2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID) | 978-1-6654-1541-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/MSIEID52046.2020.00029

# Detection of Anomaly Stock Price Based on Time Series Deep Learning Models

Wenjie Yang
Shanghai Pinghe School
Shanghai, China
1498930151@qq.com

Ruofan Wang
Shanghai Pinghe School
Shanghai, China
frankwangcn@qq.com

Bofan Wang
Shanghai Pinghe School
Shanghai, China
2581433823@qq.com

***Abstract*—Anomaly detection is a critical task for financial market, investors, and regulatory authorities, where conventional methods employ rule-based models. With the development of machine learning and deep learning techniques, it becomes more promising to detect anomalous trading behaviors from data. Here we present a deep learning model based on time series LSTM model to detect anomalous behaviors in Chinese stock market. The model is composed of 1dConv-LSTM neurons, which can predict time series stock price data from historical data. We analyzed the price of 14 stocks variations ranging from 2015/01/05 to 2019/12/31 and used univariate and multivariate time series models to generate MAE less than 4.0 consistently. The proposed method improved MSE to 0.0171 on validation datasets. Our model successfully predicts the anomalous price behaviors of '601318' stock in the range of 2019-02-13. Our method provides an automatic way of predicting anomaly stock price behavior in Chinese stock market.**

***Keywords-Finance, DL, LSTM, Time Series, Anomalous Stock Price***

## I. INTRODUCTION

The Luckin Coffee fraud scandal in the last few months has been spread worldwide. According to the internal investigation, it shows that the fabrication of sales began in April 2019, which included inflating costs and expenses by almost $200 million, as well as booking $300 million in false revenue. After the announcement, Luckin's stock has slumped 32% [1]. Such a scandal made it harder for other Chinese companies to debut in the United States and thus lose a huge amount of US potential investors due to the untrustworthiness of Chinese companies. Besides, the event magnifies the defect of the Chinese stock market that requires the mechanism to oversee anomalous behaviors of stocks and judge which is a fraud for regulatory authorities. Only if the Chinese authorities establish the mechanism to identify anomalous stock price fluctuations and investigate hoax can investments be secure, and the stock market be more stable.

Various detection systems to monitor abnormal stock price changes have been developed [2][3]. Previously, most of the detection methods rely on a prediction-based method or rule-based method. Since mid-1997, the National Association of Securities Dealers (NASD) in the United States developed Advanced Detection System (ADS) that has been used to monitor trades and quotations in the NASDAQ stock market [2]. The ADS uses two pattern matches to detect abnormal behaviors. The system relies on a rule matcher, which detects predefined suspicious behaviors, and a time-sequence matcher, which looks for temporal relationships between events that exist in a potential violation pattern.

According to Time Series Contextual Anomaly Detection for Detecting Market Manipulation in Stock Market, the author proposed Contextual Anomaly Detection (CAD) method, which aims to use unsupervised way to exploit the behavior of similar time series to predict the expected values. The biggest problem of such a method is its lack of precision and accuracy of the result due to unknown and unlabeled data. It generally has a recall about 7% [3].

we proposed a deep learning model that can learn the historical trend from previous traded stocks prices and make predictions automatically. Our method will introduce a way of automatically flag anomaly stock price and volume change mechanism in Chinese stock market.

## II. DATASET

We analyzed **13** stocks in total, including companies that become listed in Shanghai and Shenzhen Stock exchanges. We gathered these data from dataset YouKuang [4], using codes such as

DataAPI.MktEqudGet(secID=u600448.XSHG",beginDate=u"20150101",endDate=u"",field=u"",pandas="1")

and saving data in the form of CSV for further analysis. Besides, we recorded the Open, Close Price, Turnover Value, Deal Amount, etc of each stock from 2015 to 2020.

Data is split into training and testing data according to the 'tradeDate', where it starts from 2015/01/05 to 2019/12/31. The trading date range is 1219 days, where we set days = 1000 as the split. The first 1000 days are set to training data, and the last 219 days are validating data.

978-1-6654-1541-5/20/$31.00 ©2020 IEEE
DOI 10.1109/MSIEID52046.2020.00029

Table 1. Example columns of dataset

| secID | ticker | secShortName | exchangeCD | tradeDate | preClosePrice | actPreClosePrice | openPrice | highestPrice | lowestPrice | closePrice | turnoverVol | turnoverValue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 600000.XSHG | 600000 | Shanghai pudong development bank | XSHG | 2015-01-05 | 15.69 | 15.69 | 15.88 | 16.25 | 15.56 | 16.07 | 513568709 | 8182820911 |
| 600000.XSHG | 600000 | Shanghai pudong development bank | XSHG | 2015-01-06 | 16.07 | 16.07 | 16.0 | 16.68 | 15.82 | 16.13 | 511684535 | 8311084820 |
| 600000.XSHG | 600000 | Shanghai pudong development bank | XSHG | 2015-01-07 | 16.13 | 16.13 | 15.9 | 16.17 | 15.53 | 15.81 | 385716820 | 6114241100 |
| 600000.XSHG | 600000 | Shanghai pudong development bank | XSHG | 2015-01-08 | 15.81 | 15.81 | 15.87 | 15.88 | 15.2 | 15.25 | 330627172 | 5101310595 |
| 600000.XSHG | 600000 | Shanghai pudong development bank | XSHG | 2015-01-09 | 15.25 | 15.25 | 15.2 | 16.25 | 15.11 | 15.43 | 491999937 | 7692348549 |

## III. METHODOLOGY

### A. Models

The model is tested for 13 random selected stock tickers, where we slice the data using 1D conv method and used the data to feed into the LSTM model that can predict the time series of 'stockPrice'.

#### 1) Conv1D-LSTM

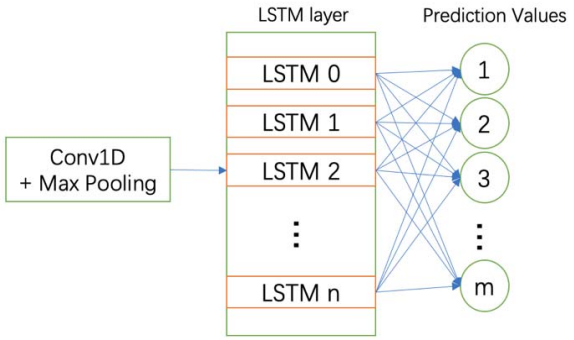


Figure 1. Structure of Conv1D-LSTM Model Prediction.

Conv1D-LSTM model is used to derive features from datasets, especially in time series data, As shown in Figure 1 [5][6][7][8]. The univariant and multivariant data was feed into the Conv1D layer to slice into readable length data for LSTM inputs. Using Conv1D layer can 'chop' time series data into the correct format to feed into LSTM. Here we window each 30 days as a block of data with a batch size equals to 32. The Conv1D uses 10 filters with a kernel size equal to 5 and a stride step 1. After this layer, sequential data is partitioned into various types of LSTM plus dense neural layers [9].

We demonstrate how the OpenPrice looks like in a stock, 'MinSheng', as shown in Figure 2. The volume and open price can be used to construct multivariant datasets.

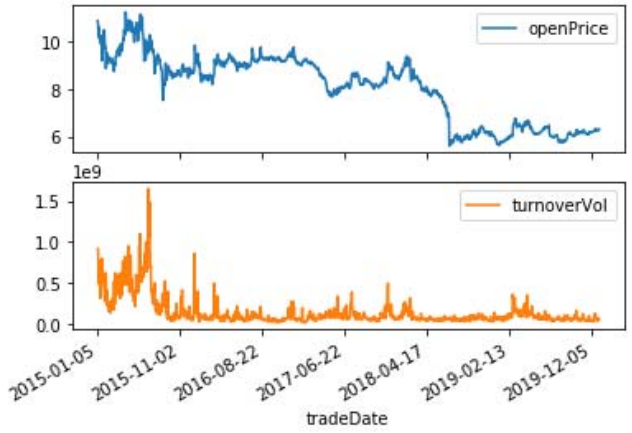


Figure 2. Time-Series data of MinSheng Stock showing the open price and turnover Volume.

#### 2) Non-Linear Activation Layer

Often, non-linear activation layer is employed after convolutional layer and fully-connected layer, since non-linear operation can help keep every change made in linear operation. There are many types of non-linear function, such as Logistic Function, Hyperbolic Tangent Function, Rectified Linear Function (ReLU), etc. In this article, we use ReLU as the activation function:

$$ReLU(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \quad (1)$$

which makes calculation speed faster than the other functions.

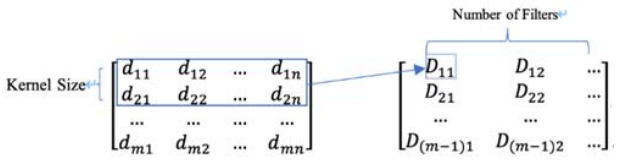


Figure 3. An Example of Conv1D-LSTM

*3) Long Short Term Memory Networks (LSTM)*

LSTM model is an autoregressive model much more efficient than RNN. The following graph (Figure 4) demonstrates the inner structure of this model:

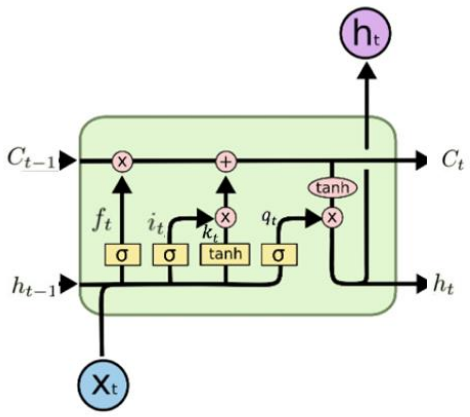


Figure 4. The Inner Structure of LSTM Unit (Picture Originally from Github [10])

The yellow rectangle represents Neural Network Layer, the pink circle represents pointwise operation and the arrows mainly stand for vectors.

TABLE I. TABLE TYPE STYLES

| Var | Meaning |
|---|---|
| X | Hadamard Product of the Information |
| + | Summation of Information |
| $\sigma$ | Sigmoid Layer |
| Tanh | Hyperbolic Function Layer |
| $h_t$ | the Output Vectors of current LSTM Cell |
| $C_t$ | the Current Vectors of Cell States |
| $X_t$ | Current Input Vectors |
| $b_f, b_i, b_k, b_q$ | Bias Vectors |
| $W_f, W_i, W_k, W_q$ | Weight Matrices |

There are four steps to pass through LSTM:

The first step is to determine which part of the input vector $h_{t-1}$ should be removed. Let $f_t$ denotes the activation value of forget gates at time $t$, then $f_t$ depends mainly on $h_{t-1}$ and $X_t$. The sigmoid layer forces the output $f_t \in [0, 1]$, where no information will pass if $f_t = 0$:

$$f_t = \sigma\left[W_f \times (h_{t-1} + X_t) + b_f\right] \quad (2)$$

The second step is to add additional information to the previous matrix, which consists of two parts: using a sigmoid layer to help decide which information will pass and computing the candidate vector which passes the Tanh layer. Let it, kt denote the activation function of the input gate and the candidate vector, respectively:

$$i_t = \sigma[W_i \times (h_{t-1} + X_t) + b_i] \quad (3)$$

$$k_t = \tanh\left[W_k \times (h_{t-1} + X_t) + b_k\right] \quad (4)$$

The third step is to combine the first and the second step by multiplying the information and their corresponding weights (activation values):

$$C_t = f_t \times C_{t-1} + i_t \times k_t \quad (5)$$

The last part is to calculate the output of this LSTM model. Similar to the second step, denote qt to be the activation value of output:

$$q_t = \sigma[W_q \times (h_{t-1} + X_t) + b_q] \quad (6)$$

$$h_t = q_t \times \tanh(C_t) \quad (7)$$

*B. Workflow*

As there are about 1219 data in a stock, thus each stock is divided into 1000 training data and 219 val/test data. First, we plot the whole data to see the stock's trend. Then, we train the training data through one Conv1D-LSTM layer, two LSTM layers and three Dense layers with "ReLU" as activation functions and figure out the best learning rate and assign a proper callback method for exponential learning rate decay. Later, we decide to consider "open price" and "turnover volume" as two important factors of stock price and use multistep model to build the corresponding neural network -- two LSTM layers and a Dense layer with "ReLU" activation function. By comparing predicted values with the actual values, anomalous stock prices can be found. Figure 5 below reflects the exponential decline of the learning rate when training the data.

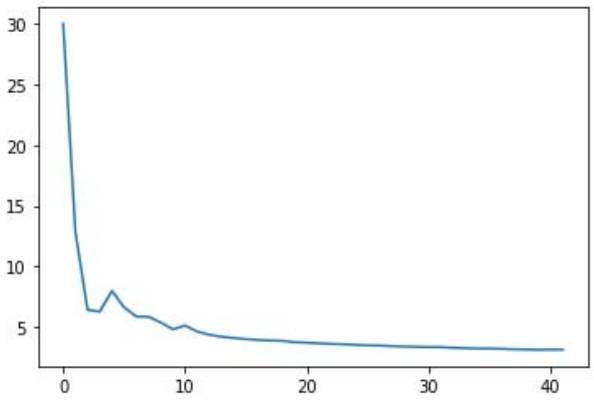


Figure 5. The Change of Learning Rate in Stock 'Beidouxingtong'

## IV. Result

### A. Model Performance

In order to evaluate the model performance in training and val/test dataset, we use MAE and MSE as our metrics. We use the first 1000 days open price and volume as multivariant data and use the validation data set of 1001 to end for model tuning purpose. Figure 6 shows a predicted example of stock '603718' using the Conv1D-LSTM model.

The naïve baseline model by shifting the date of one day gives a MAE more than 6.0 on all datasets. The example data '603718' has a MAE less than 4.5 where its prediction on validation captures the trend of open price.

### B. Predicted anomaly behavior

We randomly picked 13 stocks and used the range after 1250 to test if the model performance is statically different with previous period performances. After training, the model can predict unseen data from previous datasets. We used this model prediction as a predictor of possible anomaly behavior. If the deviation of MAE is larger than 30% on the training set, we mark the predicted period problematic. Using this strategy, we identify possible anomaly behavior of '601318' in the period of test as shown in Figure 7.

We can see that on 2019-02-13 where is set to 0 in the right, we observed abnormal volume and price jump that differs a lot from previous trends. This shows a possible point of anomaly behavior.

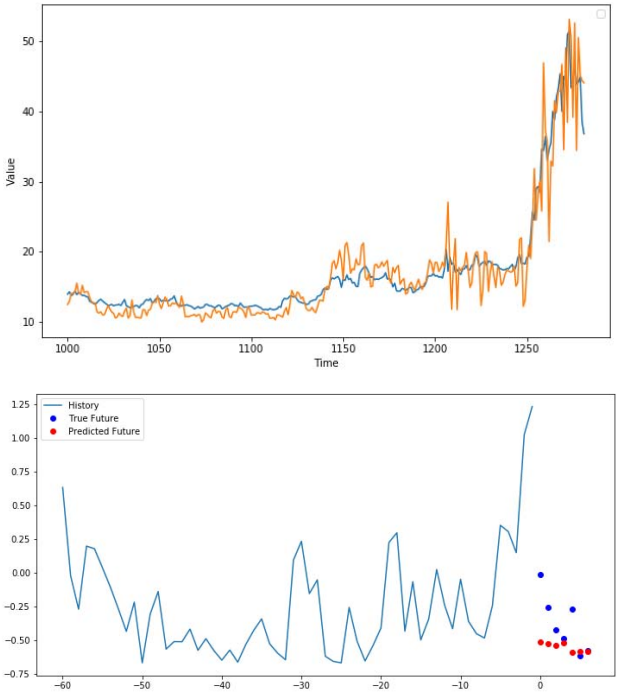


Figure 6. Model Prediction on Validation Dataset of 603718

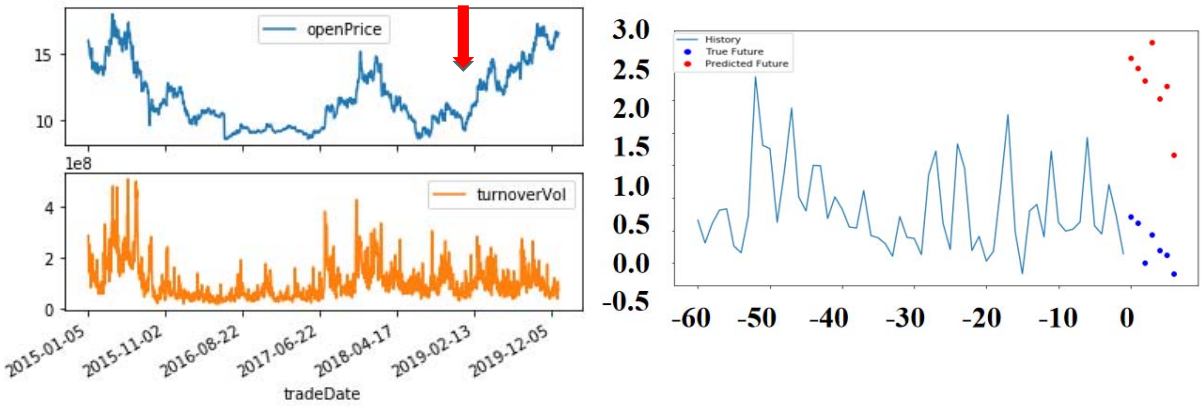


Figure 7. Anomaly Detection of 601318 in a 7 Days' Period

## V. Conclusion

In this paper, we proposed a deep learning model that can learn the historical trend from previous traded stocks prices and make predictions automatically. This model can lower the MAE to a low level, beyond which anomaly behavior of stock is possible. We demonstrated that in 2019-02-13 period of 7 days, it is possible to have high volume and price jump. This deep learning method can be an effective predictor of anomaly stock price behavior in China.


## Acknowledgment

This paper was completed under the guidance of Teacher Yu Yan. Thanks for the teacher's encouragement to all of our team members in the whole process and his timely help when we encountered problems, so that we have the motivation to go on! The specific division of the team is Yang Wenjie, who is responsible for the writing of Methodology, the control of global workflow, and the operation and summary of most data. Bofan Wang was responsible for searching the literature related to the research topic and looking for flaws in his research methods in order to optimize our model. Wang Ruofan mainly

contributed to participating in the topic discussion and running and processing some data.

## REFERENCES


[1] Luckin Scandal Is Bad Timing for U.S.-Listed Chinese Companies. Available at: https://www.bloomberg.com/news/features/2020-07-29/luckin-coffee-fraud-behind-starbucks-competitor-s-scandal. (Accessed: 2nd September 2020)

[2] Kim, Y. & Sohn, S. Y. Stock fraud detection using peer group analysis. *Expert Syst. Appl.* **39**, 8986–8992 (2012).

[3] Golmohammadi, K. & Zaiane, O. R. Time series contextual anomaly detection for detecting market manipulation in stock market. in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* 1–10 (IEEE, 2015). doi:10.1109/DSAA.2015.7344856

[4] QUER. *https://uqer.datayes.com/* (2020).

[5] Park, K. Prediction of Tier in Supply Chain Using LSTM and Conv1D-LSTM. *J. Soc. Korea Ind. Syst. Eng.* **43**, 120–125 (2020).

[6] Jain, S., Gupta, R. & Moghe, A. A. Stock Price Prediction on Daily Stock Data using Deep Neural Networks. in *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)* 1–13 (IEEE, 2018). doi:10.1109/ICACAT.2018.8933791

[7] Kotteti, C. M. M., Dong, X. & Qian, L. Rumor Detection on Time-Series of Tweets via Deep Learning. in *MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM)* 1–7 (IEEE, 2019). doi:10.1109/MILCOM47813.2019.9020895

[8] Polamuri, S. R., Srinivas, K. & Mohan, A. K. Multi model-Based Hybrid Prediction Algorithm (MM-HPA) for Stock Market Prices Prediction Framework (SMPPF). *Arab. J. Sci. Eng.* (2020). doi:10.1007/s13369-020-04782-2

[9] Fischer, T. & Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.* **270**, 654–669 (2018).

[10] Understanding LSTM Networks. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.