# Volatility Prediction of Shanghai Stock Exchange Composite Index Based on LSTM

**Yanjiao Yang**

**Abstract.**Forecasting volatility of stock market has always been the emphasis of research due to the benefits for both investors and the market. Recently, the diversity of data has brought challenges to the prediction of stock volatility with traditional models. Artificial intelligence has gradually been applied to the field of financial risk management. Deep learning methods have shown advantages in stock forecasting compared with traditional methods. This paper uses Shanghai Stock Exchange Composite Index and selects 6 aspects of Baidu Index to represent public focus on economics. The Long Short-Term Memory Model is applied to predict the stock volatility. RMSE and MAPE are selected as loss functions to measure the forecast deviation. With $1.08 \times 10^{-4}$ RMSE and 26.6% MAPE, the Long Short-Term Memory (LSTM) model predicts the volatility effectively and demonstrates great promise of using financial time series to forecast volatility.

**Keywords :**Volatility Prediction, Stock Exchange Composite Index, Baidu Index, LSTM

## 1. Introduction

As the core of the modern market economy, the financial industry has an influence of all walks of life, and it can stimulate the national economy and allocate capital resources efficiently. A lot of investors participate in stock investment which has become quite a common way of financial management nowadays. Generally, changes in stock market reflect not only the performance of enterprises but also the situation of the national economy. As for both the market and investors, the effective analysis of stock market is of great practical significance.

Stock volatility measures the uncertainty of asset returns and reflects the change of financial asset prices.[1-2] It is an important indicator of capital asset pricing, risk management and portfolio in the financial market. Forecasting stock volatility is of guiding significance for investors to avoid risks, which has always been a focus both in academia and industry. In the era of big data, there is a massive growth of financial data, and under many situations traditional economic models fail to deal with large amount of data and to help people make accurate predictions.

Artificial intelligence is the simulation of information on the process of thinking and consciousness.[3] With the development of artificial intelligence, a research boom in machine learning and deep learning has been set off. A series of algorithms such as neural network, perceptron and support vector machine were developed.[4-5] Artificial intelligence can be applied to financial risk control, marketing analysis and quantitative investment. Compared with traditional models, machine learning can better obtain the characteristics of the object and has greater advantages in data processing.[6-7] The stronger ability of nonlinear fitting and generalization of machine learning brings greater possibility to predict stock volatility .[8]Generally, machine learning can not only output better results but also reduce the labor costs to a certain extent.

Investors are important participants in the stock market whose primary goal is to maximize the revenue. Reliable analysis and prediction can help investors make more rational and accurate judgments, reduce their loss and bring higher investment returns. Forecasting the stock volatility also brings benefits for the whole market. By monitoring the smooth operation of stock market to some extents, it can reduce market risks and promote a more healthy and stable stock market. In general, knowing the change of stock volatility and predicting its trends are helpful to both the investors and the financial market.

Traditional volatility forecasting method has been relatively mature. The ARCH model was proposed by Engle in 1982, [9] and later Taylor proposed the stochastic volatility (SV) model .[10]These two types of models have become the focus of modern econometrics research because they can reflect the characteristics of the variance of financial data time series. Bollerslev et al. proposed the GARCH model based on the ARCH model. [11] Hamilton used Hidden Markov Model (HMM)for the prediction of economic cycles in the economic field . [12] With the arrival of big data era, traditional forecasting methods have been challenged by artificial intelligence models. Artificial neural network was applied to financial time series in the 1990s.[13] A forward multi-layer artificial neural network was used by Hammad to forecast stocks, which turned out to have higher accuracy.·[14] Kim directly used support vector machines for stock forecasting, and demonstrated through experiments that this method is more effective than traditional neural network methods.[15] In 1997, Hochreiter and Schimidhuber put forward Long Short-Term Memory (LSTM) model. [16] LSTM and CNN can be applied to predict stock market, and trading strategies can be established according to the predictions.[17]

This paper is structured in five sections. Section 2 includes the input data sources such as the Baidu Index and technical indicators of the stock market, and makes an initial analysis of the input data. In section 3, the LSTM model is introduced including its advantages and basic steps. The results of the model and further discussions are included in section 4. Conclusions of forecasting SSE volatility using LSTM and Baidu Index are made in section 5.

## 2. Data Sources

The input data in this article is divided into two categories. First, Baidu Index is used as an indicator to represent the public focus on certain fields of economics. Six domestic aspects (bankruptcy, credit cards, finance & investing, insurance, jobs and policy) are regarded as the main factors that can reflect the public focus in macroeconomics. Since Baidu Index was developed by Baidu in 2011, the data of Baidu Index range from Jan. 1, 2011 to Dec. 31, 2020. The second category of data is the technical indicators of the stock market. It mainly includes the opening, closing, high and low price of the Shanghai Stock Exchange (SSE) Composite Index on each trading day from January 1, 2011 to December 31, 2020. The SSE Index reflects the liquidity of stock market, and it can be easily accessed and collected. Figure 1 shows the fluctuation of Baidu Index of the 6 aspects from January 1, 2011 to December 31, 2020. There is a steep increase of 'jobs', 'finance & investing' and 'bankruptcy' index during 2012 and 2013, which corresponds with the economic situation in China. In 2012, the economy growth slowed down. In order to prevent the economy growth from declining too quickly, the government increased the infrastructure investments as well as public expenditure and focused on cutting overcapacity and deleveraging.

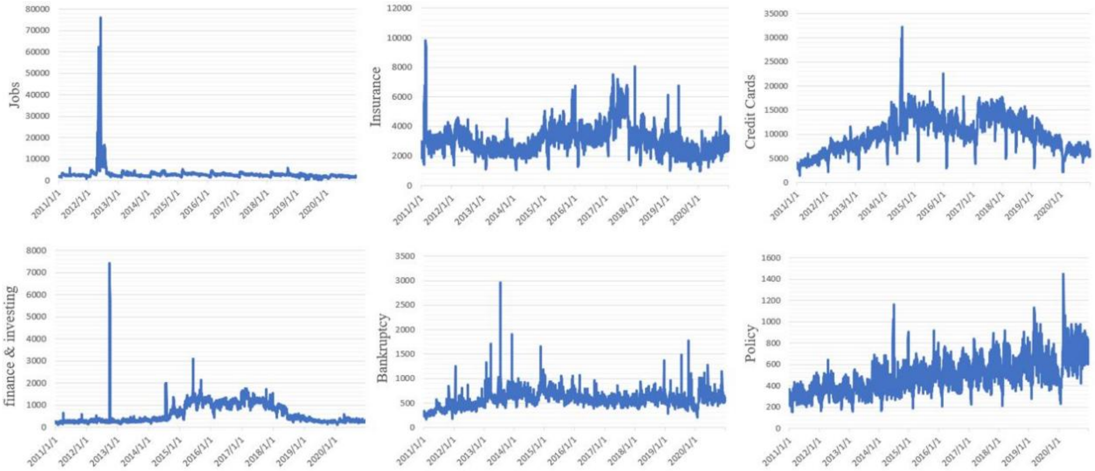

**Figure 1.** Baidu Index of jobs, insurance, credit cards, finance & investing, bankruptcy and policy

Stock volatility usually measures the fluctuation of financial asset prices, describes the changes of market in a certain period and reflects the uncertainty of asset returns. In this work, the estimation of daily volatility $\sigma_t$, which is introduced by Garman and Klass ,[18] uses the method of moments estimators and considers the daily opening, closing, high and low prices.

$$u = \log\left(\frac{Hi_t}{Op_t}\right)$$

$$d = \log\left(\frac{Lo_t}{Op_t}\right)$$

$$c = \log\left(\frac{Cl_t}{Op_t}\right)$$

$$\sigma_t = 0.511(u-d)^2 - 0.019[c(u+d) - 2ud] - 0.383c^2$$

The fluctuation of volatility $\sigma_t$ is shown in Figure 2. The volatility peaked in 2015 when the decline in China's stock market was particularly prominent and triggered the circuit breaker. Generally, the stock was greatly volatile between 2015 and 2016 and remained relatively stable after 2017.

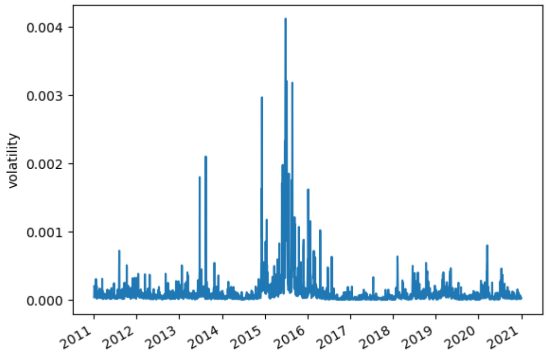

**Figure 2.** daily volatility $\sigma_t$

This paper uses the data from 2011 to 2020 and applies a deep learning model (LSTM). The training set consists of 70% of the whole data. With the timestep set to be 20 and a hidden layer which has 10 outputs, the LSTM model is constructed to forecast the stock volatility in China.

## 3. Models

LSTM is an artificial recurrent neural network (RNN) architecture, which can solve the long-term dependence problem of general RNN. The reason why the long-term dependence exists is that the characteristics of the relatively long-time sequence have been covered after multi-step calculations. All RNNs have a chain form of repeated neural network modules. There is one simple structure in the repeating module of a standard RNN while in LSTM there are four neural network layers interacting in a special way. LSTM introduces a memory unit in each neuron in the hidden layer and uses three gate control units (an input gate, an output gate and a forget gate ) to control the state of the memory unit.

There are several advantages for LSTM to predict the stock volatility. It is easier for neural network models to model complex nonlinear relationships. Due to the influence of many factors, the financial time series exhibit nonlinear characteristics. By using neural network models the accuracy of prediction can be improved to certain extent. Besides, LSTM can solve the problem that traditional BP and RNN cannot solve. Traditional BP neural network ignore the information in the time dimension. There are vanishing gradient problem and gradient explosion problem that bother RNN model. The two key problems in RNN are caused by the cyclic multiplication of the weight matrix, and multiple combinations of the same function can lead to extremely nonlinear performance. Compared with BP and RNN, LSTM has relative insensitivity to gap length, can deal with the vanishing gradient problem and achieve better results in longer sequences.

A common LSTM unit consists of 3 gates: an input gate, an output gate and a forget gate. The forget gate decides how much previous information is discarded. The input gate indicates how much new state is updated to the memory unit while the output gate determines the current output.

The first step in LSTM is to decide the information that is going to be discarded from the cell state, which is achieved by the forget gate. The mathematical expression is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f),$$

where $\mathrm{f_t}$ represents the forget gate, $h_{\mathrm{t-1}}$ indicates the output of previous unit state which is joined together with the current input $\mathrm{x_t}$ to form an input matrix. $\mathrm{W_f}$ is the weight matrix of forget gate, $\mathrm{b_f}$ means the bias and $\sigma$ denotes the sigmoid function.

The next step is to determine the information that is going to be stored in the cell state, which can be realized through a tanh layer and an input gate. The input gate determines how much information should be updated. The tanh layer creates a vector of new candidate values $\widetilde{\mathrm{C}}_\mathrm{t}$ that can be added to the state. The mathematical expression is as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\widetilde{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

The update of the old cell state is showed as follows:

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t}$$

where $\mathrm{C_{t-1}}$ is the old cell state and $\mathrm{C_t}$ is the new state.

Finally, the output information is determined based on the current state and the output gate.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \times \tanh{(C_t)}$$

where $\mathrm{W_o}$ is the weight matrix of the output gate, $\mathrm{b_o}$ is the bias of the output gate and $h_\mathrm{t}$ represents the output value of the cell unit.

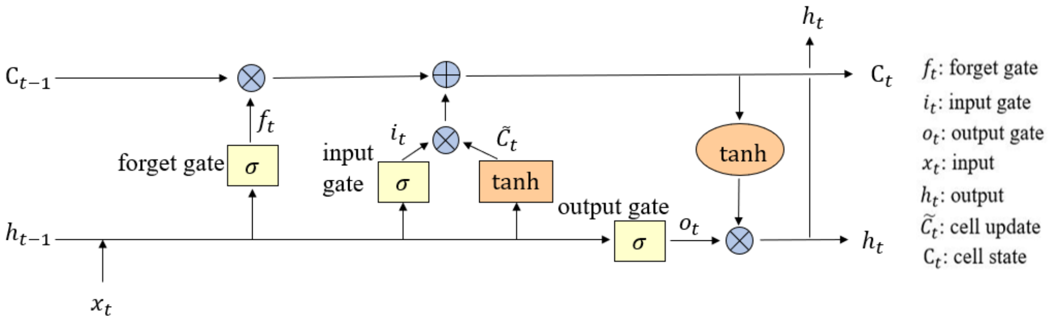

**Figure 2.** Structure of LSTM

## 4. Results and Discussion

Based on the LSTM model, the data are divided into a training set and a testing set. 70% of the obtained data is used as the training set while 30% is used as the test set. The training set ranges from Jan. 4, 2011 to Sep. 12, 2017, and the test set ranges from Sep. 13, 2017 to Dec. 31, 2020. In LSTM model, there are several key parameters that influences the prediction of volatility including the timestep, the hidden layer and the hidden size. To determine the timestep is to decide how much historical data is applied to forecast volatility of the following trading day. The timestep of this paper is set to be 10. Since too many hidden layers can cause the problem of over-fitting and increase the difficulty of training, here the LSTM model selects 3 hidden layers with 1 output. The training steps of LSTM model are as follows: The first step is to recombine the volatility sequence according to the timestep and obtain the target sequence. Then the sample interval is divided due to the target sequence. The training set trains parameters of the model, and the remaining samples form the validation set to evaluate the training effect.

After the volatility is predicted, the deviation of forecast can be obtained. However, there is no universally accepted loss function to measure the forecast deviation. This paper selects Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) as loss functions to evaluate the deviation comprehensively. The equation of loss functions are as follows:

$$\mathrm{RMSE} = \sqrt{M^{-1} \sum_{m=H+1}^{H+M} (\sigma_m - \hat{h}_m)^2}$$

$$\mathrm{MAPE} = M^{-1} \sum_{m=H+1}^{H+M} \left| \frac{\sigma_m - \hat{h}_m}{\sigma_m} \right|$$

Figure 4 compares the prediction with the observed volatility. The LSTM model is trained using Adam method. The result is achieved with 32 examples in a batch and 100 epochs. The validation set contained 20% of the training data.

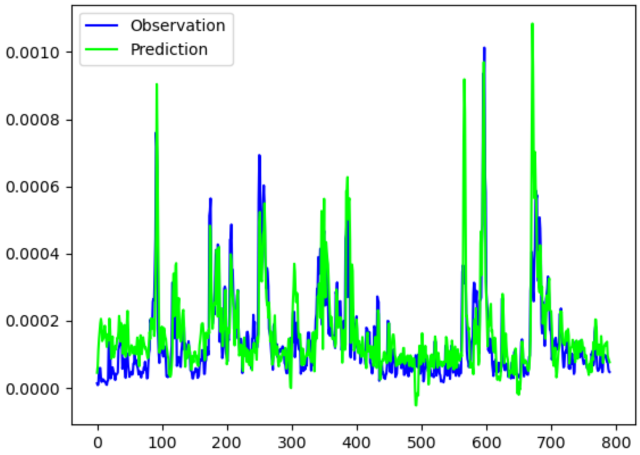

**Figure 3.** Prediction and observation of volatility

With all the features calculated after 100 epochs, the MAPE of test set is 34.04% while the result of train set is 26.6%. As a common measurement of the forecast deviation, MAPE represents the average of percentage error. According to the trend of MAPE in Figure 5, the average of percentage error keeps decreasing as the number of features increases. Since MAPE stabilizes around 25% after 8 iterations, the parameter of iteration is set to be 8.

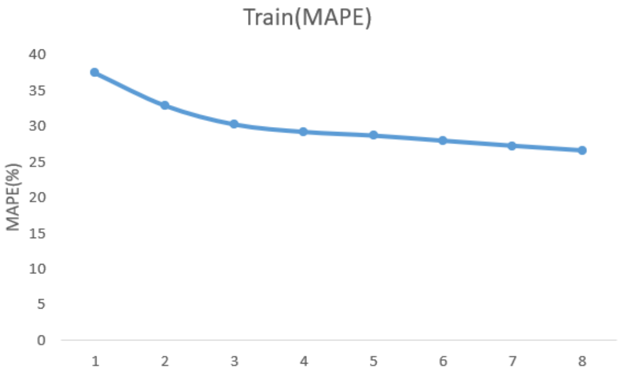

**Figure 4.** Tendency of MAPE

Table 1 shows the results of loss functions (RMSE and MAPE). With $1.08 \times 10^{-4}$ RMSE and 26.6% MAPE. It can be concluded that the LSTM model is accurate and robust. The predicted volatility through the model can reflect the real value of volatility to certain extent. Compared the real results with the predicted results, LSTM is capable of describing the peaks and valleys.

**Table 1.** Loss function RMSE and MAPE

| Model | RMSE | MAPE |
|---|---|---|
| LSTM | $1.11\times 10^{-4}$ | 26.6% |

## 5. Conclusions

In this work, 6 aspects of Baidu Index are regarded as the representative of public focus in macroeconomics. The high, low, opening and closing price of SSE range from Jan.1, 2011 to Dec. 31, 2020, reflecting the situation of stock market. The SSE Index and Baidu Index are used as input data for LSTM model to forecast the volatility. The model selects a hidden layer with 10 outputs, and is trained on 70% of the data set. RMSE and MAPE are chosen as the loss functions to measure the forecast deviation. According to the results, LSTM model is effective in prediction. It makes use of the nonlinearity of data and demonstrates great promise of using financial time series to forecast volatility.

Further improvement of this work includes several aspects: (1) Considering more fields of Baidu Index. More domestic trends indicate more information of the public focus and brings more useful data that can be input into the LSTM model. (2) Using higher frequency of SSE. This paper selects daily high, low, opening and closing price of SSE as input data. With higher frequency of SSE, the situation of stock market can be better represented. (3) Applying more models to predict the stock volatility and comparing the RMSE and MAPE to measure the deviation of each model. Different models have respective advantages when forecasting the volatility. By comparing the loss function, the most robust and accurate model can be obtained.

## References

[1] A A M F, B M I, A E K. On forecasting daily stock volatility: The role of intraday information and market conditions[J]. International Journal of Forecasting, 2009, 25( 2):259-281.

[2] Brailsford T J, Faff R W. An evaluation of volatility forecasting techniques[J]. Journal of Banking & Finance, 1996, 20(3):419-438.

[3] Ertel W. Introduction to artificial intelligence[M]. Springer, 2018.

[4] Schmidhuber J. Deep Learning in neural networks: An overview[J]. Neural Networks, 2015.

[5] Huang W, Nakamori Y, Wang S Y. Forecasting stock market movement direction with support vector machine[J]. Computers & Operations Research, 2005, 32(10):2513-2522.

[6] Gencay R. Non-linear Prediction of Security Returns with Moving Average Rules[J]. Journal of Forecasting, 1996, 15(3):165-174.

[7] Alhaj A M, Terada H. Parallel Implementations of Back Propagation Networks on a Dynamic Data-Driven Multiprocessor[J]. IEICE transactions on information and systems, 1994, 77(5):579-588.

[8] In Ce H, Trafalis T B. Short term forecasting with support vector machines and application to stock price prediction[J]. International Journal of General Systems, 2008, 37(6):677-687.

[9] Engle R F. Autoregressive conditional heteroskedasticity with estimates of the variance for U.K. information [J]. Econometrica, 1982, 50(3): 987-1007.

[10] Taylor S J. Modelling stochastic volatility [J]. Mathematica Finance,1994, 4(2) :183-204.

[11] Bollerslev T. Generalized autoregressive conditional heteroscedasticity[J]. Journal of Econometrics, 1986, 31(2): 307-327

[12] Hamilton J D . A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle[J]. Econometrica, 1989, 57(2):357-384.

[13] Varfis A, Versino C. Univariate economic time series forecasting by connectionist methods[C]. IEEE ICNN 1990: 342-345.

[14] Hammad A, Hall A E L . Forecasting the Jordanian Stock Prices Using Artificial Neural Networks[M]. 2007.

[15] Kim K J. Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1/2):307-319.

[16] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[17] Stoean C, Paja W, Stoean R, et al. Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations[J]. PLoS ONE, 2019, 14(10): e0223593.

[18] Mark B. Garman, Michael J. Klass. On the Estimation of Security Price Volatilities from Historical Data[J]. The Journal of Business,1980,53(1).