ISSN: 2321-2152

IJMECE

International Journal of modern
electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

DEEP LEARNING BASED STOCK PRICE PREDICTION MODEL USING SENTIMENT ANALYSIS

[1]A Vasavi Sujatha, [2]Adithi Chittupolu, [3]Anooja Vuppala,[4] Asma Karim

[1]Assistant professor in Department of Information Technology Bhoj Reddy Engineering College for Women

[2,3,4] UG Scholars in Department of Information Technology Bhoj Reddy Engineering College for Women

Abstract:

Reducing investing risks and increasing rewards are possible with an accurate stock price prediction. This work builds the MS-SSA-LSTM model by combining the multi-source data influencing stock prices and using deep learning, swarm intelligence algorithm, and sentiment analysis. In order to create a distinct sentiment dictionary and determine the sentiment index, we first crawl the posts on the East Money forum. Following that, the Long and Short-Term Memory network (LSTM) hyperparameters are optimised by the Sparrow Search Algorithm (SSA).Lastly, LSTM is utilised to predict future stock prices by integrating the sentiment index with fundamental trading data. Tests show that the MS-SSA-LSTM model performs better than the others and has a high degree of generalizability. The average improvement in R 2 between MS-SSA-LSTM and regular LSTM is 10.74%. We discovered that: 1) Including the sentiment index can improve the forecasting capabilities of the model. 2) SSA is used to optimise the hyperparameters of the LSTM, improving the prediction effect and providing an objective explanation of the model parameter values. 3) Short-term forecasting is better suited to China's financial sector due to its extreme volatility.

Keywords: Sentiment analysis, sentiment dictionary, LSTM model, deep learning, and sparrow search algorithm.

1. Introduction:

Many people decide to enter the financial sector as a result of the development of China's stock market and the quick expansion of online finance. These individuals recognise the significance of investing. For individuals and businesses, accurate stock price prediction can lower investment risks and increase investment returns. Using statistical techniques, early researchers built a linear model to fit the time series trend of stock prices. The conventional techniques include GARCH, ARIMA, ARMA, and so forth. These traditional techniques often only record regular and structured data. But conventional forecasting techniques necessitate unrealistic assumptions.As a result, applying statistical techniques to characterise nonlinear financial data is difficult.Financial time series problems are addressed by multi-layer ANN and artificial neural networks (ANN).However, there is need for improvement in the following areas with the standard neural network method. The ability to generalise is weak, tends to overfit fast, and becomes localised. Improved models are needed to address these issues since a large number of samples must be trained. In order to create the suggested model, we mix the data from multiple sources that influence stock prices and use deep learning, swarm intelligence, and sentiment analysis.

2. Literature Survey:

Using the ARMA model, we examined the existence and alterations in long memory features in the volatility dynamics and returns of the London Stock Exchange and S&P 500. Multifractal analysis has emerged as a significant tool in the recent past for explaining financial market complexity, which is difficult for efficient market theory's linear approaches to adequately capture. The weak version of the efficient market hypothesis in financial markets suggests that price returns are sequences with no correlation to one another. Put differently, pricing need to behave like a random walk. We compare the random walk hypothesis with alternatives that allow for either multifractality or unifractality. According to a number of studies, stock return volatility typically shows clustering, hefty tails, and long-range dependence. Given that self-similar stochastic systems exhibit long-range dependency and heavy tails, it has been proposed that return volatility modelling use self-similar methods to capture these features. Using the ARMA model, the current study forecasts Time Series Stock Returns for the S&P 500 and the London Stock Exchange on a monthly and annual basis. Based on genuine, known data, the ARMA model for the S&P 500 outperforms the London stock exchange, according to statistical analysis of the S&P 500. It can also provide medium- or long-term predictions. The London Stock Exchange's statistical analysis demonstrates that the ARMA model performs better for monthly stock returns than it does for yearly returns. The London Stock Exchange and the S&P 500 are comparable in terms of efficiency and financial stability, even throughout boom and bust cycles.

In order to reduce the risk associated with gold purchases, this study provides an inside look at the implementation of the ARIMA time series model to anticipate the future gold price in Indian browsers based on historical data from November 2003 to January 2014. Thus, to provide guidelines for the investor regarding the best times to buy or sell yellow metal. The Indian economy is being affected by factors such as shifting political landscapes, global cues, and high inflation, among others. As a result, researchers, investors, and speculators are looking for alternative financial instruments to diversify their portfolios and reduce risk. This financial instrument has gained a lot of momentum in recent times. In the past, gold was only bought in India during weddings or other special occasions, but now days, it has become more significant inin the eyes of investors as well, making it vital to forecast the price of gold using an appropriate technique.

Predicting stock prices is a popular endeavour, despite the abundance of forecasting techniques.However, they frequently have issues including poor prediction accuracy, a propensity to dip into a local minimum, and other issues.in an effort to increase stock price forecasting's accuracy.The established modified ARIMA model is based on wavelet analysis of stock price forecasting techniques.The monthly average closing price of the Shanghai composite index was then calculated using this model.and contrasted the forecast outcomes using alternative techniques.The outcomes demonstrate how successful the suggested strategy is.

The GARCH model and several of its variations have been used extensively in practice and in the financial literature. Innovations to GARCH processes are generally considered to be equally and independently distributed, with mean zero and unit variance (strong GARCH) for the sake of quasi maximum likelihood estimation. For the ex ante forecasting of GARCH innovations and consequently, stock returns, higher order dependence patterns may be used under less restrictive assumptions (weak GARCH, lack of unconditional correlation). This research tests the independence of successive GARCH innovations using rolling windows of empirical stock returns. The time variation of serial dependency is shown in rolling -values from independence tests, which are helpful in indicating one-step-ahead directions of stock price fluctuations. Gains from ex ante forecasting are reported for nonparametric innovation predictions, particularly when the sign of the innovation predictors is paired with the sign of linear return forecasts and/or independence diagnostics (-values).

We investigate the applicability of various ARIMA-GARCH models for the purpose of modelling and predicting the conditional mean and volatility of weekly spot prices for crude oil in eleven foreign markets during the period of 1/2/1997–10/3/2009. Specifically, we examine how well four volatility models—GARCH, EGARCH, APARCH, and FIGARCH—performed in out-of-sample forecasting from January to October 2009. The APARCH model performs better than the others in most circumstances, albeit the forecasting results are not entirely consistent. Furthermore, compared to classic conditional variance, conditional standard deviation better represents the volatility in oil returns. Ultimately, shocks to conditional volatility decay exponentially, in line with the covariance-stationary GARCH models rather than the FIGARCH alternative's slow hyperbolic pace.

3.Proposed System:

Our proposed model, named MS-SSA-LSTM, employs sentiment analysis, swarm intelligence algorithms, and deep learning to estimate future stock values by combining data from multiple sources that impact stock prices. In order to create a distinct sentiment dictionary and determine the sentiment index, we first crawl the East Money forum postings. Next, the hyperparameters of the Long and Short-Term Memory networks (LSTM) are optimised by the Sparrow Search Algorithm (SSA). Lastly, LSTM is utilised to predict future stock prices by integrating the sentiment index with fundamental trading data. The model is trained and tested using six sample data sets of individual equities from the Chinese financial market. To compare the performance of the suggested model, we do a number of experiments on various models.

Integrating data from several sources, such as sentiment analysis and fundamental trade data, is the focus of our methodology. By using a comprehensive methodology, the model might potentially increase its forecast accuracy and capture a wider range of market effects.

A more sophisticated knowledge of market sentiment is made possible by our model's use of sentiment analysis from East Money forum discussions. The model can more accurately represent sentiment-driven changes in stock prices by taking investor feelings into account.

Sparrow Search Algorithm (SSA) is our model's effective method for optimising LSTM hyperparameter. Better hyperparameter configurations can be found with the aid of SSA's swarm intelligence technique, which could increase prediction accuracy.

4.Prediction Model:

4.1 Libraries

Tensor Flow

TensorFlow is an open-source software framework that is freely accessible and helps with differentiable programming for a variety of applications, including dataflow. Notable for its adaptability, TensorFlow functions as a symbolic math library and is essential to many machine learning applications, especially those involving neural networks. This all-inclusive tool is used in production at Google as well as for research projects, demonstrating its versatility and robustness.

Numpy

As a popular and adaptable tool for array processing with a wide range of uses, Numpy stands out. This feature-rich package is essential for many computational tasks and plays a key role in scientific computing with Python. Fundamentally, Numpy provides a high-performance multidimensional array object, providing a productive framework for managing intricate data.

Pandas

An impressive tool for high-performance data manipulation and analysis is the open-source Python library known as Pandas. Python's strong data structures, which were first primarily used for data preparation and munging, found limited utility in data analysis. Pandas effectively addressed this gap, revolutionizing the landscape of data processing and analysis. This powerful library facilitates the execution of five fundamental steps in the data processing and analysis pipeline: load, prepare, manipulate, model, and analyze. Regardless of the data's origin, Pandas proves instrumental in seamlessly navigating these steps, providing a comprehensive framework for handling diverse datasets.When combined with Pandas, Python's flexibility finds many uses in both academic and professional realms, including finance, economics, statistics, analytics, and more. The amalgamation of Python with Pandas has considerably enhanced the capacities of data scientists and analysts to address intricate data-related predicaments in diverse sectors.

Matplotlib

One of the most well-known Python 2D plotting libraries, Matplotlib, is known for creating excellent figures that can be published in a variety of hardcopy formats and interactive settings on a wide range of devices. Its adaptability extends to integration into Python scripts, the Python and IPython shells, Jupyter Notebooks, web application servers, and many graphical user interface toolkits. Matplotlib aims to achieve a balance between simplifying simple activities and facilitating the completion of difficult tasks. This flexibility is obvious in its potential to construct a wide array of visualizations, including plots, histograms, power spectra, bar charts, error charts, scatter plots, and more, with little lines of code. The pyplot module within Matplotlib furnishes a MATLAB-like interface for straightforward plotting, especially when coupled with IPython. However, for users seeking more advanced capabilities, Matplotlib provides a comprehensive object-oriented interface. This interface allows for meticulous control over elements such as line styles, font properties, axes attributes, and more. With Matplotlib, users, whether novice or power users, can effortlessly create visually appealing and informative plots, making it a highly versatile tool in the Python data visualization landscape.

Scikit-learn

A wide range of supervised and unsupervised learning techniques are provided by Scikit-learn, which has a single Python interface for easy use. Distributed under a permissive simplified BSD license, Scikit-learn is readily available across numerous Linux distributions, fostering its widespread adoption for both academic and commercial applications. This robust machine learning library empowers users with a versatile toolkit for implementing various algorithms, promoting the development of effective and efficient models across different domains.

4.2 Algorithms

MLP

An example of an artificial neural network is the Multilayer Perceptron (MLP) algorithm, which is made up of several layers of nodes that are feedforward coupled to one another. MLPs are one of the most common and versatile types of neural networks used in machine learning for various tasks, including classification, regression, and time series prediction.

LSTM+GRU

Recurrent neural networks (RNNs) of the LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) algorithms are two varieties that are intended to solve the vanishing gradient issue in conventional RNNs and more accurately capture long-term dependencies in sequential data.

LSTM

A kind of recurrent neural network (RNN) architecture called the LSTM (Long Short-Term Memory) algorithm was created to get around the drawbacks of conventional RNNs in terms of identifying long-term dependencies in sequential data. To put it simply, LSTM is a type of specialised neural network cell that learns and remembers patterns over a range of time steps by storing memory over time.

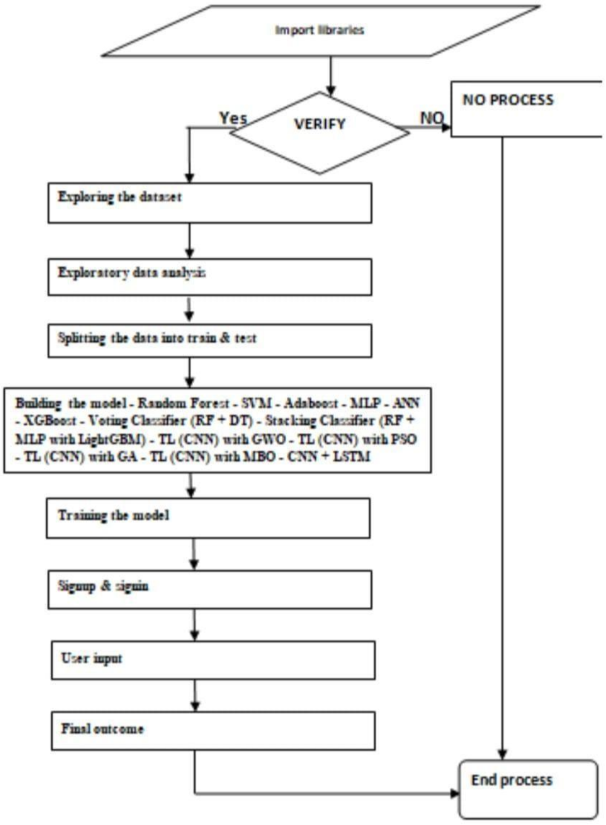
Voting classifier(RF+AdaBoost)

A voting classifier is a machine learning estimator that aggregates the results of each base estimator to make predictions after training a variety of base models or estimators. Voting decisions for each estimator output can be merged to form the aggregating criteria.

Convolutional Neural Network

In the field of machine learning, convolutional neural networks (CNNs) are a particular family of deep learning algorithms designed to handle image recognition and processing tasks with proficiency. A CNN is made up of several layers, including convolutional, pooling, and fully linked layers, among other crucial parts. Because of its complex architecture, CNNs are the go-to option for jobs involving image analysis and classification because they are excellent at extracting features and patterns from visual input.

5. Work Flow Diagram


Import libraries
VERIFY
Yes
NO
NO PROCESS
Exploring the dataset
Exploratory data analysis
Splitting the data into train & test
Building the model - Random Forest - SVM - Adaboost - MLP - ANN - XGBoost - Voting Classifier (RF + DT) - Stacking Classifier (RF + MLP with LightGBM) - TL (CNN) with GWO - TL (CNN) with PSO - TL (CNN) with GA - TL (CNN) with MBO - CNN + LSTM
Training the model
Signup & signin
User input
Final outcome
End process

- **Exploring the dataset**

This module is dedicated to loading data into the system, laying the groundwork for subsequent analyses.

- **Investigative Data Analysis**

An essential phase in the data analysis process is exploratory data analysis, or EDA. It entails the preliminary analysis and visual aid production of the data in order to comprehend the primary features, recognise trends, and ascertain correlations between variables. Data scientists and analysts can create hypotheses, direct more research, and acquire insights from the data with the use of EDA.

- **Train and Split Data**

This pivotal module divides the dataset into training and testing sets, a fundamental step for assessing the performance of predictive models.

- **Building the model**

Various deep learning and machine learning models are created at this critical stage in order to forecast results. Convolutional Neural Network (CNN) + LSTM, Random Forest, AdaBoost, MLP-ANN, XGBoost, Voting Classifier, Stacking Classifier, and Support Vector Machine (SVM) are all part of the ensemble. Every algorithm's correctness is carefully computed.

- **Training the Model**

In this segment, the system focuses on the training of the detection model specifically tailored for dementia prediction. This process involves feeding the model with labeled datasets, enabling it to learn and discern patterns indicative of dementia.

- **User Input**

Here, users input the relevant data required for making predictions within the system.

- **Sign Up and Sign in**

This is the process by which a user creates a new account on a website, app, or platform. Sign in is the process of logging into an existing account on a website, app, or platform.

- **Prediction**

The final and crucial module, where the system processes the user-input data through the generated models, providing a final prediction that is then displayed to the user.

6. Conclusion

Emotions among shareholders impact the stock price, and the LSTM network's hyperparameters are often selected based on subjective experience. As a result, we suggest the MS-SSA-LSTM stock price prediction model. We use a variety of data sources, such as sentiment from shareholders and transaction history. When compared to the use of simply fundamental stock indicators, this improves predictive performance. The manual adjustment of LSTM parameters makes it difficult to provide the best forecasts. The LSTM's hyperparameters are optimised using SSA in order to address this. In the erratic financial market, this approach improves forecasting, flexibility, and understanding of the model. Reduced investor risk, high-precision price prediction, and rapid and precise data understanding are all provided by SSA-optimized LSTM.. Comparative tests conducted on six distinct stocks using a range of models validate the robust accuracy, dependability, and flexibility of the MS-SSA-LSTM in the stock market. Its appropriateness for short-term forecasts in China's unstable financial market is demonstrated by the fact that optimal prediction results are obtained in 5–10 time increments. Sentiment analysis needs to be improved in the future. Different emotional indications, such as grief, fear, rage, and disgust, should be incorporated into our research. In the interim, additional data sources for estimating market mood might be added, such official WeChat and Weibo accounts.

7. References

[1] M. M. Rounaghi and F. N. Zadeh, "Monthly and annual forecasting of time series stock returns using ARMA model: Investigation of market efficiency and financial stability between S&P 500 and London stock exchange," August 2016, pp. 10–21 in Phys. A, Stat. Mech. Appl., vol. 456, doi: 10.1016/j.physa.2016.03.006.

[2] G. Bandyopadhyay, "ARIMA model for gold price forecasting," 2016, pp. 117–121 in J. Adv. Manage. Sci., vol. 4, no. 2; doi: 10.12720/joams.4.2.117-121.

[3] "Analysis and prediction of Shanghai composite index by ARIMA model based on wavelet analysis," by H. Shi, Z. You, and Z. Chen Journal of Mathematical Practice Theory, 44(3), 66–72, 2014.

[4] H. Herwartz, "Empirical assessment of stock return prediction under GARCH," Int. J. Forecasting, vol. 33, no. 3, pp. 569–580, July 2017, doi: 10.1016/j.ijforecast.2017.01.002.

[5] "International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models," by H. Mohammadi and L. Su Energy Economics, September 2010, vol. 32, no. 5, pp. 1001–1008, doi: 10.1016/j.eneco.2010.04.009.

[6] "Recurrent support and relevance vector machines based model with application to forecasting volatility of financial returns," by A. Hossain and M. Nasser 2011, pp. 230–241, J. Intell. Learn. Syst. Appl., vol. 3, no. 4, doi:10.4236/jilsa.2011.34026.