ISSN 2959-6130

BILSTM Model for Machine Learning Stock Prediction and Portfolio Construction in the Context of Covid-19 Pandemic

Kaiting Li[1, *]

[1] School of Financial Management, Shanghai University of International Business and Economics, Shanghai, China
*Corresponding author: 21049006@suibe.edu.cn

Abstract:

This study used Bidirectional Long Short Term Memory (BiLSTM) neural network model to learn the stock prices of six representative companies from different industries, and evaluated their performance during the period significantly affected by the COVID-19 pandemic from 2019 to 2021. These companies represent industries such as finance, industrial manufacturing, energy, technology, biopharmaceuticals, and consumer goods. A set of characteristic indicators, such as moving averages, momentum indicators, and trading volume based indices, were designed based on historical stock data to provide input data for the BiLSTM model and enhance its ability to capture temporal dynamics of data. The predictive ability of the model is evaluated through performance indicators such as mean square error. This study also applies Monte Carlo simulation to construct investment portfolios based on model predictions, exploring the optimal asset allocation to maximize Sharpe's ratio or minimize volatility at different stages of the COVID-19 crisis. The research results indicate that the BiLSTM model effectively predicts stock price trends, especially when market conditions are stable, by utilizing its ability to simultaneously process past and future input data. Despite facing challenges in predicting accuracy during extreme market fluctuations, such as those experienced in the stage of the outbreak of the pandemic, the overall performance of the model is good. The investment portfolio constructed using these predictions performs better than the S&P 500 index, demonstrating superiority in both return and risk levels, highlighting the effectiveness of combining machine learning technology with financial investment portfolio management strategies.

Keywords: Stocks prediction; Machine learning; Investment portfolio.

1. Introduction

In the financial market, stock price prediction has always been a challenging and highly anticipated research field. Predicting stock prices not only helps investors formulate reasonable investment strategies, but also provides strong support for risk management in financial institutions.

In recent years, with the increase of data volume and the improvement of computing power, the application of machine learning technology in the prediction of financial time series has gradually become a research hotspot. Many researchers are committed to improving stock price prediction methods based on machine learning models. A study used three different machine learning methods to predict the prices, and the final results showed that each neural network model had its advantages and disadvantages under different evaluation criteria [1]. Research has found that a hybrid model combining DBN and LSTM can better explore the complicated patterns and nonlinear features of stock prices, thereby improving the prediction performance of this model [2]. In accordance with the application of deep reinforcement learning algorithms for portfolio optimization has shown that this method can significantly improve the return on investment portfolio and effectively reduce risk [3].

The LSTM neural network was firstly proposed in 1997, aiming to solve the gradient vanishing and exploding problems of traditional Recurrent Neural Networks when dealing with long-term dependency problems. LSTM can retain important information in long-term sequences and update and forget it when needed by introducing gating mechanisms [4]. In recent years, LSTM neural network methods have been continuously used in financial research. By using the LSTM network for stock price prediction, the results show that LSTM has a significant advantage in capturing the long-term dependence of stock prices [5]. Another study applied LSTM combined with CNN to multivariate financial time series prediction, further validating its effectiveness in processing complex financial data [6].

The BiLSTM is an extension of traditional LSTM, which can simultaneously consider the forward and backward information of time series data, thereby more comprehensively capturing the temporal dynamic characteristics of the data. A study built a price prediction model according to BiLSTM, and the results showed that the model outperformed traditional LSTM models in capturing different fluctuations of financial time series [7]. In addition, BiLSTM has unique advantages in handling the nonlinear and complex dynamics of financial markets, especially when combined with attention mechanisms and convolutional neural networks, the predictive performance of the model is significantly improved [8].

Feature engineering plays a crucial role in price prediction issue. Through selecting and constructing appropriate features, the performance of prediction models can be significantly improved. A study proposes a three-stage feature engineering method that combines genetic algorithms and extreme gradient boosting algorithms. This method significantly enhances the prediction correctness of the model during the feature construction stage through data preprocessing and feature selection [9]. Another study validated the effectiveness of feature indicators in practical applications. This study utilizes genetic algorithms to optimize feature selection and combines machine learning regression models for stock price prediction. The results show that using optimized feature sets, the predictive accuracy of this model is significantly better than that of the model without feature selection [10]. This indicates that effective feature selection and construction can significantly enhance the performance of machine learning methods in stock price prediction.

The Monte Carlo method originated in the 1940s and was proposed by scientists such as Stanislaw Ulam and John von Neumann in solving complex physical problems. This method estimates the solution of mathematical problems through a large number of random sampling and is widely used in fields such as numerical integration, optimization, and financial engineering. A study applied Monte Carlo simulation method for risk assessment of investment portfolios, and the results showed that the method can effectively identify and quantify investment risks [11].

This study used a BiLSTM model to predict the stock prices of six representative companies from different industries. By introducing some reasonable feature indicators, the predictive ability of the model is enhanced. Monte Carlo simulation is also applied to construct investment portfolios based on model predictions, exploring the optimal asset allocation strategy at different stages of the COVID-19 crisis by optimizing the Sharpe ratio or minimizing volatility.

2. Method

2.1 Data

This study first selected six different representative industries, and then selected one representative company from each of the six industries, namely JPMorgan Chase, from the financial industry. Caterpillar Corporation, from the industrial and manufacturing industries. BP, from the energy industry. Apple Inc., from the technology industry. Pfizer, from the biopharmaceutical industry. Coca Cola Company, from the consumer goods industry. This study obtained the open, high, low, close, and volume data of these companies from January 1, 2010 to December 31, 2021 through Yahoo Finance.

2.2 Indicator Construction and Feature Engineering

This study establishes an indicator system for stocks by analyzing the factors influencing stock prices. Calculate the values of the corresponding indicators by downloading the data and input them as features into the neural network. The specific indicators include: Simple Moving Average, Exponential Moving Average, Rate of Change, Moving Average Convergence Divergence, Stochastic oscillator, Relative Strength Index, Williams% R, Accumulation/Allocation Line, On-Balance Volume, Money Flow Index, Average Directionality Index, Positive Directional Indicator, Negative Directional Indicator, Bollinger Bands, Price Lag Term(including Lag 1, Lag 2, Lag 3), Commodity Channel Index, Average True Range, Detrended Price Oscillator, Ulcer Index, Vortex Indicator, Relative Vigor Index

In addition, considering the differences between different data, this article adopts the Z-score method to standardize the data. The specific formula is:

$$X^* = \frac{(X - \bar{X})}{\sigma} \tag{1}$$

Among them, X is the true value of original data, $\bar{X}$ is the average value of original data, σ is the standard deviation of original data, and X^* is the normalized value of original data.

2.3 Machine Learning Methods and Model Construction

The BiLSTM neural network model belongs to a kind of recursive neural network, that has the ability to process both historical and future information simultaneously, making it perform well in predicting current data points. Unlike traditional LSTM models that can only process input data in one direction, BiLSTM effectively analyzes data from two directions through its bidirectional structure, significantly improving the processing efficiency and predictive performance of the model. The BiLSTM neural network structure model is composed of two independent LSTM layers: one is responsible for forward processing of input data, and the other is responsible for backward processing. The outputs of these two layers are then merged to form the final model output. This structure helps BiLSTM models improve accuracy in multiple tasks by capturing comprehensive features of data.

In this study, the structure of the BiLSTM neural network model used is as follows:

Input layer: Accepts and processes time series data containing multiple technical indicators.

The first BiLSTM layer: This layer has 60 neurons, using ReLU as activation function to enhance nonlinear processing ability of neural network. This layer outputs the results of the entire time series so that all information can be used for subsequent processing.

The second BiLSTM layer: This layer also has 60 neurons, using ReLU as activation function, but only returns the output of the last time step of the sequence.

Dropout layer: A dropout rate of 30% is set to reduce overfitting of the model during training, achieved by randomly closing 30% of the activation units in the process of training.

Fully connected layer: The layer containing 100 neurons is located after the Dropout layer and also uses the ReLU activation function.

The model uses Adam optimizer with a learning rate of 0.001. Using mean square error as the loss function. Batch size for model training is set to 64, and Epochs is set to 100.

In terms of dataset usage, the data from 2010 to 2018 is used as the training and validation sets, divided in a 9:1 ratio, while the data from 2019 to 2021 is used as the testing part. This division of training, validation and testing helps evaluate the accuracy of the model on untrained data.

2.4 Model Evaluation Indicators

The main evaluation indicators used in this article include mean square error, root mean square error, and mean absolute error.

Mean squared error (MSE): The average square of the difference between predicted results and actual data. MSE provides an indicator for quantifying the size of errors, smaller values indicate higher accuracy of model predictions. The indicator is very useful because it squares the error and emphasizes the impact of large errors. The calculation formula is:

$$MSE = \frac{\sum_{i=1}^{n}(a_i - \widehat{a_i})^2}{b} \tag{2}$$

Root Mean Square Error (RMSE): The square root of MSE, providing an error evaluation in the same units as the original data. Similar to MSE, the smaller the RMSE value. A small deviation between the predicted results of the model and the actual data means a high accuracy of the prediction. This calculation formula is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \widehat{a_i})^2}{b}} \tag{3}$$

Mean Absolute Error (MAE): The average absolute value of the difference between the predicted value and the actual value. MAE provides an intuitive measure of error, as it does not give greater weight to larger errors like MSE. Smaller MAE values indicate higher accuracy of prediction results. The calculation formula is:

$$MAE = \frac{\sum_{i=1}^{n}| a_i - \widehat{a_i} |}{b} \tag{4}$$

These statistical indicators together provide a comprehensive perspective for us to estimate the predictive performance of the model, and each measurement method reflects the correctness and reliability of this model in prediction from different perspectives.

Among them, b indicates the total number of test samples, a_i shows the true value of the i-th test sample, and $\widehat{a_i}$ is the predicted value of the i-th sample

2.5 Construction of Investment Portfolio

In the study, Monte Carlo simulation is used to construct investment portfolios, randomly generating 10000 different portfolio configurations. Through this method, a large number of possible asset allocation portfolios can be covered from a statistical perspective, ensuring the comprehensiveness of the analysis. The asset weights of each investment portfolio are randomly assigned, simulating diversity and uncertainty in the market.

In order to more accurately evaluate the performance of investment strategies under different market conditions, the data is divided into three time periods based on years, which reflect the different impact stages of the COVID-19 pandemic on global financial markets. This move aims to analyze the performance and adaptability of investment portfolios built based on BiLSTM neural network prediction data in actual market environments during different periods.

In terms of portfolio selection, this study particularly focuses on two types of investment portfolios: one is the portfolio with the highest Sharpe ratio, that is, the portfolio with the highest expected return per unit risk; Another type is the portfolio with the lowest risk, which is a combination that minimizes risk while maintaining a certain level of return. Through these two strategies, the investment effects under different risk preferences can be compared. In addition, the risk free rate is calculated using the annual yield of treasury bonds for each year.

Finally, the cumulative returns of these investment portfolios were calculated and compared with the cumulative returns obtained by directly investing in the S&P 500 index. This comparison can help evaluate the performance of custom investment portfolios relative to traditional market benchmarks, thereby verifying the effectiveness and rationality of the constructed models and strategies. This analysis not only demonstrates the absolute performance of the investment portfolio, but also provides a test of strategy robustness in different market environments.

3. Results

3.1 Results of Prediction

Fig.1 shows the predicted results of BiLSTM model of six stocks from 2019 to 2021, and Table 1 shows the testing error.

Fig. 1 six prediction results of JPM, CAT, BP, AAPL, PFE, KO

Table 1. Testing error of six stocks

	JPM	CAT	BP	AAPL	PFE	KO
MSE	9.8980	16.6993	1.3231	46.4191	1.4439	0.8606
RMSE	3.1461	4.0865	1.1503	6.8132	1.2016	0.9277
MAE	2.5568	3.3640	0.9332	5.4008	0.8470	0.7271

From the stock price prediction curves and testing error of six stocks, it can be observed that the BiLSTM model has satisfactory predictive performance for various types of stocks. These stocks range from technology stocks to consumer stocks, from energy stocks to industrial stocks, covering multiple important sectors of the market. Despite facing extreme conditions such as the epidemic, the model can still track the trend of stock price changes well, indicating that the model has good performance in stock price prediction and a certain degree of generalization ability.

In addition, these prediction results are of great significance for the formulation of investment strategies. The improvement of stock price prediction accuracy can promote wiser investment strategies, especially when the market is unstable or there are expectations of significant changes. Investors can use this information to optimize their investment portfolio, reduce risks, and seek maximum returns.

3.2 Results of Portfolio

According to the results of predictions, calculate the weights of six stocks corresponding to maximizing the Sharpe ratio and minimizing risk using a Monte Carlo model. The obtained results are shown in Fig. 2 and Table 3. Then bring predicted weight back to original data to test the portfolio performance that can be obtained with that weight. And compare with the results of direct investment in the S&P 500. These results are listed in Table 4.

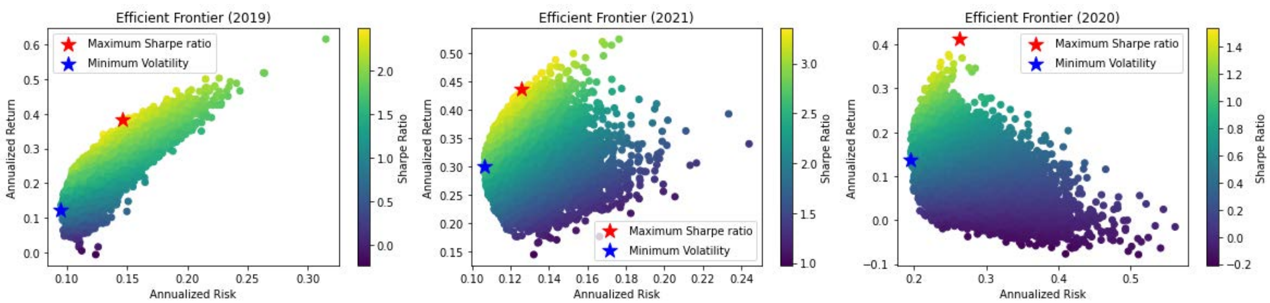

Fig. 2 three portfolio

Table 3. Weights of six stocks

	JPM	CAT	BP	AAPL	PFE	KO
Maximum Sharpe Ratio(2019)	0.4920	0.0276	0.0024	0.2331	0.0376	0.2073
Minimum Volatility(2019)	0.1692	0.0010	0.2270	0.0172	0.2296	0.3560
Maximum Sharpe Ratio(2020)	0.0798	0.1770	0.0056	0.6073	0.0621	0.0683
Minimum Volatility(2020)	0.0425	0.1202	0.0089	0.1704	0.3152	0.3429
Maximum Sharpe Ratio(2021)	0.2734	0.0195	0.1002	0.2258	0.3716	0.0095
Minimum Volatility(2021)	0.1376	0.0879	0.0764	0.2793	0.1579	0.2610

Table 4. Strategy effectiveness

	Return	Risk	Sharpe Ratio
Maximum Sharpe Ratio (2019)	0.4602	0.1504	2.4725
Minimum Volatility(2019)	0.1544	0.1142	1.1370
S&P 500(2019)	0.2574	0.1252	2.0559
Maximum Sharpe Ratio(2020)	0.5137	0.3928	1.2344
Minimum Volatility(2020)	0.1741	0.3263	0.6302
S&P 500(2020)	0.1957	0.3456	0.5663
Maximum Sharpe Ratio(2021)	0.4693	0.1434	2.6765
Minimum Volatility(2021)	0.3445	0.1250	2.3355
S&P 500(2021)	0.2642	0.1305	2.0245

2019: Market performance before the pandemic

In 2019, the investment portfolio showed excellent risk adjusted return, especially in the maximum Sharpe ratio portfolio, where JPM and AAPL had higher weights, demonstrating their advantages in the stable market stage. Specifically, the return rate of the maximum Sharpe ratio combination is 0.4602, with a risk of 0.1504. The Sharpe ratio is as high as 2.4725, significantly higher than the 2.0559 Sharpe ratio of the S&P 500 index. During this year, the minimum volatility portfolio also showed lower risk (0.1252), despite a lower return rate (0.1544), it remains a reliable choice in investments with lower volatility.

2020: Market turbulence in the early phase of the pandemic

In 2020, with the global outbreak of COVID-19, the market experienced extreme fluctuations. The model has adjusted the investment portfolio to adapt to this high volatility environment. Despite the overall market downturn, the maximum Sharpe ratio portfolio still maintains a high Sharpe ratio of 1.2344, which is significantly superior to the 0.5663 of the S&P 500. Especially with the increasing weight of KO and PFE in the investment portfolio, it shows that these stocks have performed relatively steadily during periods of market turbulence. Although the risk of the minimum volatility portfolio has increased to 0.3263, it can still provide relatively reliable returns in such an unstable market environment.

2021: Gradually Controlled Epidemic and Economic Recovery

In 2021, with the popularization of vaccines and the gradual opening of the economy, the market began to recover. During this year, the Sharpe ratio of the maximum Sharpe ratio portfolio increased to 2.6765, significantly higher than the 2.0254 of the S&P 500, indicating strong returns from a reasonable investment portfolio during the market recovery period. Meanwhile, both the risk and return of the minimum volatility portfolio have increased, and slightly better than the S&P 500 index, indicating that conservative investment strategies can also bring stable growth as the market gradually stabilizes.

4. Discussion

This study shows that although the model performs well for most of the time, it can still perform well under certain extreme market conditions. For example, during the sharp decline in the early stages of the epidemic, there was a certain deviation between predictions and reality. This may be due to the lack of similar extreme events in historical data, resulting in the model not fully capturing the characteristics of such sudden market fluctuations.

The analysis period spanned the COVID-19, which had a significant impact on the stock market. Especially for stocks such as PFE that are directly related to the epidemic, the model shows higher predictive accuracy during the epidemic, possibly because the market's response to such stocks is more consistent and obvious. For other stocks such as BP, which are greatly affected by macroeconomic and oil prices, the model faces certain challenges in capturing price fluctuations caused by unexpected events.

The impact of unexpected events such as the epidemic on the market is multidimensional, not only including an increase in volatility, but also the differentiation of industry and individual stock performance. For example, technology stocks have shown strong growth momentum during the pandemic, while some traditional industries are facing challenges.

When implementing predictions, the bidirectional structure of BiLSTM helps the model learn information from past and future data, which traditional unidirectional LSTM cannot achieve. This structure is particularly suitable for predicting stock prices in financial markets, as it can explore the dynamic characteristics of the time series data more comprehensively.

Machine learning and Monte Carlo simulation methods can effectively adapt to different stages of the market, optimizing returns and risks by dynamically adjusting the allocation of investment portfolios. In three different years, the method not only achieved high returns during periods of market stability, but also remained robust during periods of extreme market volatility.

In addition, the research also reveals the potential application of machine learning in financial investment. Through real-time data analysis and prediction, machine learning models can continuously learn about market changes, adjust strategies in a timely manner, and find the optimal solution in a constantly changing market environment.

5. Conclusion

This study used the BiLSTM model to learn the stock prices of six representative companies from different industries from 2010 to 2018, and then predicted the stock data from 2019 to 2021 affected by the epidemic. The results indicate that the neural network model can effectively predict the price trends of various stocks, especially during periods of relatively stable price changes. Although the predicted performance has declined during extreme market fluctuations such as the early outbreak of the epidemic, the overall behavior of this model is still satisfactory. On account of the predicted results, this study used Monte Carlo simulation to construct an investment portfolio model with maximum Sharpe ratio and minimum risk, and compared it with the S&P 500 index during the same period. The results showed that the investment portfolio constructed in view of the prediction results of BiLSTM model outperformed S&P 500 index, achieving higher returns or lower risk levels. This indicates that the investment method that combines machine learning methods and portfolio models is effective.

References

[1] Sheth, D., Shah, M. Predicting stock market using machine learning: best and accurate way to know future stock prices. *Int J Syst Assur Eng Manag* 14, 1–18 (2023).

[2] Zhang, X., Gu, N., Chang, J., & Ye, H. (2021). Predicting stock price movement using a DBN-RNN. *Applied Artificial Intelligence, 35*(12), 876–892.

[3] Chen, Wei, et al. "Mean–variance portfolio optimization using machine learning-based stock price prediction." *Applied Soft Computing* 100 (2021): 106943.

[4] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

[5] Moghar, Adil, and Mhamed Hamiche. "Stock market prediction using LSTM recurrent neural network." *Procedia computer science* 170 (2020): 1168-1173.

[6] Widiputra, Harya, Adele Mailangkay, and Elliana Gautama. "Multivariate cnn-lstm model for multiple parallel financial time-series prediction." *Complexity* 2021 (2021): 1-14.

[7] Yang, Mo, and Jing Wang. "Adaptability of financial time series prediction based on BiLSTM." *Procedia Computer Science* 199 (2022): 18-25.

[8] Zhang, Jilin, Lishi Ye, and Yongzeng Lai. "Stock price prediction using CNN-BiLSTM-Attention model." *Mathematics* 11.9 (2023): 1985.

[9] Yun, Kyung Keun, Sang Won Yoon, and Daehan Won. "Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process." *Expert Systems with Applications* 186 (2021): 115716.

[10] Yun, Kyung Keun, Sang Won Yoon, and Daehan Won. "Interpretable stock price forecasting model using genetic algorithm-machine learning regressions and best feature subset selection." *Expert Systems with Applications* 213 (2023): 118803.

[11] Iqbal, Roihans Muhammad, and Hari Purwanto. "Risk analysis of investment costs in PPP projects using Monte Carlo simulation." *Logic: Jurnal Rancang Bangun dan Teknologi* 22.1 (2022): 13-21.