SI630 Project Report

Chongdan Pan
pandapcd@umich.edu

Abstract

This report is for a SI630 (Natural Language Processing) Project. It covers the main idea, technique as well as models used in predicting the cryptocurrencies market behavior based on text gathered from Twitter. Then experiments are designed to evaluate the performance of the models as well as help generate insights about NLP's role in the finance area.

1 Project Goals

For now, Natural Language Processing has become the hottest topic in the technology area, because it can be used as a great tool to understand human behavior. For example, NLP can help us understand the emotion of individuals and generate some insights into collective behavior.

The typical collective behavior of humans is trading in a market. People have used different methods such as graphs and quantitative models to understand the market's behavior, and I believe NLP can definitely be a useful tool. A lot of financial institutions have set up teams trying to dig some valuable information from the market through NLP. For traders and investors, NLP may be used for them to find out the hot spots in the market and make a profit. For governors, NLP can be used to understand the emotion of the market and evaluate the risk.

People's passion have already shown NLP's potential in the finance area. Therefore, this project will do some exploratory analysis based on NLP technology, and try to dig some valuable information out of the market.

2 NLP Task Definition

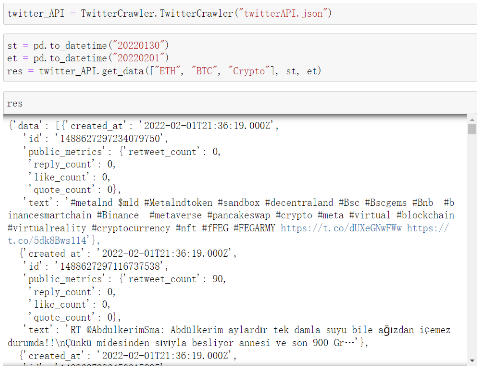
There is a lot of valuable information in the market, and they're represented in various indicators, such as volatility, price movement, and trading volume. Therefore, the output can either be numerical values of these indicators or indicators variables showing if the price will go up or down tomorrow.

The input of the task can be any text related to the market, such as comments, articles. It's worth mentioning that all these texts should appear before the market's behavior, otherwise they're just the feedback of the market behaviors.

3 Data

It's obvious that any information can be related to the market, and there are a huge amount of assets to be traded in the market. Therefore, it'll be unrealistic to cover all this information and markets in our model. This project will mainly cover the market behavior of crypto assets since all data is open to the public.

The input will mainly be retrieved from Twitter since it's always a hot place for people to share their opinions. I'll call Twitter V2 API to gather all tweets related to specific tags in a certain time period.



```
twitter_API = TwitterCrawler.TwitterCrawler("twitterAPI.json")
```

```
st = pd.to_datetime("20220130")
et = pd.to_datetime("20220201")
res = twitter_API.get_data(["ETH", "BTC", "Crypto"], st, et)
```

```
res
```

```
{'data': [{'created_at': '2022-02-01T21:36:19.000Z',
   'id': '1488627297234079750',
   'public_metrics': {'retweet_count': 0,
    'reply_count': 0,
    'like_count': 0,
    'quote_count': 0},
   'text': '#metalnd $mld #Metalndtoken #sandbox #decentraland #Bsc #Bscgems #Bnb  #b
inancesmartchain #Binance  #metaverse #pancakeswap #crypto #meta #virtual #blockchain
#virtualreality #cryptocurrency #nft #fFEG #FEGARMY https://t.co/dUXeGNwFWw https://
t.co/5dk8Bws114'},
  {'created_at': '2022-02-01T21:36:19.000Z',
   'id': '1488627297116737538',
   'public_metrics': {'retweet_count': 90,
    'reply_count': 0,
    'like_count': 0,
    'quote_count': 0},
   'text': 'RT @AbdulkerimSma: Abdülkerim aylardır tek damla suyu bile ağızdan içemez
durumda!!\nÇünkü midesinden sıvıyla besliyor annesi ve son 900 Gr…'},
  {'created_at': '2022-02-01T21:36:19.000Z',
```

Figure 1: Result of Twitters V2 API

Typically I'll use the create time of each tweet and its text, but more fields may be used for the model.

On the other hand, the market from Yahoo finance will be set as the label of the model. It will have fields like close price and trade volume. I can even generate more technique indicators based on these fields. The frequency of the market data is per day, which means I'll set all tweets in the previous day as a label. If the model will be used to predict a higher frequency, Binance API can be used to gather data every minute.

4 Related Work

1. (Shahzad et al., 2021). This paper used Linear Regression, Deep Neuro Network as well as LSTM to predict the price of bitcoin-based on tweets. It uses sentiment analysis to see whether the tweets have a positive or negative effect on bitcoin's price. However, it doesn't get me any clear result or conclusion about each model's performance. On the other hand, it only considers the plain text of each tweet, while I think we also need to consider who posts the tweets. Therefore, I may work more on the preprocessing and evaluate the performance of various models.

2. (Laskowski and Kim, 2016). This paper was published in 2016, where bitcoin and blockchain are not quite popular right now. It focuses on getting data from social media and looking for the correlation between bitcoin price and specific hashtags. It turns out bitcoin-price talk has a much higher correlation to bitcoin than other channels. This result can give me some insight into how to choose keywords for getting the data. On the other hand, it also gives me a rough idea about the size of the data as well as the time and memory required for processing. In this paper, the data size is 200GB and uses 1.95 GB RAM on average. However, I think it's a reasonable size for my project since I won't get data from so many channels and computers are more advanced now.

3. (Huang et al., 2021). This paper focuses on applying LSTM on sentiment analysis of people's posts on Weibo, one of the biggest social media in China. Similar to other papers, LSTM is used to predict if the price will go up or go down. The most interesting part of this paper is that they construct a vocabulary set with unique characteristics related to the crypto market. However, they manually construct the vocabulary set and identify the keywords related to the crypto market, which looks quite inefficient to me. Therefore, I think tf-idf probably can be useful for us to identify the key pattern.

4. (Wong, 2021). This paper used Naive Bayes and LSTM model to do sentiment analysis. It turns out Naive Bayes does a better job than LSTM in distinguishing similar tweets' relation to the crypto prices. However, there is no general threshold for these two models to do a good classification on whether the tweet is negative or positive, which means these models are better at doing a relative comparison between two tweets. As for classification, LSTM has 51% accuracy while Naive Bayes only has 50%, which means they're not working well. The result looks much more like a random guess, but it can serve as a baseline for the project.

5 Evaluation

The evaluation will be quite intuitive for now. The model can output the classification result such as whether the price will go up or go down tomorrow. Since we'll retrieve the price data, we can shift it in time and use an indicator variable to represent the price movement. In this way, a softmax function can be used to evaluate the loss. What's more, it can even be used for multinominal classification by setting some threshold in the relative change of the price.

As for the accuracy, it's even inconsistent within the related papers. The accuracy is very different, ranging from 50% to 80%, I think it may depend on the data set as well as the tokenism.

6 Discussion

This is project is going to integrate the ideas from the related papers and try to build a robust model with high prediction accuracy. What's more, since the crypto market has extremely high volatility, predicting a high-frequency with multinominal results may be more meaningful and applicable in the real scenario.

7 Work Plan

The project mainly include three parts.

1. Getting the data. It will take less than two weeks for me to get familiar with twitter's API and get historical tweets.

2. Language Processing. This is the core part of the project. As the course progress, I'll apply the different method to process the data and turn it into some matrix representation that can be directly sent to the neuron network model. It'll take the rest of weeks of the semester.

3. Model Design and Training. Training should start from the moment I've got the data I need because I can use a simple logistic regression. However, to make the model better, I may need to do some modifications to the structure of the neuro network or change some optimizers and training hyperparameters.

References

Xin Huang, Wenbin Zhang, Xuejiao Tang, Mingli Zhang, Jayachander Surbiryala, Vasileios Iosifidis, Zhen Liu, and Ji Zhang. 2021. Lstm based sentiment analysis for cryptocurrency prediction.

Marek Laskowski and Henry M. Kim. 2016. Rapid prototyping of a text mining application for cryptocurrency market intelligence. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 448–453.

Muhammad K. Shahzad, Laiba Bukhari, Tayyeba Muhammad Khan, S. M. Riazul Islam, Mahmud Hossain, and Kyung-Sup Kwak. 2021. Bpte: Bitcoin price prediction and trend examination using twitter sentiment analysis. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 119–122.

Eugene Lu Xian Wong. 2021. Prediction of bitcoin prices using twitter data and natural language processing.