SageCrypto: A Chatbot that Knows Cryptocurrency

Cheng Shu

1 Overview

In this final project, I developed a Telegram chatbot based on a finetuned DialoGPT model. As its name SageCrypto indicates, this bot has a relatively good understanding of the domain knowledge of cryptocurrency and can conduct multi-turn dialogues that mimic a real human response.

The full app consists of a back-end with a fine-tuned DialoGPT Model, which is trained with self-defined Reddit data and hosted on HuggingFace, as well as a front-end Telegram bot interface that utilizes Python Flask API and is hosted on Heruku.

2 Back-end

2.1 The Language Model

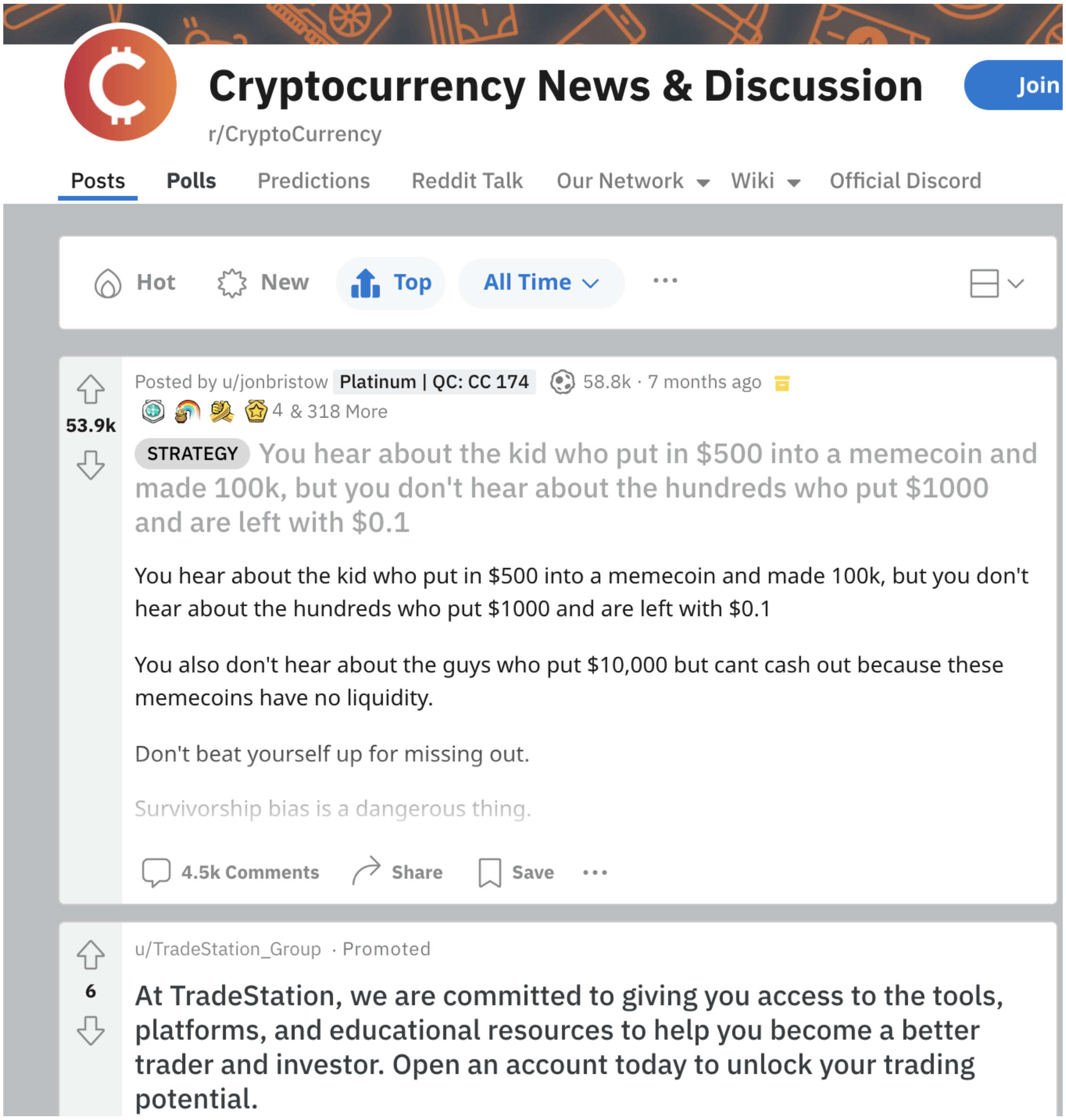
My back-end language model is built upon Microsoft DialoGPT models (Zhang et al., 2020). Microsoft DialoGPT (dialogue generative pre-trained transformer) is a large, tunable neural conversational response generation model that's based on OpenAI GPT-2. DialoGPT is original trained on 147M multi-turn Reddit comment dialogues from 2005 to 2017. It has shown that conversational systems can leverage DialoGPT to generate more relevant, contentful and context-consistent responses. DialoGPT provides three pretrained models: small (117M) medium (345M) and large (762M) models. In this project, I finetuned the DialogGPT small and medium model with data related to cryptocurrency.

2.2 Data

The data I used for finetuning is also from reddit. I chose four reddit communities (see an example in Figure 1) that are most relevant and popular among people interested in cryptocurrency. They are *CryptoCurrency* of 4.1m members, *ethereum* of 1.2m members, CryptoMarkets of 676k members and Crypto_Currency_News of 121k.

I explored three ways to scrape this data: use official Reddit API and access through PRAW (a python package), and use a python PushShift package to download reddit data without limits. While the second option seems more appealing, it turned out that downloading a full copy of a given subreddit is too time consuming and the downloaded dataset is also too large for training. Hence, I opted to the PRAW package. Concretely, I used PRAW to scrape the full comments history of the 20 most popular of all time threads of each subreddit (80 in total). As we already know, Reddit threads form a tree-like structure, as anyone can comment on any comment. But in the current research, I only kept the main sequence of the discussion thread and removed all peripheral branches. The full size of the final data is 3.3Mb.

To prepare the data for training. I need to further specify the "context" and target. As required by the DialoGPT model, the context is critical in helping the model to give context-aware responses. In my model, I define context as the six comments before the current comment (target), and the title of the thread which can also be seen as the root of the comments tree. The rationale behind this definition is for both considering the immediate context and the "macro" context as oftentimes, people participated in the discussion because they saw the title first.



Cryptocurrency News & Discussion

r/CryptoCurrency

Posted by u/jonbristow **Platinum | QC: CC 174** 58.8k · 7 months ago

You hear about the kid who put in $500 into a memecoin and made 100k, but you don't hear about the hundreds who put $1000 and are left with $0.1

You hear about the kid who put in $500 into a memecoin and made 100k, but you don't hear about the hundreds who put $1000 and are left with $0.1

You also don't hear about the guys who put $10,000 but cant cash out because these memecoins have no liquidity.

Don't beat yourself up for missing out.

Survivorship bias is a dangerous thing.

u/TradeStation_Group · Promoted

At TradeStation, we are committed to giving you access to the tools, platforms, and educational resources to help you become a better trader and investor. Open an account today to unlock your trading potential.

Figure 1: CryptoCurrency Subreddit Top Thread Page

2.3 Training and Deployment

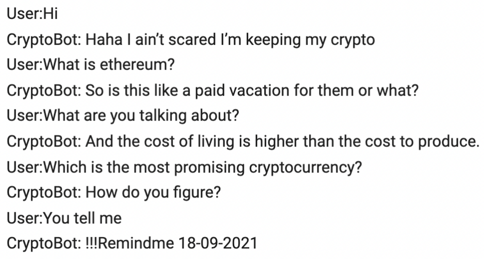
The training is mostly based on the scripts provided by DialogGPT and this tutorial. I trained two models, one based small model and the other medium model. A finetuned large model is expected to have the best performance but my Google Colab account does not have enough GPU resource for training a large model.

Once the models finished trained. They were deployed to HuggingFace: novusshu/DialoGPT-medium-crypto and novusshu/DialoGPT-small-crypto. Both models are now public and be accessed from HuggingFace UI and programmablely.

2.4 Results and Comparison

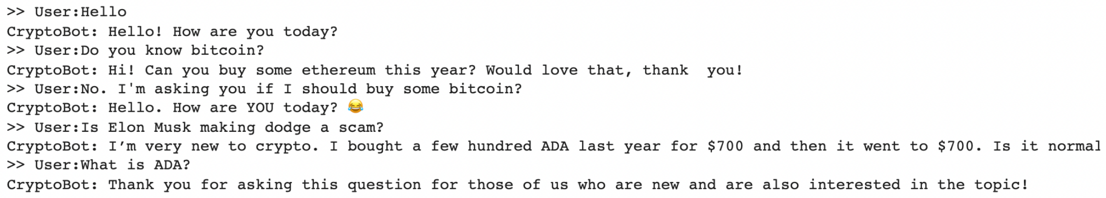
Figure 2 and Figure 3 shows respectively an example dialogue from the finetuned small model and from the finetuned medium model. Also for comparison, I generated a similar dialogue from the original DialoGPT large model as shown in Figure 4. We can see that a few features from my finetuned model at face value.

1. The bots (small and medium) are so focused on cryptocurrency that they don't do much daily talk;
2. The medium has clearly better "understanding" than the small model;
3. The original large is surprisingly good, at least for a relatively popular topic it had data before.



User:Hi
CryptoBot: Haha I ain't scared I'm keeping my crypto
User:What is ethereum?
CryptoBot: So is this like a paid vacation for them or what?
User:What are you talking about?
CryptoBot: And the cost of living is higher than the cost to produce.
User:Which is the most promising cryptocurrency?
CryptoBot: How do you figure?
User:You tell me
CryptoBot: !!!Remindme 18-09-2021

Figure 2: An example dialogue from the novusshu/DialoGPT-small-crypto model



```
>> User:Hello
CryptoBot: Hello! How are you today?
>> User:Do you know bitcoin?
CryptoBot: Hi! Can you buy some ethereum this year? Would love that, thank  you!
>> User:No. I'm asking you if I should buy some bitcoin?
CryptoBot: Hello. How are YOU today? 😂
>> User:Is Elon Musk making dodge a scam?
CryptoBot: I'm very new to crypto. I bought a few hundred ADA last year for $700 and then it went to $700. Is it normal
>> User:What is ADA?
CryptoBot: Thank you for asking this question for those of us who are new and are also interested in the topic!
```

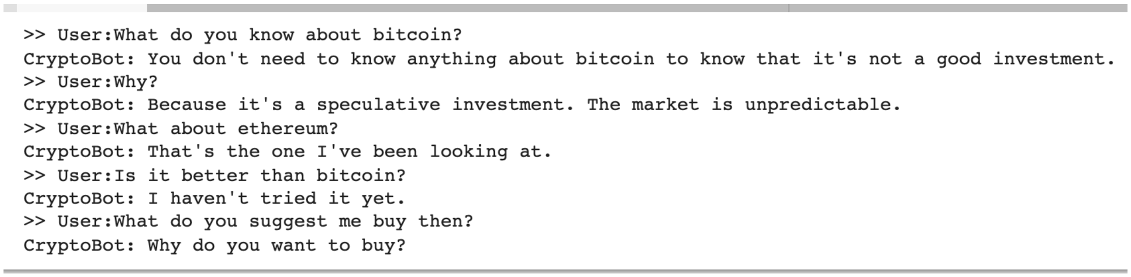
Figure 3: An example dialogue from the novusshu/DialoGPT-medium-crypto model

2.5 Limitations

The main limitation hosting on Hugging Face Server with a free tier account is the resource, which significantly affects the model response time and the limit for messages. And the model can occasionally become "mad" and response with the same sentence again and again, but this seems the innate problem with the DialoGPT model.

3 Front-end

The front-end is also public accessible. Anyone with a Telegram account can add the bot by its user name: sageC_bot and start talking immediately.



```
>> User:What do you know about bitcoin?
CryptoBot: You don't need to know anything about bitcoin to know that it's not a good investment.
>> User:Why?
CryptoBot: Because it's a speculative investment. The market is unpredictable.
>> User:What about ethereum?
CryptoBot: That's the one I've been looking at.
>> User:Is it better than bitcoin?
CryptoBot: I haven't tried it yet.
>> User:What do you suggest me buy then?
CryptoBot: Why do you want to buy?
```

Figure 4: An example dialogue from the microsoft/DialoGPT-large model

3.1 Telegram

I chose Telegram as the frontend bot application mostly because it is the de facto official IM for the cryptocurrency community. Telegram is an encrypted messenger service and is a popular tool for communication and marketing in the community. Most important, it provides a very comprehensive bot API for 3rd party developers.

The steps for developing a bot starts with a free Telegram account. And then we contact the BotFather for setting up a new bot with some basic info, and use the assigned API credentials for further development, which in this project, is a simple flask app.

3.2 Flask

For the telegram bot to use Hugging Face models, the most important thing is to get the credentials of Hugging Face, Telegram and Heruku right. I used the huggingface_hub library for integrating a Hugging Face model inference API object as well as the telegram library as the bot API wrapper for accessing various endpoints. The Hugging Face model returns a JSON format response and it's easy to parse for reply message.

3.3 Results

Users accessing from Telegram results can experience significant delay when starting a conversation, but as the conversation continues, the delay will no longer perceivable. This should be due to the treatment at the Hugging Face server. The conversation content is no different from what we tested directly from the backend, hence I'm not demonstrating here again.

Right now, the bot is only capable of responding to individual talks. Future development will add functionality for channel management and group messages moderation.

4 Future Plans

This project is quite an interesting journey for me to get hands-on experience with the GPT models. The most unsatisfactory thing is the limited computing power, i.e., GPU resources. Each epoch is taking too long and a finetuned DialoGPT large model is unavailable. However, there are still a lot of specificity I need to further attend to on the model side. It seems training with Reddit data is far from ideal, and I need to develop with new data and perhaps new model specification. Other thoughts for future features include integration of pricing info and expand to other chatbot frontends such as Discord.

References

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv:1911.00536 [cs]*.