

Team Transformers: Final Report

CSE 481 N

Brock Grassy, Dan Le, Kaushal Mangipudi

June 2020

Abstract

In this Natural Language Processing Capstone Project we aim to develop a chatbot that behaves like a depressed patient seeking help. The chatbot would be a training tool that therapists could use to practice talking to and helping patients. We found that training a double headed GPT-2 model on therapy conversations is an effective strategy, but the quality of the resulting chatbot is heavily dependant upon the size of the training dataset. There exists real potential for a conversational model trained on a very large dataset to serve as an effective imitation of depressed patients.

1 Introduction

Mental illness is prevalent throughout the United States and there is a lack of mental health care providers equipped to respond to clients within the critical time frame [1]. Training professional psychotherapists is too expensive and time-consuming [2]. Additionally, practice sessions need to be supervised and require the presence of actors due to real clients being more severe than they appear in initial screenings [3]. We wanted to provide a way for helpline counselors and volunteers to practice and train without affecting people who could respond negatively to poorly asked questions. To alleviate the barriers to training counselors and therapists, we developed a chatbot that aims to mimic a typical helpline client.

Tanana et. al [2] created an automated patient called ClientBot which interacted with participants and gave real time feedback. Their focus was on how participants responded to feedback and if it helped them develop more desirable counseling skills. While evaluation is an important step in helping train therapists, we lacked the therapy expertise and access to experts that would be necessary to develop an evaluation system that would give meaningful feedback for therapists to improve. Hence, we opted to focus on improving the chatbot itself so that it will be able to respond to questions well and coherently.

In this paper we go over an approach that uses GPT-2 and the modifications we made involving sentiment. Finally, we review feedback from psychiatrists about the chatbot and explore possible avenues for expanding the chatbot's capabilities.

2 Model Details

We relied heavily on Huggingface's Conversational AI model for this approach. The basis for this approach is the GPT2DoubleHeadedModel, which consists of a GPT-2 Transformer and two heads that consist of two linear layers, one head for language modeling and the other for multiple choice classification. The model takes in information about the subject and generates text in the form of a 'persona', and then uses that context along with the dialogue history to generate responses to user input and serve as a conversation partner. We used model and then finetuned it using our transcript data from the Distress Analysis Interview Corpus (DAIC) dataset.

The training data consists of conversations between a therapist and patients seeking mental health treatment. Each conversation consists of a collection of descriptive sentences that contain information about the patient (e.g. ‘I am a writer’, ‘I regret dropping out of school’, etc.) that constitute a persona, the actual conversation between the therapist and the patient, and for each actual sentence said by the patient, a series of alternative ‘candidate’ sentences that the model must distinguish the actual sentence from. For example, in response to the question ‘how are you feeling,’ the patient might have actually said ‘I’m feeling down because I can’t find a job,’ and alternative candidate sentences could be ‘I’m feeling great’ or ‘I’m feeling excited about talking with you’. The task the model is trained on is to, given a persona and a question the therapist has asked, distinguish the actual sentence the patient said from the alternative candidate sentences. In the aforementioned example, the training task would be to, given the persona of the patient in the conversation and the question ‘how are you’, predict that the patient said that they’re feeling down instead of feeling great or excited.

3 Dataset

To train our model, we made use of the DAIC-WOZ database. The database is part of a larger corpus of clinical interviews, the DAIC, that was assembled by the USC Institute for Creative Technologies. The specific subset of the data that we focused on was comprised of 189 transcripts of conversations between research participants and a computer interviewer called Ellie. Ellie had built in responses which were selected by a human supervisor. Because of this, the dataset is rich with answers to the same questions, but the types of questions asked were limited and not fully representative of a helpline call.

To convert the DAIC data into the format used by the model for training, we added personas to each transcript and candidate sentences for each line spoken by patients. We manually wrote personas for each transcript by reading the transcripts and writing descriptive sentences about patients based off what they said over the course of the conversations. For example:

i am from santa monica	i like horses and nature
i don't like how crowed los angeles is	i have ptsd
i studied psychology	i have depression
my dream job is to help other people	

We generated a pool of candidate questions by iterating through all the transcripts, and for each question asked by a therapist, we would create a set of all the different answers given by various patients to the question. For example, for the question ‘do you have any regrets,’ the set of potential candidate answers might consist of ‘I regret dropping out of school’ and ‘I regret getting married so early.’ Then, when generating training data from a transcript, for each question the therapist asked, we would randomly select 4 candidate answers from the pool of answers that correspond to that question. The vast majority of questions therapists asked had at least 4 different answers to them from the 189 transcripts we have, but for those that did not, we randomly selected answers across all questions to supplement the existing answers to get a total of 4 candidates.

4 Manually Tweaking Sentiment

Initial iterations of our finetuned GPT-2 model proved adequate when answering questions, but failed to accurately convey the level of depression that we were expecting the bot to have. As such, the bot failed our evaluation tasks that specifically measured its ability to simulate someone seeking out mental health help. In order to improve performance on these tasks, we decided to include sentiment information directly as input into our model. A standard paradigm used in NLP tasks to include additional information in a model is to directly append that information to the input data [4]. To accomplish this, for each input into our model we

appended a placeholder token and the sentiment classification (either positive, negative, or neutral) to our input's history array.

We used two different strategies to classify sentiment for our input data. For both, we made use of the pretrained Vader sentiment analysis tool provided in the nltk Python module [5]. Vader provides a function that takes a sentence and outputs a “compound” sentiment score that is between -1 and 1, with negative scores corresponding to negative sentiment and positive scores corresponding to positive sentiment. For each conversation in our training data, we calculated the sentiment score for each line uttered by the participant. We classified the sentiment scores into three categories: between -1 and -0.05 as “negative”, between -0.05 and 0.05 as ‘neutral’, and between 0.05 and 1 as “positive”. These inputs were appended to our input data through the previously described method. Including sentiment in this manner and passing in requests for negative sentiment to the model at runtime noticeably impacted our model's behavior. Generally, this change directly caused the bot to output responses rife with negative-sentiment words. However, this impacted the consistency of the bot throughout the course of conversations as all responses tended to be negative regardless of content.

To change this behavior, we modified the sentiment inputs to aggregate over each conversation as a whole. To accomplish this, for all conversations in our input data we found the proportion of messages with negative sentiment as defined by the Vader sentiment model. We then split these conversations into positive, negative, and neutral categories based on these proportions, and passed these inputs into our model. The goal of this strategy was to reflect how conversations typically remain consistent in mood throughout and prevent the volatile changes in sentiment that resulted from the previous strategy. Conversations generated from this model tended to exhibit variance in the sentiment of individual messages while preserving an overall negative sentiment. This architecture was used in the final version of the bot that was evaluated by psychiatrists.

5 Results

For the evaluation of the model we had psychiatrists interact with the chatbot as if they were therapists and the chatbot was a depressed person. The categories were coherence (how comprehensible the responses were), consistency (how consistent was the bot's overall narrative), depression (how depressed it sounded), ease of use (how easy it was to use the bot), and usefulness (how useful they thought the bot would be in its current state for training helpline therapists). The bar graph shows the average rating given among 8 psychiatrists where 1 is a low rating and 5 is a good rating and the error bars represent the standard deviation.

The bot received mediocre scores from the psychiatrists. The chatbot did not sound depressed enough, with an average rating of 2.625 and coherence and consistency are only slightly better. In the free text responses, the most common concerns were that “it wasn't very empathetic” and “its responses almost made no sense or were vague”. The standard deviation for each category was about 1.5, meaning that psychiatrists either rated the model very highly or very poorly, which matches the quality of the recorded conversations.

Conversation A

Therapist: Can you give me more details about your life stressors?

Chatbot: uh my finances.

Therapist: How so?

Chatbot: uh i have a lot of uh bills to pay.

Therapist: Do you feel your finances cause you the most stress in your life?

Chatbot: no.

Therapist: Great. What are some things within your control that can help you financially?

Chatbot: uh my rent payments.

Therapist: Have you considered cutting your financial costs by having a roommate?

Chatbot: no.

Therapist: Is this something you'd be open to try?

Chatbot: no.

Therapist: What are examples of some other stressors?
Chatbot: uh traffic.

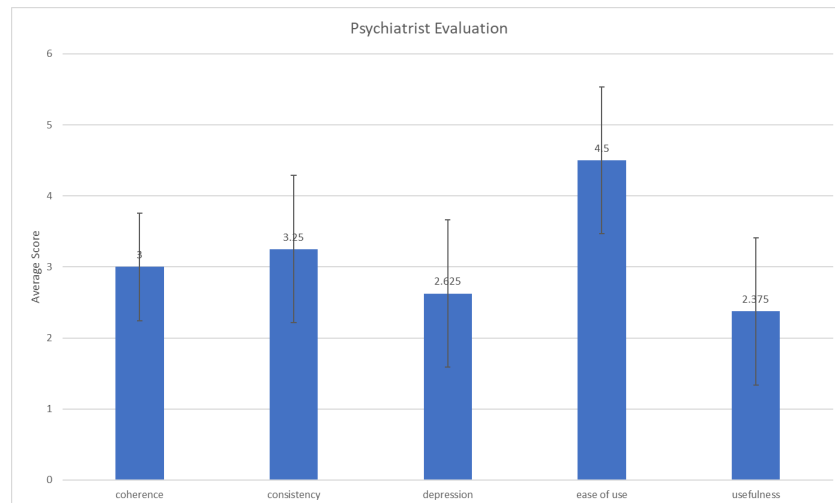
Therapist: Do you have the option to avoid traffic by working or studying remotely?
Chatbot: i do.

Conversation B

Therapist: How are you feeling?
Chatbot: sad.
Therapist: Why are you sad?

Chatbot: uh i'm um divorcing my wife
Therapist: why do you want to divorce your wife?
Chatbot: um she's a very strong and caring person.

Figure 1: Graph of the average ratings psychiatrists gave the chatbot



6 Discussion

Overall, our model was inconsistent. It performed well in some of the conversations with the psychiatrists, but it did poorly in others. One common theme among the conversations where the bot was rated highly was that the questions the psychiatrists asked resembled the questions asked in the training data. Conversation A is an excerpt from a conversation with the chatbot that a psychiatrist rated highly. In this conversation, the psychiatrist mostly asked questions that were asked in the training data. They talked to the patient about financial difficulties, rent, and traffic, which are all common issues in Los Angeles, which is where all the conversations in the training data took place. On the other hand, the chatbot tended to do poorly when asked questions about things that were not discussed in the training data. Conversation B is an example of an interaction that was poorly rated by a psychiatrist. The first two questions are answered properly by the chatbot. These questions appear in our training data quite often. However, the followup question "Why do you want to divorce your wife?" is met with a nonsensical response. This question is specific to the individual and did not appear in the training data at all due to Ellie's question bank being limited. Because of this the chatbot was only able to create a response about his wife but not in the way the psychiatrist intended.

7 Related Work

In "Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health," Althoff et. al. [6] analyzed a large set of counseling conversations to determine what

counseling strategies lead to better patient outcomes. They came up with methods to empirically measure things like counselors' adaptability, creativity, and ability to deal with ambiguity, move a conversation along, and change patients' perspective on their lives, and then used that data to determine some effective counseling strategies. They found that overall, counselors who did better in these 5 categories tended to have better outcomes when working with patients. If we were to implement an automated feedback system that gave therapists feedback on their conversation with the chatbot, their techniques for measuring these factors would prove very useful. We could measure how well therapists did in each of these categories and let them know which categories they did well and poorly in.

8 Limitations and/or Future Work

The biggest obstacle we faced throughout the project was the size of our dataset. GPT-2 is a fairly data hungry model that needs lots of training data, but the DAIC dataset we used only contained a bit under 200 conversations that were each about 15 minutes long. This meant that our model was not able to generalize - it performed well when it was asked questions that were similar to those it encountered in the training data, but they are a small subset of all possible questions the chatbot could encounter. We believe that if we were to train our chatbot model on a far larger dataset, it would be exposed to more questions and be able to answer them consistently. There is a dataset of conversations from an anonymous help line called crisistextline.org that consists of over 100 million messages. We were unable to get access to it for use in our project, but it would be interesting to take that data, process it, and train our chatbot model with it to see how much better the resulting chatbot would perform. Another interesting improvement to the chatbot model would be to add a depression severity parameter that therapists could use to control how depressed the patient sounded. This would allow therapists to practice speaking with patients with varying levels of depression. Another potential avenue to explore in the future is to develop a model to evaluate therapists' conversations with the chatbot and provide them with feedback to improve. Althoff et. al. [6] have done some work in empirically evaluating conversations between therapists and patients, and their findings could be incorporated into an evaluation model that would measure therapists' performance along dimensions like adaptability, creativity, ability to deal with ambiguity, etc. and tell them what categories they do poorly in.

9 Conclusion

In this paper, we sought to develop a chatbot that would talk like depressed patients. Ideally, psychiatrists would be able to practice talking to patients by having conversations with the chatbot. Unfortunately, our model could only effectively respond to the questions that appeared in the DAIC data. The fact that the model performed well on personas and questions encountered in training is a good sign, as it shows that the model is able to learn how to imitate a depressed patient from transcripts of therapy conversations. We believe that a model trained on a far larger dataset would be able to answer a wide range of questions therapists could ask them and serve as an effective tool to help therapists improve their counselling skills.

10 Acknowledgments

We are very grateful to the Valley Medical Center Psychiatry and Counseling Clinic for testing our model and giving us expert insight on how closely our chatbot resembled depressed patients. We would like to thank Ashish Sharma for meeting with us to discuss the applications of machine learning in therapy. His insight was invaluable in helping us understand the ethical implications of our work and the state of the field, and he inspired us to focus on developing a tool to help train therapists instead of treating patients. We would also like to acknowledge Kelvin Lu for working with us to debug hardware issues and giving us guidance on how to implement the sentiment knobs. Noah Smith and Tal August gave us constant feedback throughout our project, and Team Null and Team Trigram gave us valuable feedback on our final paper.

11 Bibliography

References

- [1] Kathleen Thomas, Alan Ellis, Thomas Konrad, Charles Holzer, and Joseph Morrissey. County-level estimates of mental health professional shortage in the united states. *Psychiatric services (Washington, D.C.)*, 60:1323–8, 10 2009.
- [2] Michael J Tanana, Christina S Soma, Vivek Srikanth, David C Atkins, and Zac E Imel. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *J Med Internet Res*, 21(7):e12529, Jul 2019.
- [3] Michael Helge Rønnestad and Nicholas Ladany. The impact of psychotherapy training: Introduction to the special section. *Psychotherapy Research*, 16(3):261–267, 2006.
- [4] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. 2020.
- [5] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015.
- [6] Tim Althoff, Kevin Clark, and Jure Leskovec. Natural language processing for mental health: Large scale discourse analysis of counseling conversations. *CoRR*, abs/1605.04462, 2016.