Abstract

Mental health is a universal problem that we all struggle with throughout our lives, however, many of us do not have access to the resources to aid us. When such is the case, it can be very helpful to have an anonymous chatbot to talk to. This aids the user in being able to seek help with their personal problems without having to go through the traditional method of speaking with a therapist. To do so, we can employ a bot that is automated and available 24/7 to help users when they need it the most. This can be used in various scenarios, such as when a User is going through a tough time and they need advice, or when a user would like to maintain anonymity and still seek help, as well as if therapy is too expensive at the moment. This bot is a transformer model, more specifically the sharded version of falcon-7b, and it has been fine-tuned on a mental-health conversational dataset. The dataset consists of interactions between a human and a mental-health assistant, and is a great resource to train our model on.

Overview

The Problem and why it is interesting:

Mental health is a universal problem that we all struggle with throughout our lives, a constant battle for many. However, many of us do not have access to the resources to aid us in such difficult times. When this is the case, it is very helpful to have an anonymous chatbot to talk to. This aids the user in being able to seek help with their personal problems without having to go through the traditional method of speaking with a therapist. This can be a positive for many as the AI chatbot is free, always available, and it offers accessibvle, stigma-free support. Furthermore, the chatbot is anonymous which allows for an open and safe conversation for the user. Seeking help for mental health can also be difficult due to limiting factors such as cost and stigma, and again the chatbot is both free and stigma-free.

The Approach and the rationale:

The technical portion of the approach is to employ a bot that is automated and available 24/7 to help users when they need it the most. This bot is a transformer model, more specifically the sharded version of falcon-7b, and it has been fine-tuned on a mental-health conversational dataset. The dataset consists of interactions between a human and a mental-health assistant, and is a great resource to train our model on. This technical approach makes sense as the goal is to make conversational-style mental health bot. Thus, fine-tuning a transformer model that is used for conversational tasks already, on a carefully selected mental-health dataset, will yield us desirable results. This is the standard approach when trying to make a chatbot for a specific use-case.

The ethical portion of the approach is to ensure that the interactions are empathetic and comforting. Thus, choosing the correct dataset is key. We want to ensure that the dataset has real-life mental-health conversations that tackle sensitive topics with care. Furthermore, we want to ensure that these conversations are also anonymous to adhere to the AI ethical guidelines. The rationale behind this is that we cant to create a safe space for sensitive mental health discussions, while still balancing ethical guidelines and user privacy and trust.

Experiment Setup

Dataset Description:

The dataset chosen is the “mental_health_chatbot_dataset”, in English. This dataset was chosen because it is designed specifically for the task at hand, which is to train a conversational AI model (chatbots), for the purpose of mental health (our goal). The dataset consists of conversational pairs of questions and answers which pertain to various parts of mental health. Furthermore, all of these conversations are anonymous, thus this adheres to our earlier requirement of wanting to use anonymized data to follow proper AI ethical guidelines. This dataset has already been pre-processed to ensure that there are no unwanted characters present.

Every dataset pair is structured using a text column which is made up of a conversational pair between a patient which is labeled as ‘<HUMAN>’ and then a therapist which is labeled as ‘<ASSISTANT>’. Example of such a pair from the dataset:

Question (HUMAN): "What is a panic attack?"
Answer (ASSISTANT): "Panic attacks come on suddenly and involve intense and often overwhelming fear. They’re accompanied by challenging physical symptoms, like a racing heartbeat, shortness of breath, or nausea. Unexpected panic attacks occur without an obvious cause. Expected panic attacks are cued by external stressors, like phobias. Panic attacks can happen to anyone, but having more than one may be a sign of panic disorder, a mental health condition characterized by sudden and repeated panic attacks."

This dataset has been created through aggregating many popular healthcare websites such as WebMD, Mayo Clinic, and so on. These are all reliable and well-trusted sources within the online health community, thus further validating our choice of this particular dataset. Furthermore, this dataset focuses on sensitive mental health topics as per our requirement from earlier, and has been carefully curated to ensure that all conversational pairs are ethically sourced and are also sensitive towards the issues they contain. Finally, this dataset has also been previously trained on a falcon model, thus, integrating it into our workflow should not be a problem.

Implementation and Model architechture:

To initiate the implementation we must setup necessary libraries as well as our environment. We will be working on Google’s cloud notebook application Colab. Furthermore, we will be using an A100 GPU as it will help speed up training times significantly, making use of NVIDIA’s tensor-core technology. First, we must install various python libraries such as bitsandbytes, torch, transformers, peft, accelerate, datasets, and loralib. Then, we must ensure that our CUDA environment is using GPU acceleration. Continuing to our model selection, we will be using *vilsonrodrigues/falcon-7b-instruct-sharded,* which is a large pre-trained language model. We are using a sharded model here because we want to manage the size of the large-scale neural network. Each of these shards contain some portion of the model's parameters. This is useful as we can manage memory efficiently and avoid memory constraints which may arise while using GPUs. Continuing, we also implement a BitsAndBytes configuration for quantization as this will help us optimize the model's performance even more. We then initialize a tokenizer for our model and we enable gradient checkpointing. We then apply Low-Rank Adaptation or LoRA to adjust the specific layers of the model for a given rank and adaptation rate.

We then move onto the dataset processing. We first load the 'mental_health_chatbot' dataset discussed earlier, and process and tokenize it in order to create prompts which we can use for training. We then configure the training parameters. Here, lets discuss why the specific training parameters were chosen. For the number of epochs we chose 4 as this is a balanced choice. We also want to ensure that overfitting does not occur as falcon is already good at conversational tasks, thus we want to merely fine-tune it for mental health purposes. Continuing, we found that 8 for the batch size achieves a perfect balance between memory usage and efficient training. We set the weight decay to 0.01 to prevent overfitting. Choosing an appropriate learning rate was very important. We had to ensure that the model did not learn too quickly and succumb to overfitting and not too slowly as this could cause underfitting. We chose 1e-4 as it is usually a balanced learning rate for fine-tuning models like falcon. It helps the model learn at a decent rate and avoids overfitting or underfitting. Finally, we enabled F16 as it allows for faster computation on the A100 like quantization, and gradient accumulation steps are set to 8 to provide a healthy balance between memory usage and an effective batch size. We have chosen adamw_torch as the optimizer as it has shown consistent efficiency in large-scale training. We then train the model on the preprocessed dataset. In order to generate responses we utilized the configured tokenizer as well as customized generation settings.

Experiment Results & Discussion

Primary Results:

The primary goal of the experiments was to fine-tune a transformer model to provide sensitive and thoughtful advice and responses to mental-health prompts. We can see that this is true and the model is able to generate contextually relevant responses to various mental-health related prompts. A sample prompt we tried was, "How do I deal with sadness?". The model replied with the following:

```
<HUMAN>: How do I deal with sadness?
<ASSISTANT>: Dealing with sadness can be challenging, but there are
several ways to cope with it effectively. Here are some tips to help you
cope with sadness:
```

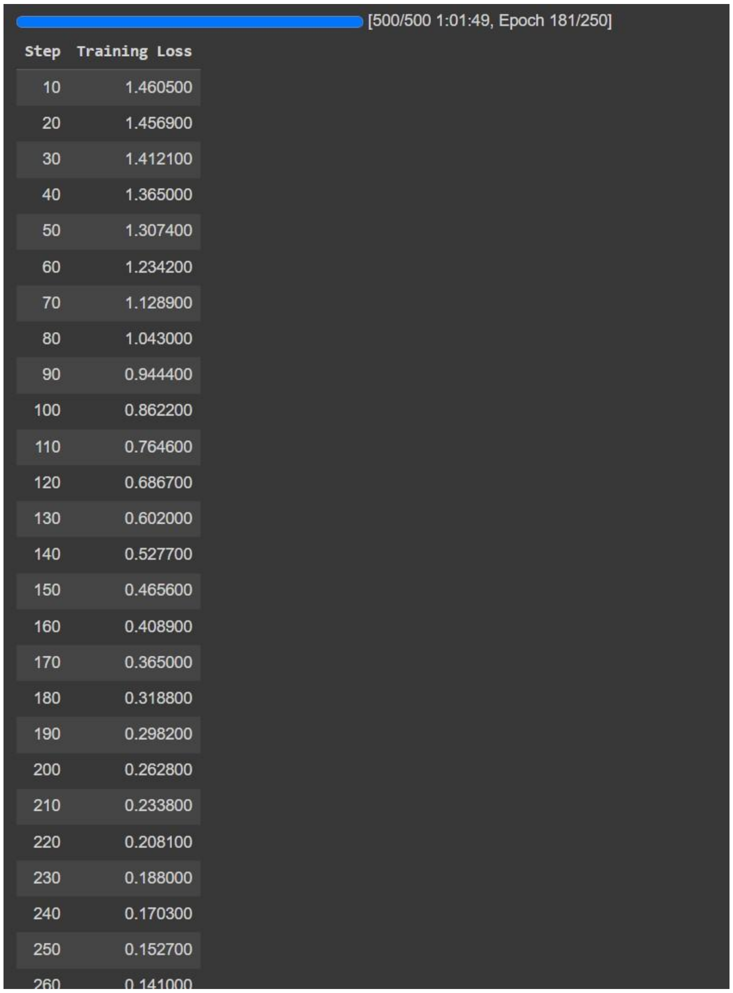
```
1. Reach out: Talk to someone you trust, whether it's a friend, family
member, or a support group member. Sharing your feelings with someone who
understands can make a big difference in your healing process.

2. Practice self-care: Make sure you're getting enough sleep, eating well,
and engaging in activities that bring you joy. Taking care of your
physical health can also help alleviate some of the emotional distress
caused by sadness.

3. Engage in relaxation techniques: Meditation, yoga, or other relaxation
exercises can help you manage stress and emotional pain. It's all about
finding the right balance between taking care of yourself and addressing
your feelings.

4. Set small goals: Break down your tasks into smaller, achievable goals.
This can help you feel more in control of your life and…
```

We can see that the response is detailed, empathetic, and provides applicable advice. It first encourages the user to reach out to people you trust for support. Furthermore, the response emphasizes the that the user should pratice self-care through techniques such as relaxation, as well as setting small goals that are achievable and make you feel good. Another point of reference we have is the loss decreasing steadily over a number of epochs. This indicates that the model was learning the data well.



[500/500 1:01:49, Epoch 181/250]

Step	Training Loss
10	1.460500
20	1.456900
30	1.412100
40	1.365000
50	1.307400
60	1.234200
70	1.128900
80	1.043000
90	0.944400
100	0.862200
110	0.764600
120	0.686700
130	0.602000
140	0.527700
150	0.465600
160	0.408900
170	0.365000
180	0.318800
190	0.298200
200	0.262800
210	0.233800
220	0.208100
230	0.188000
240	0.170300
250	0.152700
260	0.141000

Supplementary Results:

We can now discuss the justification behind choosing certain parameters for generation. First, ‘max_new_tokens’ was set to 200. This is the maximum length of the generated response. We wanted to ensure that the responses were concise and yet still detailed enough for the user to receive the help they needed. We set the ‘temperature’ parameter to 0.7 as we wanted a balance between the randomness and relevance of the response to ensure that it was still contextually correct. ‘Top_p’ was set to 0.7 as well to help generate a diverse response that was still understandable and relevant. ‘Num_return_sequences’ was set to 1 as we wanted the model to generate one sequence per prompt to ensure a quality response. The token and padding id settings were set to handle the sequence endings to stabilize the structure of the generated response.

Comment on Results:

Overall, I am quite pleased with the results. They accomplish the task which was set at the beginning of this project: to generate sensitive, yet thoughtful and realistic responses to mental-health questions and situations. We wanted to create a safe space for the user, and we can see by the generated response, that it meets that criteria and more. Hopefully, this can be useful to anyone going through a tough time that needs some support.

Conclusion:

We have managed to create a mental-health bot that is a safe space for anyone going through a tough time. Furthermore, this bot is capable of generating thoughtful and realistic responses to sensitive mental health topics. We were able to achieve this through selecting an appropriate model for conversational use (falcon-7b sharded), selecting an ethically sourced and reliable mental-health conversational dataset, and then combining the two through carefully chosen hyperparemeters to create a fine-tuned mental health bot. We then tested the bot on a sample prompt and analyzed that it had achieved the criteria we set out at the beginning of the project.

References

Fine-tuning Falcon-7b on MidJourney