Stock Price Prediction System*

Chanyoung Lee[1], Yujin Seo[2], and Donghun Jung[3]

[1] Sungkyunkwan University, Computer Science and Engineering, Republic of Korea
[2] Sungkyunkwan University, Department of Mathematics, Republic of Korea
[3] Sungkyunkwan University, Department of Physcis, Republic of Korea

Abstract. This project aims to predict the stock price of some selected stocks in KODEX 200 and S&P 500 on a day-by-day basis. Leveraging data sourced from the Korea Exchange (KRX) and Nasdaq, we will undertake the training and evaluation of a diverse set of machine learning models, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformer. We focus on predicting the stock price of the next a few days. In a contrast to the general stock price prediction system of many stock firms, we do not analyse the fundamental of the company. We focus on the short-term changes of the stock price, and attempts to reflect the recent issues via VADER.

Keywords: Stock Price Prediction · Machine Learning

1 Introduction

In today's capitalist world, stocks present an attractive avenue for financial gain, offering significant profit potential. Investors, regardless of their depth of understanding of the stock market, are constantly seeking reliable methods to predict future stock prices, as their financial well-being often hinges on these predictions.

Predicting stock prices is fundamentally about discerning the real value of a company. Knowing a company's true value helps in determining whether its stock price will rise or fall from its current level. However, this is a complex task influenced by a myriad of factors, including global events and economic conditions.

Historically, professional fund managers have been pivotal in guiding investments in the stock market. However, their strategies and predictions are susceptible to personal biases. This limitation has spurred interest in leveraging machine learning for more objective and unbiased stock price forecasting.

While numerous services today use machine learning for this purpose, many are not open-source and are primarily geared towards professional investors. These systems, often complex and tailored for experienced users, focus more on asset allocation than on precise stock price prediction. In contrast, our project targets the general public, who rely more on simple chart analysis and intuition, without delving deep into a company's fundamentals or news events.

* Supported by LINC

Our proposal is to predict the short-term movement of blue-chip stocks, a strategy known to be effective over periods ranging from a few days to a few months. We aim to make short-term predictions, not in real-time, to assist general users and simplify the prediction process.

To achieve this, we have employed machine learning techniques, utilizing models such as Long Short-Term Memory(LSTM), Gated Recurrent Unit(GRU), and Transformer. These models have been instrumental in successfully predicting stock prices for the next few days by capturing short-term trends. In addition, we have integrated VADER, a sentiment analysis tool, to reflect recent issues and capture both short- and long-term stock price changes based on current events.

Our system has been deployed as a publicly accessible web service, offering users easy access to our predictive model. This deployment significantly contributes to investment decision-making, lowering the entry barrier for those interested in stock market investments. Empirical evaluations of our system have shown promising results in capturing short-term trend of stock price movements, thereby validating our approach and offering a valuable tool for investors.

2 Design for the Proposed System/Solution/Service

2.1 LSTM

Overview of LSTM LSTM is a sophisticated architecture within Recurrent Neural Networks (RNNs), designed to address the limitations of traditional RNNs in capturing long-range dependencies in sequential data, a challenge often compounded by the vanishing gradient problem. The hidden layer of an LSTM consists of memory cells interconnected through a series of gates: the input gate, forget gate, and output gate, which collectively facilitate short-term memory storage.

1. Forget gate: Marked by a blue box in the figure, the forget gate determines which information from the previous state should be discarded or retained. Mathematically, it is represented as:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{1}$$

   where f_t is the output of the forget gate layer at the time step t, W_f, U_f, b_f are the weight matrices and bias vectors for the forget gate computation, respectively.
2. Denoted by an orange box, the input gate decides what new information should be stored in the cell. It operates as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$

$$g_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3}$$

   where i_t is the output of the input gate layer at the time step t, g_t is the candidate value to be added to the output layer at the time step t, W_i, W_c,

U_i, U_c, b_i, b_c are the weight matrices and bias vectors for the input gate and candiate value computation, respectively.

3. Output gate: Shown as a gray box, the output gate determines which information from the cell should be used to generate the output. It is defined by:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{4}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$
$$h_t = o_t \odot tanh(c_t) \tag{6}$$

where o_t is the output of the output gate layer at the time step t, c_t is the status of the cell at the time step t, h_t is the output of the LSTM unit at the time step t, W_o, U_o, b_o are the weight matrices and bias vectors for the output gate computation, respectively.

Compared to the traditional RNN, LSTM performs various mathematical operations, including including element-wise multiplication and addition, to control the flow of information and perform updates to the memory cell and hidden state.

Compared to traditional RNNs, LSTMs perform a variety of mathematical operations, including element-wise multiplication and addition, to control the flow of information and update the memory cell and hidden state. This architecture enables LSTMs to effectively handle long sequential data, making them capable of capturing time-dependent patterns in the data.

Background of employing LSTM Stock prices are typically regarded as time series data, characterized by their sequential nature and the presence of temporal dependencies. LSTM is particularly well-suited for modeling such data due to their ability to capture both short-term and long-term dependencies and patterns. This capability stems from their unique architecture, which allows them to remember information over extended periods and forget irrelevant data, a critical feature for analyzing the often volatile and non-linear patterns observed in stock market data.

In the context of stock price prediction, LSTMs can learn from historical price data, identifying underlying trends and patterns that are indicative of future movements. This makes them a powerful tool for financial analysts and investors seeking to make informed decisions based on predictive analytics. The use of LSTM in this domain is driven by the hypothesis that past stock performance, along with other relevant financial indicators, can provide valuable insights into future price trajectories.

2.2 GRU

Overview of GRU GRU is a type of RNN architecture, offering a simpler alternative to LSTM. Despite their simpler structure, GRUs have proven highly

effective in various applications. They are designed to address the vanishing gradient problem, enabling RNNs to better capture long-range dependencies in sequential data. A distinctive feature of GRUs, as compared to LSTMs, is their unified approach to managing cell state and output, which simplifies the architecture.

GRUs utilize two main gates for their operation:

1. Update gate: The update gate determines how much of the past information needs to be passed along to the future.

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{7}$$

2. Reset gate: The reset gate decides how much of the past information to forget.

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{8}$$

These gates help the GRU to make decisions about what information is relevant to keep from past time steps and what can be discarded, enabling it to capture dependencies over different time scales effectively.

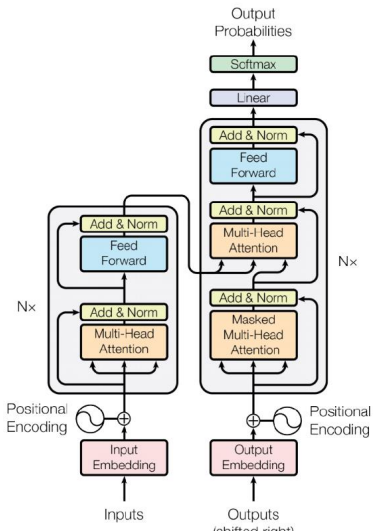
Background of employing GRU Similar to LSTMs, GRUs are employed with the expectation that they can capture both short-term and long-term dependencies and patterns in sequential data. GRU, with their simplified structure, offers an efficient way to model these complexities.

The choice of GRU for stock price prediction is motivated by its efficiency in training and its effectiveness in handling sequences where the gap between the relevant information and the point where it is needed is unknown or variable. This makes GRUs particularly suitable for analyzing financial time series data, where they can learn from historical price movements to predict future trends and patterns.

2.3 Transformer

The Transformer is a neural network architecture that has revolutionized the field of natural language processing (NLP) and has shown promising potential in time series prediction tasks. Unlike traditional RNN-based models like LSTM and GRU, the Transformer relies entirely on an attention mechanism.

Overview of Transformer Transformer operates as an Encoder-Decoder model, leveraging an attention mechanism. In the Encoder-Decoder architecture, the Encoder takes an input sequence and encodes the information into a single context vector. Conversely, the



Decoder utilizes this context vector to generate an output sequence. Within the model structure, a crucial component is the "embedding" process. This process involves the conversion of input values into a unified vector representation.

The attention mechanism employed by the Transformer is a key feature. It assigns varying weights to elements within the input sequence, placing greater emphasis on pertinent information. This emphasis is then reflected in the model's output. The Transformer employs this mechanism to comprehensively evaluate the significance of the entire input sequence when generating the output.

The utilization of the Transformer model holds the promise of delivering superior performance. It has the capacity to incorporate diverse sources of information such as news, stock indices, and corporate disclosures, distinguishing it from other models.

Background of employing Transformer The decision to employ the Transformer model in stock price prediction stems from its advanced capabilities in handling sequential data. While LSTM and GRU have made significant strides in addressing the vanishing gradient problem, they still have limitations, particularly in their sequential processing nature and in fully capturing long-range dependencies. The expectation is that the Transformer, with its advanced architecture, will outperform traditional models like LSTM and GRU in capturing the intricate patterns and dependencies inherent in stock price data, leading to more accurate and reliable predictions.

2.4 Loss function

Overview of Loss function In our model, the loss function is defined as:

$$\mathcal{L} = \mathcal{L}_H + \lambda \sqrt{\sum \left(\frac{\hat{y}_t - y_t}{y_{t-1} - y_t} \right)^2} \tag{9}$$

where $\mathcal{L}_H$ is Huber loss function, λ is a regularization parameter, y_t is the actual value at time step t, and $\hat{y}_t$ is the predicted value at time step t. The Huber loss function is given by:

$$\mathcal{L}_H = \sum_{t=1}^{n} \begin{cases} \frac{1}{2}(y_t - \hat{y}_t)^2 & \text{tf } |y_t - \hat{y}_t| \leq \delta \\ \delta |y_t - \hat{y}_t| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \tag{10}$$

The Huber loss function is chosen for its balanced approach, combining the robustness of the L1 norm with the smoothness of the L2 norm. This makes it differentiable and less sensitive to outliers in the data. Visually, the Huber loss resembles the L2 norm for smaller errors but transitions to an L1 norm-like behavior for larger errors, effectively handling outliers.

Additionally, we incorporated a regularization term into the loss function. This term is designed to provide some leniency for rapid changes in stock prices. It implies that the model is less penalized for missing rapid price changes, acknowledging the inherent volatility and unpredictability in stock price movements.

Background of employing Loss function Initially, our model employed the Huber loss function without a regularization term. However, we observed that the model tended to predict the next day's stock price as being equal to the current day's price, a pattern inconsistent with real-world stock market behavior. This phenomenon aligns with the martingale theory in probability, where the expected next value of a sequence of random variables is simply the current value.

Recognizing that stock prices are not merely a sequence of random variables but are influenced by trends and patterns, we introduced the regularization term. This addition encourages the model to recognize and respond to short-term trends in stock prices, rather than strictly adhering to the immediate past value. It allows the model to be more responsive to rapid changes, aligning its predictions more closely with the dynamic nature of the stock market.

2.5 VADER

Overview of VADER VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based model specifically designed for sentiment analysis of social media text. It excels in analyzing the sentiment of individual words and phrases, providing an overall sentiment score for the text. VADER employs a combination of lexical elements and grammatical heuristics to assess sentiment, effectively capturing both the polarity (positive or negative) and intensity of emotions expressed in the text.

One of VADER's strengths lies in its ability to adeptly handle the unique characteristics of social media text, such as slang, emoticons, and abbreviations, which are often challenging for traditional sentiment analysis models. Its performance in sentiment analysis tasks, especially in the context of social media, is notably high.

Background of employing VADER In our stock price prediction model, we integrated VADER to incorporate additional factors that could influence stock prices. While our model effectively captures short-term trends in stock prices—predicting rises or falls based on recent price movements—it sometimes struggles to anticipate when these trends might not hold.

To address this, we turned to external factors like news, which can significantly impact stock prices both in the short and long term. News sentiment, whether positive or negative, along with the intensity of the sentiment, can be pivotal in shaping stock market trends. VADER's ability to evaluate these sentiments provides a nuanced understanding of how news might affect stock prices.

We use VADER to generate sentiment scores for news articles, incorporating these scores as new factors in our model. This approach aims to enhance the model's predictive accuracy by accounting for the influence of news sentiment on stock price movements, capturing new trends that might not be evident from historical price data alone.

3 Implementation

3.1 Data Set

Stock Selection For our prediction model, we carefully selected approximately 15 stocks from both the KODEX 200 and S&P 500 indices. o add a layer of analysis, we categorized these stocks into two distinct groups: stable and unstable. This categorization was based on their presence in commercial ETF fund holdings. The table3.1 below details the specific stocks chosen for our study.

Korea Stable	**Korea unstable**	**US Nasdaq Stable**	**US Nasdaq unstable**
Item 1	Item 2	Item 3	Item 4
Item 1	Item 2	Item 3	Item 4

Normalization To standardize the stock price data for analysis, we applied a normalization process. Let P_t be the stock price of the t-th day, and P_i' denote the normalized stock price on the same day. The normalization formula is as follows: Then, we have

$$P_t' = \frac{P_t}{P_{\max} - P_{\min}} \tag{11}$$

where P_{max} and P_{min} are the maximum and minimum stock prices within the specified period, respectively. This normalization allows for a more consistent comparison across different stocks and time frames.

Training Set Our training dataset comprises various stock price metrics—open price, close price, high price, and low price—from the period of 2021 to 2022. Due to challenges in sourcing older news articles necessary for VADER analysis, the timeframe for our training data is thus constrained.

Test Set For testing, we utilized stock price data (including open, close, high, and low prices) from the period of February 17, 2023, to December 1, 2023. Our testing approach involved using a set of 30 days' worth of stock prices and sentiment scores to predict the stock price for the subsequent day.

3.2 UI/UX Design

The user experience (UX) and user interface (UI) aspects of the system were prioritized to ensure a seamless and intuitive user interaction. The frontend was designed with a user-centric approach, focusing on usability and visual appeal. The UI elements were designed to be responsive and accessible across different devices and screen sizes, enhancing the overall user experience.

3.3 Frontend

3.4 Backend

Django, a Python web framework, was employed for server-side development.

4 Evaluation

4.1 Evaluation Metric

Mean Prediction Accuracy (MPA) To assess the effectiveness of our methods, we calculate the Mean Prediction Accuracy (MPA), defined as:

$$\text{MPA}_t = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{y_i} \tag{12}$$

where N is the number of stocks, $\hat{y}_i$ and y_i represent the predicted and actual high or low prices, respectively, for the t-th day of stock i.

Mean Absolute Error (MAE) The Mean Absolute Error (MAE) is another metric used to evaluate our methods:

$$\text{MAE}_t = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{13}$$

Here, N denotes the number of stocks, with $\hat{y}_i$ and y_i being the predicted and actual high or low prices for the t-th day.

Trend Accuracy (TAC) Given our focus on short-term stock price trends, we also compute the Trend Accuracy (TAC):

$$\text{TA}_t = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\text{sign}(\hat{y}_i - \hat{y}_{t-1}) = \text{sign}(y_i - y_{t-1})) \tag{14}$$

where N is the number of stocks, $\hat{y}_i$ and y_i are the predicted and actual high or low prices for the t-th day, and $\mathbb{1}(\cdot)$ is the indicator function.

Accuracy To compare the similarity between the predicted and actual price intervals, we calculate the accuracy(ACC):

$$\text{Accuracy}_t = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{length}(\{y_{\min} < y < y_{\max} \cap \hat{y}_{\min} < y < \hat{y}_{\max}\})}{\max(\text{length}(\hat{y}_{\min} < y < \hat{y}_{\max}), \text{length}(y_{\min} < y < y_{\max}))} \tag{15}$$

where N is the number of stocks, $\hat{y}_{\min}$ and $\hat{y}_{\max}$ are the predicted minimum and maximum prices for the t-th day, and $y_{\min}$ and $y_{\max}$ are the actual minimum and maximum prices for the t-th day.

4.2 Result

		STABLE				UNSTABLE			
Model		**MPA**	**MAE**	**TAC**	**ACC**	**MPA**	**MAE**	**TAC**	**ACC**
LSTM	High	0.9922	1.2179	0.5640	0.4183	0.9856	4.2027	0.5872	0.4163
	Low	0.9921	1.1985	0.5702		0.9851	4.1493	0.5711	
GRU	High	0.9929	1.1118	0.5640	0.4340	0.9864	3.9909	0.5930	0.4199
	Low	0.9926	1.1193	0.5659		0.9862	3.8426	0.5747	
Transformer	High	0.9906	1.4721	0.4743	0.4295	0.9827	5.2046	0.4731	0.3374
	Low	0.9892	1.6502	0.5361		0.9830	4.8531	0.4616	
BASELINE		0.9815				0.9815			

As shown in Table 4.2, the performance of our models on the test set is quantified. Incorporating VADER into the model has resulted in MPA surpassing that of the baseline model, which is DP-LSTM. Notably, the model's performance varies based on the stability of the stocks. For unstable stocks, the MAE significantly increases, and ACC decreases. Both LSTM and GRU models exhibit similar performance metrics, whereas the Transformer model demonstrates slightly lower performance compared to the other two models.

5 Limitation and Discussions

5.1 Prediction Quality

As illustrated in Figure 2, our models demonstrate a tendency to follow the trend of stock prices. If the stock price has been increasing up to the current day, the model predicts a continued increase for the next day, and similarly for a decreasing trend. This indicates that the models are effectively capturing the short-term trend of stock prices rather than merely reflecting random fluctuations. However, it's noteworthy that while the models occasionally attempt to predict price rebounds, these predictions are not consistently successful.

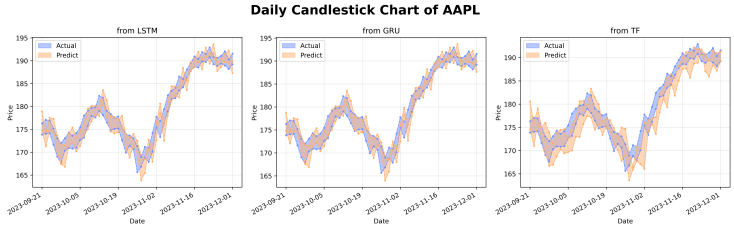


Fig. 2.

The impact of incorporating VADER for sentiment analysis on stock price prediction is not definitively clear. Although the sentiment score is more aligned with long-term trends, the models still achieve a higher MPA score compared to the BASELINE model, which does not utilize external information. This suggests a potential, albeit indirect, benefit of including sentiment analysis in the prediction process.

A challenge arises when the stock price remains relatively constant with minor fluctuations. In such scenarios, the models often predict in the opposite direction of the actual trend, leading to inaccuracies. This issue is particularly evident in the period from November 16, 2023, to December 1, 2023, as shown in Figure 2.

Regarding the comparison of LSTM, GRU, and Transformer models, LSTM and GRU exhibit similar prediction quality, but the Transformer model lags slightly behind. As indicated in Table 4.2, the Transformer model has a higher MAE and lower scores in MPA, TAC, and ACC compared to LSTM and GRU. Notably, the Transformer model's predicted price range (high to low) is broader and less consistent than the other models, leading to more frequent misses in trend prediction. This results in its lower TAC and ACC scores.

Despite these challenges, the Transformer model holds promise, particularly due to its capability to interpret long-term trends and its flexibility in integrating additional features. In our project, we limited the input features to stock prices and sentiment scores, which may not fully leverage the Transformer model's potential. Future improvements and the inclusion of more diverse features could enhance the Transformer model's applicability in real-world scenarios.

5.2 Lack of Information

A significant challenge in our project arises from the constraints in collecting sentiment scores using the VADER model. For this purpose, we sourced news articles from the Investing.com website (`investing.com`). However, we encountered a limitation in the availability and distribution of these articles. The website provides approximately 1000 news articles per stock, but this distribution is highly uneven across different stocks. For instance, stocks like AAPL and TSLA have a daily influx of articles, whereas other stocks receive considerably less coverage. This disparity in news frequency is a primary reason why we could

not effectively apply the VADER model to stocks in the KODEX 200, limiting our analysis and results.

Another constraint is the temporal availability of news articles. The website only offers access to current day's news, preventing us from utilizing historical data in our training set. The ability to access past news articles would significantly enhance our model's training process, allowing us to incorporate sentiment scores from earlier periods. This would likely improve the model's ability to understand and predict stock price movements based on a more comprehensive historical context. Furthermore, access to professional analysts' reports could potentially elevate the prediction quality of our model.

6 Related Work

Recent research trends in stock price prediction include advancements in deep learning-based regression models. [1] These models often utilize Long Short-Term Memory (LSTM) networks and innovative validation techniques, such as walk-forward validation, to enhance their predictive capabilities.

In addition, some researchers have explored Particle Filter Recurrent Neural Networks (PF-RNNs), a new RNN family explicitly designed to model uncertainty within their internal structure. [2] Unlike traditional RNNs that rely on a deterministic latent state vector, PF-RNNs maintain a latent state distribution approximated as a set of particles. To enable effective learning, researchers have introduced a fully differentiable particle filter algorithm that updates the PF-RNN latent state distribution based on Bayes' rule. Experimental results have shown that PF-RNNs can outperform conventional gated RNNs across various domains, including synthetic robot localization datasets and real-world sequence prediction tasks, which is stock price prediction.

Furthermore, recent studies have proposed novel approaches, such as the development of a sentiment-ARMA model, which combines the autoregressive moving average model (ARMA) with sentiment analysis of financial news articles. [3] This model is integrated into an LSTM-based deep neural network consisting of three components: LSTM, VADER model, and a differential privacy (DP) mechanism. The proposed DP-LSTM scheme has demonstrated the potential to reduce prediction errors and enhance model robustness. Extensive experiments conducted on S&P 500 stocks have indicated promising results, including a 0.32% improvement in mean Mean Percentage Absolute Error (MPA) and a significant up to 65.79% reduction in Mean Squared Error (MSE) for the prediction of the market index S&P 500.

7 Conclusion

In conclusion, our project has focused on predicting the short-term movements of blue-chip stocks, a strategy that has proven effective over periods ranging from a few days to a few months. To achieve this, we have leveraged advanced machine learning techniques, employing models: LSTM, GRU, and Transformer.

Additionally, we integrated VADER, a sentiment analysis tool, to incorporate the impact of recent events and news on stock prices. This integration aimed to capture both short- and long-term changes in stock prices influenced by current events. While our results have been promising, there is potential for further improvement. Enhancing the feature set, particularly for the Transformer model, could yield more accurate predictions. Access to historical news articles and professional analysts' reports would also significantly enrich our dataset, potentially leading to more nuanced and informed predictions.

Our prediction results are available on our website, `http://3.34.75.210:8000/`.

References

1. Li, X., Li, Y., Yang, H., Yang, L., Liu, X.Y.: Dp-lstm: Differential privacy-inspired lstm for stock prediction using financial news. arXiv preprint arXiv:1912.10806 (2019)
2. Ma, X., Karkus, P., Hsu, D., Lee, W.S.: Particle filter recurrent neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5101–5108 (2020)
3. Mehtab, S., Sen, J., Dutta, A.: Stock price prediction using machine learning and lstm-based deep learning models. In: Machine Learning and Metaheuristics Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 2. pp. 88–106. Springer (2021)