconsistent number of rows which could not be loaded directly. Therefore, the rows are processed according to their lengths. The text is processed so as to replace the '._comma.' with ',' and adjusting the space character and end marks to complete the sentence properly. Next, the dialogues are tokenized based on the model we use. These tokenized dialogues are structured as a list of lists. Each of this list represent a conversation. Each of this list is also a list of lists with utterances from alternative speakers. Finally, these tokenized utterance are stored as pickle files which are later used while creating data-loaders.

3.3 Training

The original GPT2 model is trained with a single input sequence. For the current use case, there are many contexts including the history of the conversation and the speaker of the utterance. To build the input for the model to fine-tune, the dialogues from each speaker along with the history are concatenated. Thus, additional special tokens <s1> and <s2> are used to distinguish between different speakers. To indicate the meaning of each segment, segment embeddings are also provided along with the input token embeddings and the positional embeddings. Thus the data for training is loaded from a data-loader with the data structured as described above. For training, GPT2LMHeadModel architecture is used from transformers python library¹. This is a GPT2 model with a linear layer which is a head that is used to give weights for tokens in the vocabulary to chose the next word. The vocabulary is extended with the new special tokens. The model is fine-tuned with AdamW optimizer with a learning

¹https://huggingface.co/transformers/model_doc/gpt2.html#gpt2lmheadmodel

rate of $5e-4$. Early stopping criterion was used to stop training where the validation loss did not decrease for 5 steps. Clipped gradients are used so as to reduce the problem of exploding gradients with max norm of the gradients set to 1. The gradients are accumulated for 2 steps before the parameters are updated. Later, the loss from the model, that is, the language modeling loss calculated as part of next token prediction task is logged to check and evaluate the model’s performance. After each epoch, the validation loss is also logged and plot is shown in figure 1.

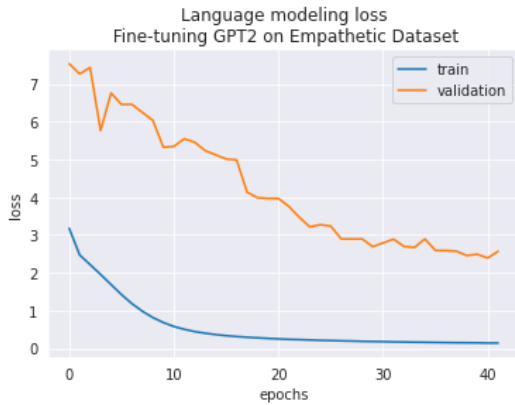


Figure 1: Training and Validation loss after fine-tuning the GPT2 model with empathetic dataset

3.4 Inference

Lastly, the fine-tuned model is used to interact with the user to act as a chatbot. For this interaction, it is important that the model selects the next tokens in order to generate a complete sequence as a response. The most widely known techniques for language generation are Greedy-decoding and beam-search. The Greedy-decoding technique chooses the highest probable word at each time step until $\langle \text{eos} \rangle$ token appears. Beam-search supports a beam of several possible sequences and finally selects the best sequence among the beams. Recently, top-k and top-p are two other sampling techniques which have become more popular for sequence generation. These sampling methods predict the next token from a filtered subset of the vocabulary. This filtering is done by taking top-k most probable tokens or top-p sampling by removing those tokens whose cumulative probability is beyond p. In this project, we have used top-p sampling method to give the response to the user input with $p = 0.9$.

4 Evaluation and Results

4.1 Evaluation

Evaluating responses of conversational agents such as chatbots, without human intervention, is strenuous and complicated. Previous works have experimented on automatic evaluation of responses using BLEU metric. BLEU (Bilingual evaluation understudy) is commonly used for machine translation. The drawback for using BLEU for conversational systems is that it considers word-overlap with the ground truth responses which is not the case with chatbots.

Therefore, the current best method to analyze and evaluate a chatbot is by human judgement. This was done by conducting a survey to about 50 people and determine the chatbot’s performance. People are given a situation like an incident and are asked to convey their emotions with the chatbot. The user then analyzes the responses from the chatbot and rate it accordingly. To make it possible to rate on different aspects, some metrics are designed to evaluate the empathetic quality of the chatbot. The metrics used for this evaluation are *dialogue quality*, *human-likeness*, *fluency*. These set of metrics constitutes the overall performance of the chatbot. Each of these metrics are rated between the range of 1 to 5. (1: very bad, 2: bad, 3:neutral, 4:good, 5: very good)

1. **Fluency:** One of the major factors for evaluating a conversation agent such as a chatbot is based on how well the chatbot is expressing itself easily and articulately. This can be determined by judging on whether the user has understood the responses and if the language seemed accurate.
2. **Dialogue Quality or Relevance:** This metric is used to measure the quality of the chatbot’s response. This can be determined by understanding if the response was a “sensible” reply or “a strange but understandable” reply or a totally “non-sensical” reply and rate accordingly.
3. **Human-likeness:** Human likeness is basically to determine if the responses from the chatbots manifests the understanding of the emotions of the person who is talking about their feelings or experience. In simple words, this metric is used to evaluate on how well the chatbots shows empathy.

4.2 Results and Discussion

Figure 1 shows the training and validation loss of the model. The model training has not been as

expected. The training loss has been decreasing constantly to a very lower value but the validation loss hasn't been decreasing as expected. After 40 epochs of fine-tuning, the chatbots i.e. original GPT2 and finetuned GPT2 have been evaluated by the people.

```
##### Start the Chatting #####
user: I'm sick
bot: I'm sick of you.
user: I am suffering from fever
bot: I am suffering from a fever.
user: What should I do?
bot: I'm sick of you.
user: bye
bot: I'm sick of you.

##### Start the Chatting #####
user: I'm super happy
bot: I'm super happy too.
user: my father is coming
bot: I'm super happy too.
user: bye
bot: I'm super happy too.
```

Figure 2: GPT2 chatbot output

Figure 2 and figure 3 illustrates some of the chats with the chatbots. It is clearly seen that the responses from the GPT2 model are not relevant to the user input. It is observed that it uses the history of the chat to replicate the answers. Although, the sentences are grammatically correct but it does not show any indications of showing empathy. On the other side, by looking at the chats from the fine-tuned GPT2 model, the responses from the chatbot are not so accurate but was comparatively better in understanding the context of the speaker. It can be seen that the fine-tuned model asked questions relevant to the user input but sometimes it just repeated from the history.

```
##### Start the Chatting #####
user: Hi
bot: Hi.
user: I'm sick
bot: I'm sick too.?I hope you are feeling good.
user: Not so, I have fever
bot: I'm worried about you because you seem so bad.
user: Need to take some rest
bot: Yeah.?Can you play a few songs to feel better?
user: good idea, thank you
bot: You're welcome.
user: Bye!
bot: Yes.

##### Start the Chatting #####
user: I'm afraid
bot: What's going on?
user: I'm alone at home
bot: You must be very stressed out.
user: no, my parents went out, I'm alone
bot: It must be very sad.
user: bye
bot: Hope you get better soon.
```

Figure 3: Fine-tuned GPT2 chatbot output

As mentioned above, these models are evaluated by people where they chat with the chatbot while given a situation. This is done for both original GPT2 model and also with the finetuned GPT2

model. Table 1 shows the average score given by people for each metric. By comparing the results, it is understood that people have found the fine-tuned model to respond relevantly to their situation and most of the people felt that the fine-tuned GPT2 model responded emphatically when compared to the original GPT2 model. The fine-tuned GPT2 model couldn't handle longer emotional inputs. But the fluency of models are almost similar because this metric is dependent on the original language model that generates the sentence.

Metric	GPT2	Fine-tuned GPT2
Fluency	3.7	3.74
Dialogue Quality	1.5	3.48
Human-likeness	1.54	3.76

Table 1: Average score for each metric explained in sub-section 4.1 for original GPT2 model and GPT2 model fine-tuned on ED dataset

5 Conclusion

This work is focused on building a chatbot with the ability to respond to a user in an empathetic manner by finetuning the GPT2 model with the empathetic dataset and evaluate the empathetic capability based on human judgement. The survey and the results showed that the fine-tuned GPT2 model yielded better empathetic capability when compared to the original GPT2 model.

References

- [1] Venkatesh Duppada, Royal Jain, and Sushant Hiray. SeerNet at SemEval-2018 task 1: Domain adaptation for affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 18–23, June 2018.
- [2] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. Caire: An empathetic neural chatbot, 2020.
- [3] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [4] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: a new benchmark and dataset, 2019.
- [5] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? pages 2204–2213, July 2018.