Multilingual Cryptocurrency Chatbot

Group 26
https://github.com/joohan-lee/CS544-Multilingual-Cryptocurrency-Chatbot
Kumari Anuska Singh, Tetsuro Kai, Joohan Lee, Junlin Liu, Behzad Ebrahimi

Abstract

A chatbot is a robot that understands customer questions and provides helpful answers quickly. With advances in natural language processing, chatbots are becoming increasingly popular in various fields. We implemented a cryptocurrency chatbot application utilizing TF-IDF, Word2Vec, BERT, and SIF embedding models in this project. To assist in answering questions in multiple languages, a translating module was developed that recognizes the language type and translates the contents if needed. The results of measuring the cosine similarity showed high accuracy of the models implemented, especially for BERT and SIF. Additionally, the response can be generated in the language entered by the user.

1 Introduction

In recent years, there has been a tremendous increase in interest in cryptocurrencies, which are used as a medium for financial transactions or simply for speculative trading. Because of their speculative nature and ease of trading once an account is established, cryptocurrencies are becoming increasingly popular, particularly among individual investors. Cryptocurrency traders come from a wide range of backgrounds, with participants ranging from market professionals to novices. Chatbots could be useful in bridging knowledge gaps among market participants and providing timely and easy-to-understand information on rapidly changing cryptocurrencies. Our group decided to create a user-interactive chatbot to answer questions about cryptocurrencies. Our group also decided to implement a function to support several types of languages other than English since users of cryptocurrencies are not limited to English-speaking countries but also include a wide range of other countries such as Japan, Korea, and India.

2 Related work

2.1 Chatbot

Along with the development of machine learning and deep learning technologies, the performance of chatbots are increasing. There exists two types of ML-based chatbots. One is intention-detection type and the other is generative type. An intention-detection model basically exploits a classifier which predicts intents from a set of users' questions and returns the regarding answer in the answer set. Most intention-detection models use a QA dataset that consists of question-answer pairs. Sentence-BERT(SBERT), a modification of the pretrained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity, shows good performance on question-answer pair regression tasks. Since it does not require to feed both sentences of a pair into the network, it has much less computational overhead and shows faster performance than BERT / RoBERTa while maintaining the accuracy from BERT. On the other hand, generative chatbots typically function in the machine translation manner.

2.2 Cryptocurrency

The cryptocurrency market has gained tremendous traction in the past few years. There are a huge number of people who just started to invest in cryptocurrency. However, because of the anonymous nature of cryptocurrencies and the absence of a centralized regulating authority, cryptocurrencies are subject to price manipulation. Therefore, there were a variety of research such as Cryptocurrency Portfolio Management with deep learning,

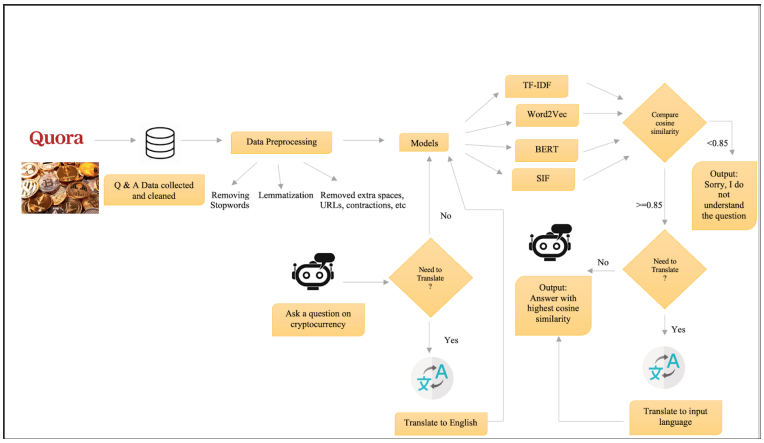

Figure 1: Chatbot System Architecture

Detecting pump and dumps, and chatbot applications for Cryptocurrency. Our research is mainly focused on the chatbot to answer general questions about cryptocurrency so that new investors are able to more readily acquire information.

3 Method

3.1 System Design

We built a command line chatbot. The user can type their question into it, and the system will then output the answer. Our chatbot system is split into two parts: Translate Model and Chatbot Model.
Translate Model: The main steps of Translate model will be designed by following steps:

- Detect the language of the user's question.
- If the input text is English, feed original text into the network without translation.
- If not, translate texts into the detected language using googletrans API.

Chatbot Model : The main step of Chatbot model will be designed by following steps:

- Train the chatbot model by using the question dataset we collect from Quora.
- Receive the text data from translate model.
- Input the text data into trained chatbot model and find the question with highest cosine similarity.
- If the highest cosine similarity is higher than 0.85, the model will output the answer.
- If the highest cosine similarity is lower than 0.85, the model will output i don't understand your question.
- If users ask the price of cryptocurrency, we return the real-time price of it. We support 50 cryptocurrencies.

Figure 1. is an example of chatbot architecture. When a user input a cyrptocurrency question is present, the translate model will directly output the input text. The chatbot model will then receive the text and respond. If the language of the input text is not English, the translate model will translate it to English and output it. The chatbot model will then receive the text and respond.

3.2 DATA

Quora is the main data sources for our model. There are a lot of questions and answers about cryptocurrency for the Quora. We collect more than 22,000 different cryptocurrency topic questions and select the most voted answer of the questions.

First, we search for the most popular cryptocurrency questions from google, and we save these questions as a list. Then we try to find the questions from Quora and crawl the answers from Quora. Moreover, we will create a keyword list like cryptocurrency, Bitcoin, and so on, and then we will crawl all questions that contain those keywords and their answers by using Requests, Beautiful Soup, and Selenium.

For each question, we selected the most voted answer. In addition to cryptocurrencies with high trading volume, crypto trading platform related terms such as FTX were also included. Then

we removed unnecessary words such as HTMLs, URLs, and extra spaces. We also changed all of the sentences to lower case so that our language model could better understand the same words. Following various experiments, we will perform stop word removal, lemmatization with NLTK, and contraction expansion with the 'contractions' library.

3.3 MODEL

For our chatbot, we used four models, which are TF-IDF, Word2Vec, BERT and SIF. We used the model to calculate the cosine similarity between user questions and all questions in the question dataset and select the questions with the highest similarity from the question dataset and output answer to the user. Moreover, we also set a threshold which is 0.85 for each model. If the highest cosine similarity is smaller than 0.85, the model will output 'there is no similar question'.

3.3.1 TF-IDF

TF-IDF works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. In our model, we used more than 16k cleaned cryptocurrency question, and we then generate the TF-IDF vector for this dataset by using packages such as TfidfTransformer. After we get the TF-IDF vector for all questions, we count the user question's cosine similarity with the other questions and output the question with the highest cosine similarity.

3.3.2 Word2Vec

Word2vec groups the vector of similar words together in the vector space. Given enough data, usage and contexts, word2vec can make highly accurate guesses about a word's meaning based on past appearances. In our model, we used more 16k cleaned cryptocurrency question to train Word2Vec model. Then we used wmdistance function to count the similarity of two questions. In order to improve the model, we are tuning the model like changing the size of window. After many attempts, we found that increase the size and window will increase the cosine similarity.

3.3.3 BERT

For the BERT model, it will encode the text to be a vector containing 768. Those 768 values contain our numerical representation of a single token — which we can use as contextual word embeddings. Then we transform them to create semantic representations of the input sequence. We did not do the data cleaning for the BERT model because it might worsen the prediction of the model. In our model, we firstly used the pre-trained BERT model 'bert-base-nli-mean-tokens', and we used all 16k cryptocurrency questions dataset to train the BERT model. Then we encode the user question and output the question with the highest cosine similarity. In order to get a better model, we used another pre-trained BERT model 'sentence-transformers/distilbert-base-nli-max-tokens'.

3.3.4 SIF

SIF which is Smooth Inverse Frequency, is a weighting scheme for improving sentence embedding performance. When encoding a sentence, it is critical to determine which words are more important. We begin by training a SIF embedding model on the "glove-wiki-gigaword-100" dataset. The SIF model is then trained using all of the questions from our own dataset. Once the model has been trained, we find the SIF embedding for the input question. Following that, we find and compare the cosine similarity of all questions to the input question. The question with the highest similarity is chosen, and the corresponding answer is output. In this way, the SIF embedding assists us in selecting the most similar question from the dataset.The chatbot correctly predicted the answer. When compared to Word2Vec and TF-IDF, cosine similarity and accuracy improved.

4 Experimental setup

4.1 Datasets

We scraped and collected over 22K cryptocurrency-related questions and answers on Quora. After data cleaning, we still have more 16k cryptocurrency-related questions. The table 1 below is the example of our dataset. To test the accuracy of all the models, we created our own

Table 1: Example of Question and Answer data

Questions	Answers
Who needs cryptocurrency exchange	Everybody who wants to buy or sell cryptocurrencies. I.e. crypto investors or users.
What should a cryptocurrency technology that is non blockchain be called	yes it can Cryptocurrency is just an incentive for people to keep the blockchain correct.
Is Dogecoin more volatile than Bitcoin.	Never ever, Bitcoin is the largest volatile crypto than all cryptos.

test data, which was approximately 2K from the approximately 16K cleaned dataset.

4.2 Google Colaboratory

We used Google Colaboratory to run our code. We chose Google Colab because of its numerous advantages. It allowed us to share files with all team members, make it easy to version on github, and because Google Colab has a great collection of snippets, we could just plug in our code.

4.3 Evaluation metrics

We primarily used cosine similarity for the evaluation metrics. We used cosine similarity for each model to find the most similar question in our dataset to the user input question. We were able to answer the questions more efficiently this way. The cosine similarity is advantageous because even if two similar documents are far apart by the Euclidean distance (due to the size of the document), they may still be oriented closer together. We also calculated the accuracies of top performing models by creating our own test data out of the scrapped dataset.

5 Results and discussion

Implementing a chatbot and using it in real-life is like a double-edged sword. It can be favorable and helpful for users if well designed and give them enough satisfaction with their dialogue. At the same time, if any of the implementation steps, from data collection to cleaning, processing, model design, training, and testing the outputs are not done properly, chatbot reactions to the user questions might be irrelevant and usually people don't give a second chance to the chatbots. Considering the final purpose of our chatbot as a cryptocurrency information source or even an investing consultant, it was really important to maintain a smooth, relevant, and credible dialogue between the chatbot and the user. As mentioned before, after a considerable amount of research and evaluating several different options, we decided to work on four different models including TF-IDF, Word2Vec, BERT and SIF. Given the type of our chatbot as an intention-detection model, gathering enough credible question answer pairs and a powerful evaluation metric seemed necessary. We used cosine similarity for each model to find the most similar question to the input(test) question. Table 2 contains three examples of data/input pairs and their cosine similarity for all models:

Dataset Questions	Input Questions	TF-IDF	Word2Vec	SIF	BERT
how do people send crypto from a trust wallet to binance	how to send crypto to binance from trust wallet	0.609	0.651	0.875	0.926
how the price of bitcoin is affecting bitcoin cash and etherium	how etherium and bitcoin cash prices are impacted by the price of bitcoin	0.588	0.600	0.885	0.963
how do i create an account on binance	how to create account in binance	1.0	0.622	0.986	0.959

Table 2: Cosine similarity in all four models

From the table above, the cosine similarity of TF-IDF and Word2Vec are so low compared with the other two models, so we speculate that the accuracy of these two models will also be very low. In this case, we only test and compare the accuracy of BERT model and SIF model.

Applying the same approach on a set of 2000 data/input question pairs, BERT method represented the highest accuracy and SIF was in second position. Figure 2 shows a comparison between the accuracy of BERT and SIF models, having 83% and 79% accuracy respectively.

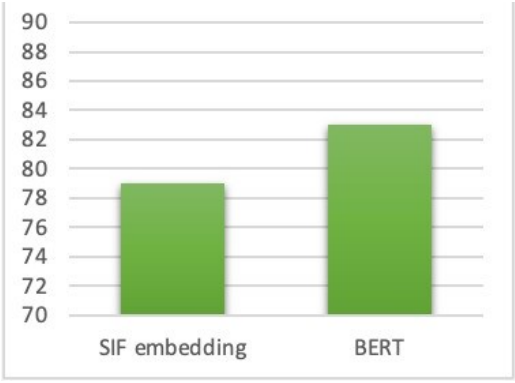

Figure 2: BERT vs SIF accuracy(%)

As mentioned in previous sections, we also considered non-English speaking users and included a translator in our model, so the input questions can be in five different languages. Although the result was relatively satisfying for us, we are not giving up our chatbot at this stage. Adding more data and improving the chatbot knowledge with an interesting user interface is our next step. We are also planning to implement a generative version of the Cryptobot and do more research on the basis of the cryptocurrency in order to provide more credible information, detect pump-and-dumps and predict the crypto price based on scientific and probabilistic methods.

6 Division of labor

- Junlin Liu - He worked on TF-IDF, Word2Vec, BERT model and Final project.
- Anushka Singh - She worked in the creating test data, SIF model, making poster and report.
- Tetsuro Kai - He mainly worked on creating a dataset by developing a scraping tool on Quora and collected 220K data.
- Joohan Lee - He worked on SBERT model and implemented a simple chatbot with translation and real-time price functions.
- Behzad Ebrahimi - He worked on Data cleaning and data processing

References

Sreelakshmi, A.S., Abhinaya, S.B., Aishwarya, N. and Jaya Nirmala, S., 2021. A Question Answering and Quiz GenerationChatbot for Education.

Nghiem, H., Muric, G., Morstatter, F. and Ferrara, E., 2021. Detecting cryptocurrency pump-and-dump frauds using market and social signals. *Expert Systems with Applications, 182,* p.115284.

Mirtaheri, M., Abu-El-Haija, S., Morstatter, F., Ver Steeg, G. and Galstyan, A., 2021. Identifying and analyzing cryptocurrency manipulations in social media. *IEEE Transactions on Computational Social Systems, 8(3),* pp.607-617.

Weizenbaum, Joseph., ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9.1 (1966): 36-45.*

Jacob Devlin, Ming-Wei Chang, et al. BERT: Pre-training of DeepBidirectional Transformers for Language Understanding. *ArXiv e-prints: 1810.04805v1, 2018*

Xie, Q., Zhang, Q., Tan, D., Zhu, T., Xiao, S., Li, B., Sun, L., Yi, P., Wang, J., 2019. Chatbot Application on Cryptocurrency.

Nils Reimers, Iryna Gurevych 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.