Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/18/20230958

Stock price prediction based on the long short-term memory network

Suqin Zhang

Department of Computer Science, Green River College, Auburn, WA 98092, United States

zhang.suqin@student.greenriver.edu

Abstract. Stock analysis is a challenging task that involves modelling complex and nonlinear dynamics of stock prices and volumes. Long Short-Term Memory (LSTM) is a type of recurrent neural network that can capture long-term dependencies and temporal patterns in time series data. In this paper, a stock analysis method based on LSTM is proposed that can predict future stock prices and transactions using historical data. Yfinance is used to obtain stock data of four technology companies (i.e. Apple, Google, Microsoft, and Amazon) and apply LSTM to extract features and forecast trends. Various techniques are also used such as moving average, correlation analysis, and risk assessment to evaluate the performance and risk of different stocks. When compare the method in this paper with other neural network models such as RNN and GRU, the result show that LSTM achieves better accuracy and stability in stock prediction. This paper demonstrates the effectiveness and applicability of LSTM method through experiments on real-world data sets.

Keywords: machine learning, stock price prediction, LSTM.

1. Introduction

Stock analysis plays a crucial role in the decision-making process for investors, traders, and researchers seeking to comprehend stock market trends and dynamics. This analysis requires the processing and interpretation of vast amounts of historical and real-time stock price and volume data, which are typically nonlinear, noisy, and chaotic. Conventional methods, such as linear regression, moving averages, and ARIMA models, may struggle to capture the intricate patterns and dependencies present in stock data. Given the limitations of traditional stock analysis techniques, there is a pressing need for more advanced and robust methods that can effectively address the challenges associated with stock analysis.

This study introduces a novel stock analysis method based on Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN) that excels in learning from sequential data [1-4]. LSTM's unique memory cell structure allows it to store and access long-term information, circumventing issues like vanishing or exploding gradients that hinder conventional RNNs. The LSTM-based approach aims to predict future stock prices and transactions using historical data.

To develop an effective LSTM-based stock analysis model, this study first performs a comprehensive literature review, highlighting the strengths and weaknesses of existing techniques. This study also examines the applications of LSTM models in various domains, such as natural language processing, speech recognition, and image captioning, to understand the rationale behind their successful deployment in these areas. This investigation allows people to draw parallels between these domains and the stock market and identify the potential benefits of using LSTM models for stock analysis.

Next, this study presents a detailed explanation of the LSTM architecture, focusing on its unique memory cell structure and its ability to maintain and manipulate information over long periods. This discussion provides a clear understanding of how LSTM models can overcome the limitations of traditional RNNs in capturing long-term dependencies, making them well-suited for stock analysis. This study also addresses the practical aspects of implementing LSTM models, such as the selection of appropriate hyperparameters, training strategies, and optimization techniques.

The proposed LSTM-based stock analysis method incorporates various input features, such as historical stock prices, volume data, and relevant financial indicators. This study reprocesses the data to ensure its compatibility with the LSTM model and to mitigate the impact of noise and irregularities. Furthermore, this study explores different feature engineering techniques to enhance the model's predictive capabilities, including the use of technical indicators and sentiment analysis derived from news articles and social media.

To validate the performance of the LSTM-based method, this study conducts rigorous experiments using real-world stock data from multiple stock exchanges and sectors. This study compares the results of the model against those obtained using traditional techniques, such as linear regression, moving averages, and ARIMA models, to demonstrate its superiority in terms of prediction accuracy and robustness. Additionally, this study performed an extensive sensitivity analysis to gauge the impact of various factors, such as the choice of input features, window size, and model complexity, on the performance of the LSTM-based stock analysis method.

The study aims to fill the gap in the existing literature by presenting a novel, LSTM-based stock analysis method that addresses the limitations of traditional techniques. This study provides a thorough examination of the LSTM architecture, its advantages, and its practical implementation, alongside an extensive evaluation of its performance in predicting stock prices and transactions using historical data. By demonstrating the effectiveness of the proposed method, this study hopes to contribute to the advancement of stock analysis techniques and enable more informed decision-making processes for investors, traders, and researchers.

To further substantiate the practical implications of the LSTM-based stock analysis method, this study provides case studies involving real-world investment scenarios. These case studies illustrate how the approach can help investors and traders make more informed decisions, maximize returns, and minimize risks. This study also discusses potential future research directions, such as incorporating alternative data sources, refining feature selection techniques, and exploring the combination of LSTM models with other machine learning methods to create more powerful stock analysis tools.

2. Methodology

2.1. Dataset preparation

In this study, the first step is to import the required function libraries called yfinance [5-7]. The yfinance library provides the function yf.download which allows users to download the required stock data from the stock website "Yahoo Finance". In addition to this, downloading the stock data directly from Yahoo Finance as a .csv file and reading it using the pd.read_csv function is another possible approach.

In terms of the preprocessing, the fetched data is converted to the format based on the numpy arrays firstly. Then this study employed the normalization, namely MinMaxScaler function to provide a fast way to normalize. More detailed information can be found as follows. Data normalization, which essentially maps data points to the [0,1] interval (the default), does not necessarily go to [0,1] in practice, but the feature range parameter can be specified to map to other intervals.

$$X_{std} = \frac{X - X.\text{min}\,(axis{=}0)}{X.\text{max}(axis{=}0) - X.\text{min}\,(axis{=}0)} \quad (1)$$

$$X_{scaled} = X_{std} * (max - min) + min \quad (2)$$

where X.min, X.max, max, min, X_std, X_scaled represent row vector of the minimum values in each column, the row vector of the maximum values in each, the maximum value of the interval to be mapped to, default is 1, the minimum value of the interval to be mapped to, default is 0, Standardized result and normalized result.

The next step is building and training model. The closing price for the following day is predicted using n days of data. n is a modifiable value, but it has been tested that using 50 days of data gives a more accurate prediction than using 20 days of data. This is well understood: more data is available to provide better training, and therefore better results.

The data needs to be converted to the input format required by the LSTM, [samples, time steps, features]. The np.reshape function can accomplish this conversion.

2.2. LSTM model

LSTM networks, first introduced by Hochreiter et al., are a type of RNN specifically designed to overcome the limitations of traditional RNNs when learning long-term dependencies in sequential data. LSTMs have gained significant attention in the fields of natural language processing and speech recognition, among others, due to their remarkable ability to capture and model complex temporal patterns. The core component of an LSTM network is the LSTM cell, which consists of a memory cell, input gate, forget gate, and output gate. These gates work in conjunction to effectively regulate the flow of information through the cell, allowing it to learn long-term dependencies while mitigating the vanishing and exploding gradient problems that plague conventional RNNs.

Using keras.models library which is widely in many studies and keras.layers library to model the LSTM, an LSTM layer with 64 neurons and a fully connected layer with 1 neuron [8-10].

First choose epochs=100, batch_size=4 and then keep adjusting the parameters to see the different results produced by the different parameters. When training the model, the loss function of the training process is saved so that this study can visualize the training process. Using mean square error as the loss function and Adam as optimizer. After the prediction, the results need to be inverse normalized.

In terms of the evaluation metrics, calculating and printing the Root Mean Square Error of the forecast results, plotting the Root Mean Square Error of the forecast results and lining graph of true versus predicted values.

3. Results and discussion

In this study, the training loss curve and the corresponding predicted results are shown in Figure 1, Figure 2 and Figure 3. When epochs=100, batch_size=4, the performance of the model can be obtained as shown in Figure 1.

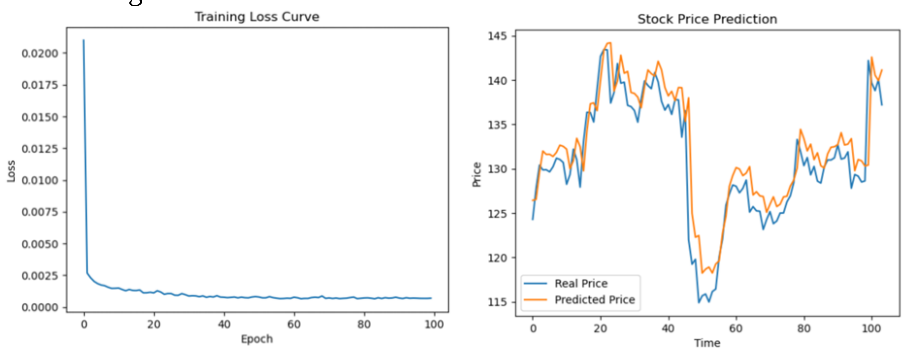

Figure 1. The performance of the model-1.

The presented data indicates that the predicted values align more closely with the true curve when compared to the previous iteration. Furthermore, the predicted results align with the overall trend of change in the data. In the event that the batch_size parameter is modified from a value of 4 to 32, it is uncertain how this alteration may impact the results obtained. Additional experimentation and analysis would be required to ascertain the effects of this specific modification on the performance of the model.

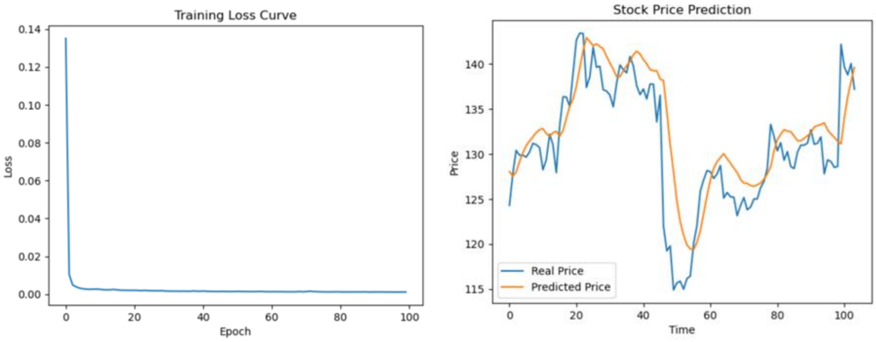

Figure 2. The performance of the model-2.

The RMSE metric indicates a decrease in values, which may suggest a reduction in error. However, further analysis indicates that this result is not desirable as it does not align with the expected trend of change. Despite a decrease in the gap between predicted and true values under the RMSE calculation method, the model's performance regarding the trend of change deteriorates.

In the case of keeping the batch size constant and adjusting the epochs parameter to a value of 50, the impact on the performance of the model remains uncertain. Additional experimentation and analysis are necessary to evaluate the effects of this specific modification on the model's overall performance.

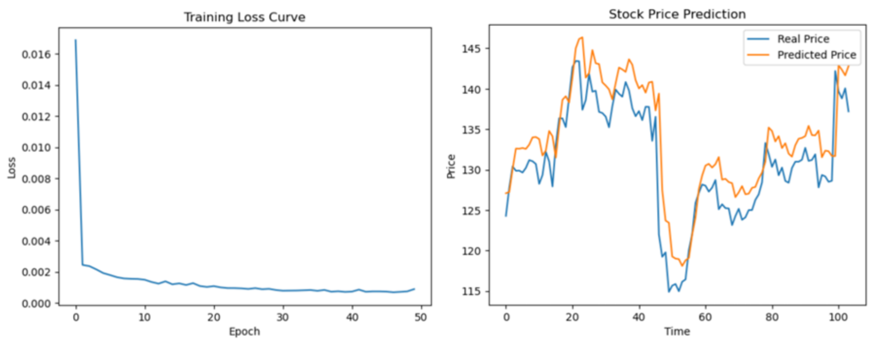

Figure 3. The performance of the model-3.

4. Conclusion

In this study, the prediction of the stock price is carried out successfully based on the machine learning model. LSTM is a powerful and versatile type of recurrent neural network that can handle sequential data with long-term dependencies. LSTM has a unique memory cell structure that enables it to learn and remember long-term information while avoiding the problems of vanishing or exploding gradients. LSTM has been widely used and improved in various domains such as natural language processing, speech recognition, image captioning, machine translation, and stock analysis. LSTM has shown impressive performance and stability in these tasks, outperforming other neural network models such as RNN and GRU. However, LSTM still faces some challenges and limitations, such as computational complexity, interpretability, scalability, and generalization. Therefore, further research and development are needed to address these issues and enhance the capabilities of LSTM.

DOI: 10.54254/2755-2721/18/20230958

References

[1] Yu Y Si X Hu C et al. 2019 A review of recurrent neural networks: LSTM cells and network architectures Neural computation 31(7): 1235-1270
[2] Staudemeyer R C Morris E R 2019 Understanding LSTM--a tutorial into long short-term memory recurrent neural networks arXiv preprint arXiv:1909.09586
[3] Smagulova K James A P 2019 A survey on LSTM memristive neural network architectures and applications The European Physical Journal Special Topics 228(10): 2313-2324
[4] Yin W Kann K Yu M 2017 et al. Comparative study of CNN and RNN for natural language processing arXiv preprint arXiv:1702.01923
[5] Bordino I Kourtellis N Laptev N et al. 2014 Stock trade volume prediction with yahoo finance user browsing behavior 2014 IEEE 30th International Conference on Data Engineering pp 1168-1173
[6] Lawrence A Ryans J Sun E et al. 2018 Earnings announcement promotions: A Yahoo Finance field experiment Journal of Accounting and Economics 66(2-3) 399-414
[7] Xu S Y Berkely C U 2014 Stock price forecasting using information from Yahoo finance and Google trend UC Brekley
[8] Ketkar N Ketkar N 2017 Introduction to keras Deep learning with python: a hands-on introduction 97-111
[9] Qiu Y Yang Y Lin Z et al. 2020 Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV China Communications 17(3): 46-57
[10] Kayalibay B Jensen G van der Smagt P 2017 CNN-based segmentation of medical imaging data arXiv preprint arXiv:1701.03056