SI630 Project Update Report

Chongdan Pan
pandapcd@umich.edu

Xinyi Ye
sylva@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 Introduction

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.

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.

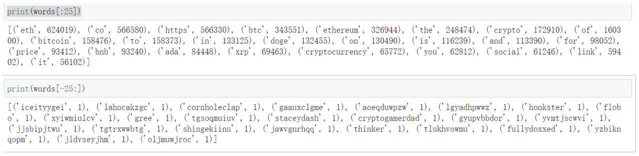
2 Data

We have data from two Twitter API and Binance market data api. The time range of the data is from 2021/01/01 to 2022/02/25.

2.1 Twitter Text

We're using the text data from Twitter to train our model. Since there are too much text on twitter everyday, and we our computer can process all of them, we use a filter when querying the data from Twitter API. Since Twitter have a limitation on how much data you can get every month, our query must be very specific. We only query for data with a hash tag ETH or Ethereum, and there is number and dollar symbol in the tweets. Besides the text, we also get the authorid and timestamp of each tweet so that we can match it with the market data.

Besides the text information, tweets may have other entities such as emoji or entities, therefore, we're using regular expression to extract the information that we're interested in. After counting all the words in the corpus, we find the most frequent words are coin's name or some stopwords, which is meaningless. On the other hand, we also find that the lest frequent words are some meaningless components of url. Therefore, we're going to set some threshold on the frequency of words and drop out those don't have much meaning.



```
print(words[:25])
```

```
[('eth', 624019), ('co', 566580), ('https', 566330), ('btc', 343551), ('ethereum', 326944), ('the', 248474), ('crypto', 172910), ('of', 1603
00), ('bitcoin', 158476), ('to', 158373), ('in', 133125), ('doge', 132455), ('on', 130490), ('is', 116239), ('and', 113390), ('for', 98052),
('price', 93412), ('bnb', 93240), ('ada', 84448), ('xrp', 69463), ('cryptocurrency', 65772), ('you', 62812), ('social', 61246), ('link', 594
02), ('it', 56102)]
```

```
print(words[-25:])
```

```
[('iceityygei', 1), ('lahocakzgc', 1), ('cornholeclap', 1), ('gaauxclgme', 1), ('aoeqduwpzw', 1), ('lgyadhpwwz', 1), ('honkster', 1), ('flob
o', 1), ('xyiwmiulcv', 1), ('gree', 1), ('tgsoqmuiuv', 1), ('staceydash', 1), ('cryptogamerdad', 1), ('gyupvbbdor', 1), ('yvmtjscwvi', 1),
('jjsbipjtwu', 1), ('tgtrxwwbtg', 1), ('shingekiinu', 1), ('jawvgnrhqq', 1), ('thinker', 1), ('tlukhvowmu', 1), ('fullydoxxed', 1), ('yzbikn
qopm', 1), ('jldvseyjhm', 1), ('oljmuwjroc', 1)]
```

Figure 1: Most frequent and lest frequent words

2.2 Binance Market Data

For market data for ETH, we're using Binance's API and fetch candlesticks of every 1 hour. For simplicity, we're only considering the close price of each interval, and we also keep other fields such as trade volume, buy volume as well as timestamp.

To make profit, we focus more on the return and other property of an asset rather than the price, therefore, we calculate the return manually by $R_t = \frac{P_{t+1}}{P_t} - 1$, and our goal is trying to find some casuality between the text and return.

We've plot the histgram of the return, and it has a range from -0.13 to 0.07 with a mean to be -0.001. Due to the heavy tail of the return, it follows a t-distribution

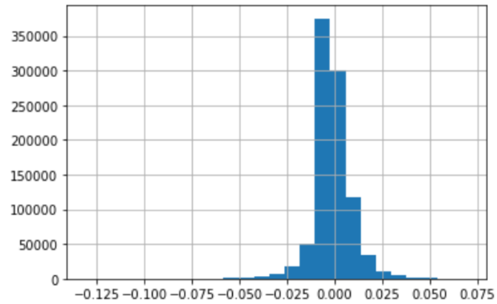
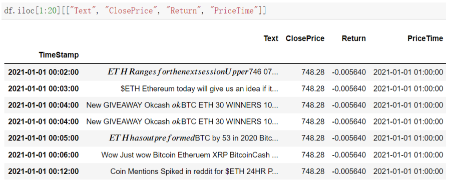


Figure 2: Distribution of return

2.3 Sample Data

For now, our goal is to use all the text related to ETH price posted on Twitter to predict the return of next hour. Therefore, we're merging the data by timestamp so that the text always happen before the return.



```
df.iloc[1:20][["Text", "ClosePrice", "Return", "PriceTime"]]
```

	Text	ClosePrice	Return	PriceTime
TimeStamp				
2021-01-01 00:02:00	$ETHRangesforthenextsessionUpper$746 07...	748.28	-0.005640	2021-01-01 01:00:00
2021-01-01 00:03:00	$ETH Ethereum today will give us an idea if it...	748.28	-0.005640	2021-01-01 01:00:00
2021-01-01 00:04:00	New GIVEAWAY Okcash okBTC ETH 30 WINNERS 10...	748.28	-0.005640	2021-01-01 01:00:00
2021-01-01 00:04:00	New GIVEAWAY Okcash okBTC ETH 30 WINNERS 10...	748.28	-0.005640	2021-01-01 01:00:00
2021-01-01 00:05:00	$ETHhasoutpreformed$BTC by 53 in 2020 Bitc...	748.28	-0.005640	2021-01-01 01:00:00
2021-01-01 00:06:00	Wow Just wow Bitcoin Etheruem XRP BitcoinCash ...	748.28	-0.005640	2021-01-01 01:00:00
2021-01-01 00:12:00	Coin Mentions Spiked in reddit for $ETH 24HR P...	748.28	-0.005640	2021-01-01 01:00:00

Figure 3: Sample Data

As shown in the figure, the data from training can be text posted from the first hour in 2021/01/01, but we're going to predict the return in the second second hour.

3 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. (Patel et al., 2020) This paper used a LSTM and GRU-based hybrid cryptocurrency prediction scheme to predict two cryptocurrencies, namely Litecoin and Monero. LSTM has been proved to be the best in until now due to their ability to remember and extract the temporal features of data. The results depict that the proposed scheme accurately predicts the prices with high accuracy, revealing that the scheme can be applicable in various cryptocurrencies price predictions.

4 Methodology

1. Embedding the words with word vectors
   Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc. Word2Vec is a method to construct such an embedding including some dependence of one word on the other words. It can be obtained using two methods: Skip Gram and Common Bag Of Words (CBOW). We will try both methods.
   CBOW Model: This method takes the context of each word as the input and tries to predict the word corresponding to the context. The input or the context word is a one hot encoded vector of size V. The hidden layer contains N neurons and the output is again a V length vector with the elements being the softmax values. Another variant is Skip Gram model. We input the target word into the network. For each context position, we get C probability distributions of V probabilities, one for each word. The model outputs C probability distributions.

2. Use sequence model such RNN and LSTM
   A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data. Sequential data is basically just ordered data in which related things follow each other. In a RNN the information cycles through a loop. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously. That is to say, Every prediction at time t (h_t) is dependent on all previous predictions and the information learned from them.
   Long Short Term Memory Network is an advanced RNN, a sequential network, that allows information to persist. It is capable of handling the vanishing gradient problem faced by RNN. A recurrent neural network is also known as RNN is used for persistent memory. The LSTM consists of three parts. The first part chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. In the second part, the cell tries to learn new information from the input to this cell. At last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp.
   Just like a simple RNN, an LSTM also has a hidden state where H(t-1) represents the hidden state of the previous timestamp and Ht is the hidden state of the current timestamp. In addition to that LSTM also have a cell state represented by C(t-1) and C(t) for previous and current timestamp respectively. The cell state carries the information along with all the timestamps.

3. Change different representation of features
   We also plan to take different features into consideration. As we all know, there are different types of cryptocurrencies. Bitcoin is the most popular and valuable cryptocurrency. Among others, some are clones or forks of Bitcoin, while others are new currencies that were built from scratch. They include Solana, Litecoin, Ethereum, Cardano, and EOS. They have close relationship with each other. When predicting the returns for one of them, we will try to consider others' influence on it.
   Besides, we will also take the influence of spokesman of the news into consideration. The information for some spokesman, like a government official or authority figure, is clearly more essential than others. We will train the weights for different spokesman along with other parameters, and add it to the final prediction.

5 Evaluation and Results

There are two steps for our work. First, we first tokenized text data from Twitter. We lowered the data. We used regular expression to filter possible features, such as removing the punctuations. We also removed the stopwords. Then we built a term-document matrix for every hour period. Every minutes we have several texts. We collected texts for every hour and built a term-document matrix for each period as the feature vector. Second, we used linear regression to predict the returns and we used logistic regression to predict the trend and returns as our two baseline models. For the first baseline model, we used the formula

$$return = \frac{final\_value - initial\_value}{initial\_value}$$

to calculate the returns for each one-hour period. Then we used linear regression to predict the returns for each one-hour period. Finally we simulated the investors' investment using the following method. For each predicted return value, we first put it into the sigmoid function and get a possibility value. Then we used this possibility as the p value of the binomial distribution and get a 0/1 value, whose meaning is whether we should buy the cryptocurrencies for this hour. We used this investment strategy to do the investment for each hour and calculated the revenue we can get using the strategy.

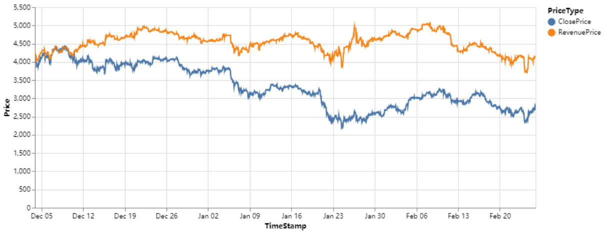


Figure 4: Revenue Price ans Closed Price

From the above picture we can conclude that using this strategy, we can get the positive earnings. The second baseline model is using logistic regression to predict the trend of cryptocurrencies. We divided the returns into ten quantiles, and gave a label for each quantile. After getting the predicted label, we used the mean value of that section as the returns and used the returns to calculate its corresponding priced.

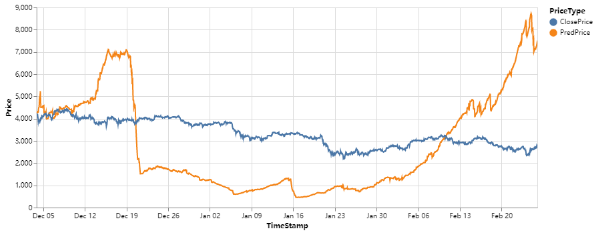


Figure 5: Revenue Price ans Closed Price

From the above picture we can conclude that the predicted trend we get from the second baseline is not so good.

6 Work Plan

The project mainly include three parts.

1. Getting the data. It will take less than two weeks for us 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, We'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 we've got the data we need because we can use a simple logistic regression. However, to make the model better, we 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.

Mohil Maheshkumar Patel, Sudeep Tanwar, Rajesh Gupta, and Neeraj Kumar. 2020. A deep learning-based cryptocurrency price prediction scheme for financial institutions. *Journal of information security and applications*, 55:102583.

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.