Cryptocurrency Predict Real Time

Le Phung Hoang Son[1], Hoang Anh Tuan[1], Nguyen Minh Sang[1], and Doan Đuc Phuong[1]

VNUHCM-University of Information Technology, HCMC, Vietnam

Abstract. Cryptocurrency is one of the most concerned issues today, having been widely used as an exchangemedium in areas such as financial transaction and asset transferverification. However, there has been a lack of solutions that cansupport real-time price prediction to cope with high currencyvolatility. To tackle this problem, we use Kafka to get streaming data from Crypto Compare API and send it to long short-term memory (LSTM) model to predict price in real-time.

Keywords: Cryptocurrency · LSTM · Kafka · Real time · API.

1 Introduction

In the past decade, cryptocurrencies as digital coins have become well known because of their extraordinary return potential in phases of extreme price growth and their unpredictable massive crashes. As a result, it engages the machine learning and data mining community to be able to predict price changes to check for volatility and better assess the danger associated with cryptocurrencies.

Because cryptocurrencies have always been highly volatile and do not behave like traditional currencies, it is therefore difficult to determine what is causing this volatility. This makes it challenging to accurately predict the future price of any cryptocurrency.

In this paper, we provide a real-time cryptocurrency price prediction system based on data from CryptoCompare. Firstly, we used a Kafka for data transmission. Secondly, we used several different solutions to handle the large volume of incoming data in a persistent and fault tolerant way. Finally, we used deep learning to leverage machine learning technology to predict the real-time price of Bitcoin.

2 Related Work

Since we deal with the issue of cryptocurrency, we have to learn the relevant knowledge about the cryptocurrency market. Next is the knowledge related to handling streaming data and solving the aforementioned problems through existing articles and research.

2.1 Kafka

Kafka is a distributed pub/sub messaging system. The side that pulls in the data is called the producer, and the subscriber that receives the data by topic is called the consumer. Kafka is capable of transmitting a large number of messages in real time, in case the recipient has not received the message, it is still stored in a backup queue and also on a secure disk. At the same time, it is also replicated in the cluster to prevent data loss.

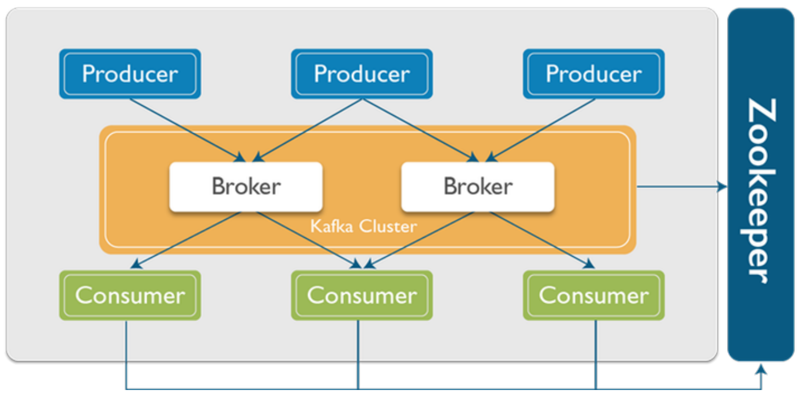

Fig. 1. A simple Kafka's structure.

2.2 LSTM Networks

Long-short network (Long Short Term Memory network), commonly known as LSTM(**?**) - is a special form of RNN that is capable of learning properties. LSTM was introduced by Hochreiter Schmidhuber (1997), and has since been refined and popularized by many people in the industry. We work extremely effectively on many different problems, gradually becoming as popular as it is today. LSTM is designed to avoid dependency problem (long term dependency). The information in the time long is the default feature of them, but we don't need to train it to be able to remember it. That is, its content can be memorized without anything being able to learn. Every regression network takes the form of a sequence of repeating modules of a neural network. With standard RNNs, these modules have a very simple structure, usually a fishy layer. The LSTM also has the same chained architecture, but its modules have a different structure from that of a standard RNN. Instead of just one layer of neural networks, they have up to four layers that interact with each other in a very special way.

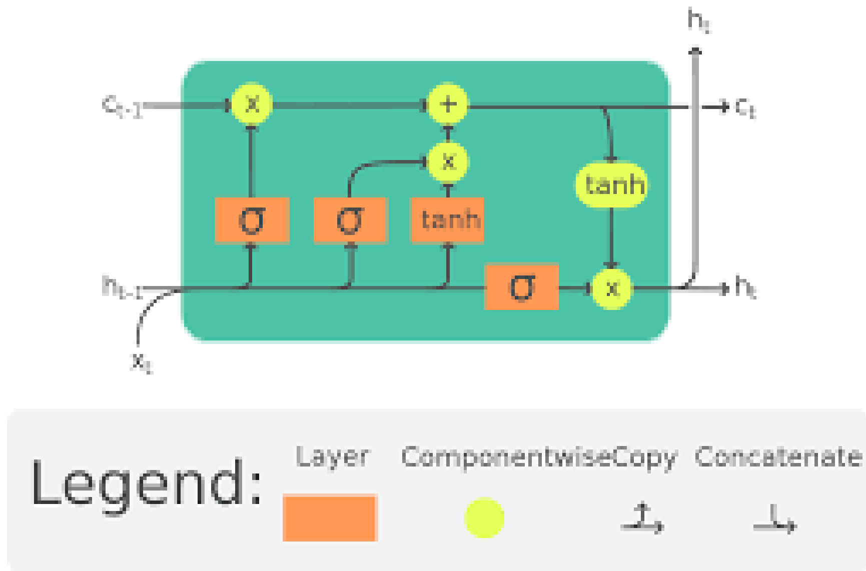

Fig. 2. LSTM Architecture

3 Method

3.1 Architecture

The figure 3 below describes the system's architecture.

Our system has 3 main modules: Producer, Consumer and Model. At the first module, we pulled data from Crypto Compare Web API which will be continously pushed in every 60 seconds. Then, we used KafkaProducer to send each row of data to the next module named Consumer. The consumer module has the role of receiving the data sent from the producer module then processing the data and finally pushing it to the LSTM model to run the prediction results. The process of taking data and transferring it to the LSTM model so that the model predicts the results goes on continuously

3.2 Deep Learning

A key element of our architecture is a deep learning model, trained to predict cryptocurrency prices. In problem prediction, we decide to choose an LSTM model, which tends to outperform other algorithms or frameworks. The LSTM is called into the consumer and receives clean data to predict the cryptocurrency price, the data being fed continuously forces the model to make continuous predictions.

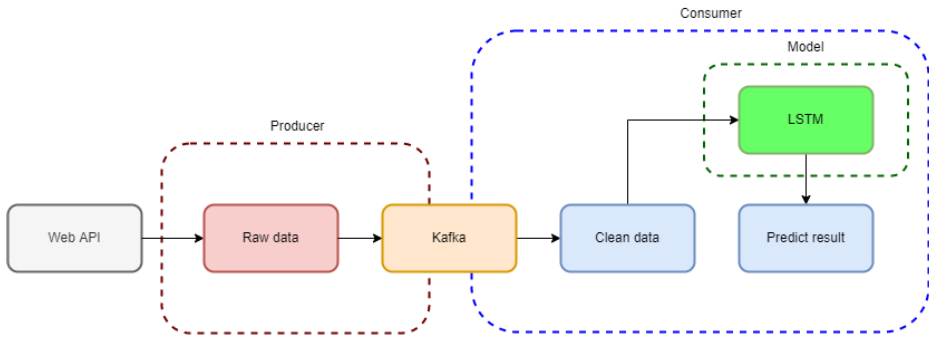

Fig. 3. Cryptocurrency Predict System Architecture

3.3 Metrics

MSE In this project, we use mean square error (MSE) as loss function. The MSE either assesses the quality of a predictor (i.e., a function mapping arbitrary inputs to a sample of values of some random variable), or of an estimator (i.e., a mathematical function mapping a sample of data to an estimate of a parameter of the population from which the data is sampled). The definition of an MSE differs according to whether one is describing a predictor or an estimator.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{1}$$

MAE In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}. \tag{2}$$

R2 In statistics, the coefficient of determination, denoted R2 or r2 and pronounced "R squared", is the proportion of the variation in the dependent variable that is predictable from the independent variable(s). It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. We do this calculation in three steps to make it easier to understand.

g is the sum of the differences between the observed values and the predicted ones. (ytest[i] – preds[i]) **2. y is each observed value y[i] minus the average of observed values np.mean(ytest).

$$R2 = 1 - (g/y) \tag{3}$$

4 Datasets

The dataset was built entirely handicraft, cryptocurrency data is taken directly from Crypto Compare's API, this data source is completely reliable and easy to manipulate. Raw data is processed in multiple steps to be suitable for the purpose of building machine learning models.

4.1 Data Processing

The raw data consists of 10,086 lines, including the entire information history of Bitcoin such as time, close price, open price, close price, high price, low price, volumn from, volumn to, conversion type and conversion symbol. We excluded the values of the conversion type and the conversion symbol because it did not significantly affect the prediction results to remove the blank values. Next, we deal with the time, which is a string of seconds indicating the date and time. For visualization, we've converted it to a string displaying the date and time. Finally, we convert the data to a matrix (x, y, z).

$$x = total\ row\ of\ data - y \tag{4}$$

In there, x is the total number of lines remaining in data, y is the number of rows of data that need to be fed into the model each time and z is the number of input features.

5 Result

The loss for the LSTM model is minimal at the learning rate of 0.01. It is not the best fit because it is almost impossible to meet both the train and the test at a time because the time series data fluctuates enormously. LSTM's loss graphs are given below. [4]
LSTM multi-feature predicting graphs are given below. [5]

6 Conclusion and future work

This study focuses on the Bitcoin closing price for the development of the predictive model. The prediction is limited to previous data. The ability to predict data

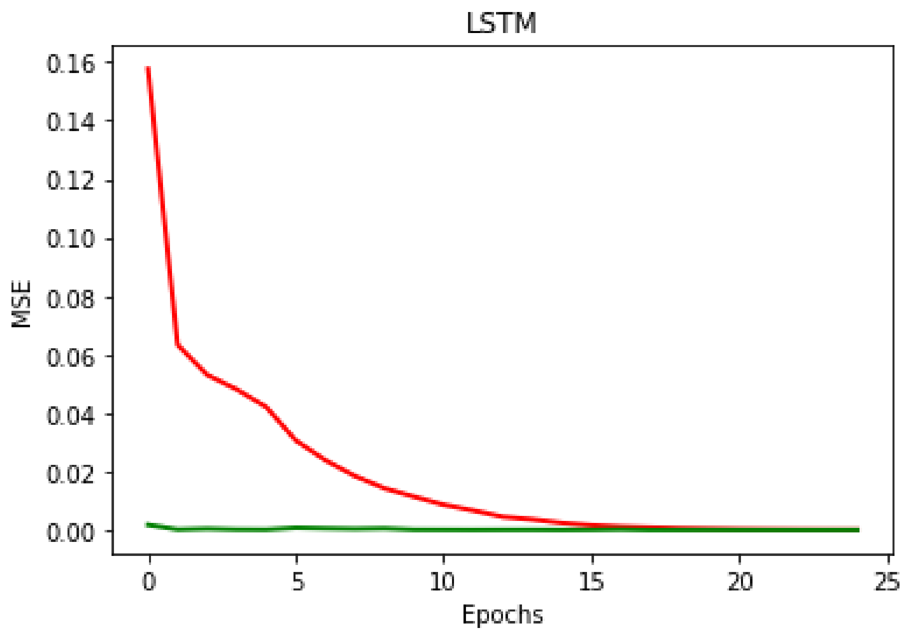

Fig. 4. Loss value of train dataset and validation dataset

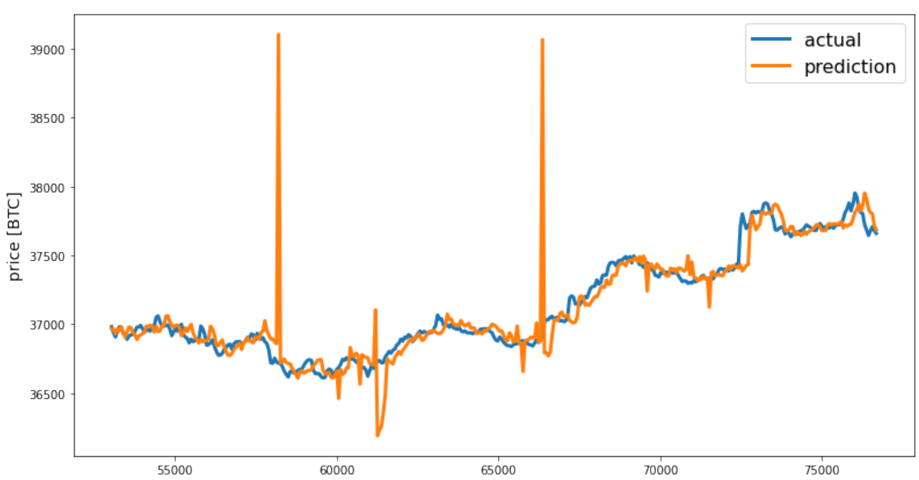

Fig. 5. LSTM multi-feature predicting graphs

streaming would improve the model's performance and predictability. The model developed using LSTM is more accurate than the traditional models that demonstrate a deep learning model. In our case, LSTM (Long Short-Term Memory) is obviously an effective learner on training data than ARIMA, with the LSTM more capable of recognizing long-term dependencies. This study uses the daily price fluctuations of the Bitcoin to further investigate the model's predictability with minutely price fluctuations in the future.

In the future we plan to pull data from CryptoCompare API on a daily or monthly. In addition, we also take emotional data from other sources such as Twitter, Facebook,... to improve accuracy. Besides, we also want to apply the self-learning model to this project. The model will automatically update the best weights for training at regular intervals eg 1 week or 1 month after more data is available the model will update the weights once, then compare and find the weight with the best results and then put it into the model, which not only improves the results of the model's accuracy, but also helps the model not to be outdated.

Bibliography

[1] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 1735–80 (12 1997). https://doi.org/10.1162/neco.1997.9.8.1735

[2] Mohapatra, S., Ahmed, N., Alencar, P.: Kryptooracle: A real-time cryptocurrency price prediction platform using twitter sentiments. pp. 5544–5551 (12 2019). https://doi.org/10.1109/BigData47090.2019.9006554

[3] Raju, S.M., Tarif, A.M.: Real-time prediction of bitcoin price using machine learning techniques and public sentiment analysis (2020)

Bibliography

(1)(2) (3)