A study on stock price prediction system based on text mining method using LSTM and stock market news

Sunghyuck Hong
Professor, Baekseok University, Division of ICT

LSTM과 증시 뉴스를 활용한 텍스트 마이닝 기법 기반 주가 예측시스템 연구

홍성혁
백석대학교 ICT학부 교수

Abstract The stock price reflects people's psychology, and factors affecting the entire stock market include economic growth rate, economic rate, interest rate, trade balance, exchange rate, and currency. The domestic stock market is heavily influenced by the stock index of the United States and neighboring countries on the previous day, and the representative stock indexes are the Dow index, NASDAQ, and S & P500. Recently, research on stock price analysis using stock news has been actively conducted, and research is underway to predict the future based on past time series data through artificial intelligence-based analysis. However, even if the stock market is hit for a short period of time by the forecasting system, the market will no longer move according to the short-term strategy, and it will have to change anew. Therefore, this model monitored Samsung Electronics' stock data and news information through text mining, and presented a predictable model by showing the analyzed results.

Key Words : IoT, Server, Data Security, Blockchain, platform

요 약 주가는 사람들의 심리를 반영하고 있으며, 주식시장 전체에 영향을 미치는 요인으로는 경제성장률, 경제지료, 이자율, 무역수지, 환율, 통화량 등이 있다. 국내 주식시장은 전날 미국 및 주변 국가들의 주가지수에 영향을 많이 받고 있으며 대표적인 주가지수가 다우지수, 나스닥, S&P500이다. 최근 주가뉴스를 이용한 주가분석 연구가 활발히 진행되고 있으며, 인공지능 기반한 분석을 통하여 과거 시계열 데이터를 기반으로 미래를 예측하는 연구가 진행 중에 있다. 하지만, 주식시장은 예측시스템에 의해서 단기간 적중이 되더라도, 시장은 더 이상의 단기 전략대로 움직여지지 않고, 새롭게 변할 수밖에 없다. 따라서, 본 모델을 삼성전자 주식데이터와 뉴스 정보를 텍스트 마이닝으로 모니터링하여 분석한 결과를 나타내어 예측이 가능한 모델을 제시하였으며, 향후 종목별 예측을 통하여 실제 예측이 정확한지 확인하여 발전시켜 나갈 예정임

주제어 : 주가분석, 빅데이터, 텍스트마이닝, 인공지능, 예측 시스템

1. Introduction

There are some three areas with stock market problems. The point of social problem, investing in foreign and institutional stocks is a long-term investment, which affects the direction of the stock market by taking advantage of the long-term investment. The reason why the return of an individual investor is lower than the average return of KOSPI is that it has significantly lower accessibility to stocks, analytical ability, and risk management ability than foreigners and institutions [1]. Individual investors tend to choose stocks based on acquaintances' investment recommendations rather than technical analysis, and if profits are made, they tend to continue to listen to acquaintances and invest.

The point of economic problem, individual investors are more likely to lose money in terms of accessibility than foreign and institutional investors, while they have a large number of transactions, which bear a significant portion of the transaction tax.

As of 2018, individual investors in the stock market repeatedly paid securities transaction taxes of more than 5.55 trillion won by repeating frequent buying and selling in order to make short-term profits. KOSPI's return increased by 8% compared to 2019, but foreigner's return reached 20%.

The current stock price prediction system analyzes past data and uses inference through time series and artificial intelligence to learn the fluctuation patterns of stock prices and use it in a way to predict the future. However, there are many factors that affect the stock price, and it is virtually impossible to predict the future from past data. In the stock market, the past trading patterns are not repeated, international relations, economic indicators, specific events, celebrity comments Every moment is affected by a new situation Even if a deep learning algorithm is applied, it is almost impossible to predict the price of stocks that change from time to time.

In order to respond to the changing stock market in real time, it uses artificial intelligence LSTM and YTextMiner to reflect securities news in real time and, based on past time series analysis data, finds the closest situation to when the stock price has risen by mathematical calculation [2, 3]. It enable to predict stock.

2. Goal of Stock Prediction System

2.1 Needs and Goals

Using deep learning-based LSTM and YTextMiner, individuals can also make technical and rational investments at the foreign and institutional level, and make reasonable investments.

It is possible to determine the optimal buying and selling point through analysis of technical stock price changes and market environment changes. Automated stock trading system reduces stock monitoring time and contributes to establishing a sound investment culture rather than speculation. Achieve an increase in the return of individual investors by inducing reasonable investment decisions without being affected by the investment propensity of foreign and institutional investors

Decide the timing of buying and selling through deep learning, reducing transaction frequency to reduce stock trading costs and taxes [4]. Contributing to the development of national industry and strengthening of national competitiveness by investing in competitive industries. Using artificial intelligence LSTM to learn past data using deep learning techniques to generate predictive data, and using YTextMiner to judge real-time stock market news as good news and bad news, and adjusting the weight when generating predictive data. Construct a more systematic stock price forecasting system based on time series data

2.1.1 LSTM model

The present research analyzes the impact on the stock price by analyzing real-time stock market news using inference through time series and analysis of past stock prices using the Recurrent Neural Network (RNN) and analyzing the effect on the stock price. The development of a program that predicts short-term stock prices by giving a specific stock price analysis model for each item to develop and apply it to securities firms and investment institutions is a circulating neural network (RNN) [5] and stock price prediction technology using news mining. The market forecast program is expected to have high profits through business models due to its various fields of application.

Predicted by analyzing additional factors such as KOSPI index, KOSDAQ index, exchange rate information, buying and selling amount by investor, transaction price, sub-indicator, stock news, and company earnings report through RNN does.

LSTM (Long Short Term Memory networks) is a type of RNN (Recurrent Neural Network), and it is the most suitable model for stock price prediction as a structure in which past data can affect the future. LSTM is able to solve the problem that the values between [-1, 1] are multiplied continuously by the Chain Rule, a long-term dependency problem, and the value becomes smaller as it goes to the front. (1997). The existing RNN simply calculated St = tanh (Uxt + WSt) when calculating Hidden State (St) in the Hidden Layer, but LSTM has a total of four calculation processes. As you can see in the LSTM's Hidden Layer below, there are four Neural Network Layers (a square shape in Fig. 1). See Fig. 1. Also, the core of the LSTM is a very minor linear operation on the horizontal line in the upper part of the neural (A) in Fig. 1, and passes through the entire chain, so that the information is transferred to the next step without change.
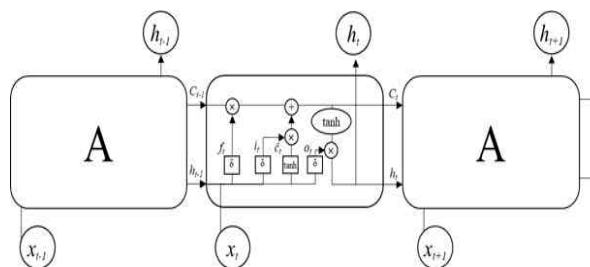

Fig. 1. Basic Structure of LSTM

2.1.2 YTextMiner System

YTextMiner is a tool that analyzes text and uses real-time stock market news in the present invention to determine whether news corresponding to a stock item by keyword is good or bad, train time series data through LSTM, and use it to adjust weights when predicting. The stock prediction system through deep learning does not recognize the market in real-time, so it is used to analyze the collected data through real-time news mining to improve accuracy with reference to the final stock price prediction.

The stock price prediction system is based on the past stock average value ((high price + author) / 2), and the reference day calculates the increment on the next day with the stock price as 0 see Table 1.

Table 1 is a Samsung Electronics stock data set from December 2019 to January 2020. If No. 1 is the reference, the reference day is the increment value, and the stock price is less than 1 or greater than -1 to normalize. Process. The increments in Table 1 are Middle (No.2) / Middle (No.1) -1, the average price (average price) of high and low prices is calculated, and then the increment is calculated.

Input values are incrementals in Table 1 and LSTM can predict the future value of the price. However, stock prices are based on stock news. For example, the CEO of Samsung is indicted or company's profile goes up or down. Therefore, LSTM is not enough for prediction for changeable stock market. Because there are lots

of investment money.

Table 1. Stock Data Set

No	Date	High	Low	Average	Incremental
1	2019-12-09	51400	50700	51050	0
2	2019-12-10	51600	50700	51150	0.001958864
3	2019-12-11	52200	51400	51800	0.012707722
4	2019-12-12	53300	52700	53000	0.023166023
5	2019-12-09	51400	50700	54350	0.025471698
	...	...	...	...	...

3. Factors on Stock Price

3.1 External Factor

External factors affecting share prices include economy, currency volume, prices, interest rates, exchange rates, and commodity prices. Of course, stock prices move before these changes in economic phenomena. In other words, if the game is expected to improve, the stock price will start to rise more than six months before the game actually improves.

A better economy doesn't mean the stocks of all industries will rise. Therefore, as the game improves, you need to choose a business with many benefits. If the amount of money increases, there is a high possibility that the stock price will rise due to the abundance of funds in the stock market, but if it increases too much, the price may fall due to the inflation of the value of the currency. Inflation factors also have different impacts on stock prices depending on the mode of inflation. A sharp rise in inflation freezes consumer sentiment, which can reduce a company's profits, leading to a fall in stock prices. On the other hand, a modest rise in inflation as the economy recovers may improve the company's performance and raise the share price.

When the interest rate rises, the interest burden of the debtor company increases, which adversely affects the company. Investors are also willing to take stable risks and choose stable

banking products rather than investing in stocks. This is a big effect on the stock price. .

An increase in the exchange rate increases the likelihood that the share price of a company focusing on exports will rise. Since one dollar is equal to 1,000 won in our country, and the value of our country's money is relatively low, it is equal to 1,300 won, so we exported 1 dollar worth of goods and received 1,000 won in the past, but now the exchange rate has risen to 1,300 won. Because it is. On the other hand, it is disadvantageous for companies that mainly import. It was because you had to pay 1,000 won in the past while importing as much as 1 dollar, but now you have to pay 1,300 won when the exchange rate has risen. As such, if there is a sign of an increase in the exchange rate, the share price of companies that mainly export is likely to rise, and the share price of companies that mainly import is likely to decline. However, not only exchange rate fluctuations, but also other economic phenomena are always changing, so we need to analyze closely how the complex influence will affect the share price.

3.2 Internal Factor

The internal factors that affect the share price include the company's management performance or sales performance, management qualities, new product development, and new business entry. Where can I get information about them? Companies that issue stocks are obliged to provide changes to important corporate content, such as financial and business conditions, to investors through the stock exchange. This is referred to as the disclosure system, and the types of disclosure include 'regular disclosure' that publishes business reports and shareholders' meeting results reports, 'temporary disclosure' on occasional changes in corporate contents, and 'inquiries about the facts of rumors and newspaper reports' Disclosure '. Disclosures can be conveniently viewed by the Financial

Supervisory Service's electronic disclosure system by listing on exchanges and by KOSDAQ registered corporations.

3.3 Test Environment

The testing sample code is based on Python 3 or lower. Samsung Electronics past data from Yahoo finance was used for analysis. The data I use is contained in the 005930.csv file and consists of the closing price of the Yahoo finance from January 2016 to January 2020. Jupyter notebook [11], Tensorflow [14], and Keras. Tensorflow are used, and it is end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications. The Jupyter Notebook is a widely open web application tool that it can be used for creating and sharing documents which contain text, visualizations, equations, and live code. In addition, Jupyter Notebook is maintained by the people at Project Jupyter. Jupyter Notebooks are a spin-off project from the IPython[15] project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that can be used as well [12-15]. Currently, Python 3.7 is available. However, Python 3.7 doesn't fully support Keras library. Thus, we used Python 3.6 instead the current version. There is a minor modification to use this code, if you use Python 3.7 in Fig. 2.

```
#Step 1: import for past stock past data
data=pd.read_csv('005930.KS.csv')
data.head()
```

```
#Step 2: Calculate average price
high_prices = data['High'].values
low_prices = data['Low'].values
average_prices = (high_prices + low_prices) / 2
```

Fig. 2. Python 2.0 source code for data import

Table 2. Key Factors Affecting on Stock Price

No	Type	Incremental(I)	Formula(I×W×100)	Weighting(W)
1	economy	1.3%	13	0.1
2	Exchange Rate	−1.5%	−15	0.1
3	KOSPI	1.0%	10	0.1
4	Interest rate	0.1%	1	0.1
5	SAP500	1.2%	6	0.05
6	NASDAQ	0.5%	2.5	0.05
7	ROE(Return on Equity)	1.1%	11	0.1
8	Currency	1.0%	10	0.1
9	operating profit	3.5%	52.5	0.15
10	PER(Price Earnings Ratio)	2.4%	36	0.15
	Sum	10.6%	127	1

127 in Table 2 means calculating all affecting factors total, so LSTM [6] will return the future value (FV), then 127/100*FV is going to be final future price for expected stock price [7]. After time goes on, the actual stock price will come and compare these future price which is calculated by LSTM and key factors affecting on stock Price. Deep learning [8] skill keeps learning its future value and news weight value total. It is getting applied for accurate price as it collects lots of data from LSTM and text miners. The weighting values in Table 2 will be changed based on the actual price and continue to calculate the future price with adopting and changing weight values.

4. Conclusion

Although the existing stock price prediction system analyzes past data and uses inference through time series and artificial intelligence to learn how to change the stock price and predict the future, the core idea of this study is that the

factors that affect the stock price is that it is virtually impossible to predict the future with data from a variety of past, so there is a difference in reflecting the result value using text mining so that you can quickly respond to new situations every moment. In order to respond to the changing stock market in real time, it uses artificial intelligence LSTM and YTextMiner to reflect stock news in real time and, based on past time series analysis data, finds the closest situation to when the stock price has risen by mathematical calculation.

REFERENCES

[1] J. P. Tang. (2015r). The Effect on KOSPI 200 Futures after Launching KOSPI 200 Option. In 2015 *International Conference on Industrial Technology and Management Science*. Atlantis Press. . doi:10.2991/itms-15.2015.347

[2] H. Xue, D. Q. Huynh & M. Reynolds. (2018). SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1186-1194). *IEEE*.

doi:10.1109/wacv.2018.00135

[3]Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R & Delen, D. (2012). Text Mining PubMed. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, *Academic Press.* 703-750.

doi: 10.1016/b978-0-12-386979-1.00030-x

[4] S. Stoikov & R. Waeber. (2015). Reducing Transaction Costs with Low-Latency Trading Algorithms. *SSRN Electronic Journal.* doi:10.2139/ssrn.2661618

[5] R. Sproat & N. Jaitly. (2017). An RNN Model of Text Normalization. Interspeech 2017. doi: 10.21437/interspeech.2017-35

[6] Q. Chen, B. Liang & J. Wang. (2019). A Comparative Study of LSTM and Phased LSTM for Gait Prediction. *International Journal of Artificial Intelligence & Applications, 10(4),* 57-66. doi: 10.5121/ijaia.2019.10405

[7] R. Aviandy.. (2016). Analysis of Financial performance affecting Stock Price on Pharmaceutical Industry. *Business and Entrepreneurial Review, 7(1),* 69. doi:10.25105/ber.v7i1.1184

[8] S. H. Koh. (2018). A Comparative Study on the Excess Returns of Growth Stocks and Value Stocks in the Korean Stock Market. *Journal of the Korea Convergence Society, 9(7),* 213-222.

[9] S. H. Koh. (2015). Convergent Momentum Strategy in the Korean Stock Market. *Journal of the Korea Convergence Society, 6(4),* 127-132.

[10] S. H. Koh. (2016). A Converging Approach on Investment Strategies, Past Financial Information, and Investors' Behavioral Bias in the Korean Stock Market. *Journal of the Korea Convergence Society, 7(6),* 205-212.

[11] Supplemental Information 2: GSS Calculations Jupyter Notebook. (n.d.). doi:10.7287/peerj.preprints.26665v2/supp-2

[12] R. Bandi & J. Amudhavel. (2018). Object Recognition Using Keras with Backend Tensor Flow. International *Journal of Engineering & Technology, 7(3.6)*, 229. doi:10.14419/ijet.v7i3.6.14977

[13] M. L. Hetland (2010). Python Algorithms: Mastering Basic Algorithms in the Python Language. doi:10.1007/978-1-4302-3238-4

[14] K. Detlefsen. (2018). Monte Carlo Option Pricing in Tensorflow. *SSRN Electronic Journal.* doi:10.2139/ssrn.3214058

[15] B. Zhao, X. Li, & X. Lu. (2018). HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* doi:10.1109/cvpr.2018.00773

홍 성 혁(Sunghyuck Hong)   [종신회원]

· 2007년 8월 : Texas Tech University, Computer Science (공학박사)
· 2012년 3월 ~ 현재 : 백석대학교 정보통신학부 부교수
· 관심분야 : 블록체인, 사물인터넷 보안, 경량보안프로토콜
· E-Mail : shong@bu.ac.kr