Machine Learning in the Analysis and Forecasting of Financial Time Series

Machine Learning in the Analysis and Forecasting of Financial Time Series

Edited by

Jaydip Sen and Sidra Mehtab

Cambridge
Scholars
Publishing

Machine Learning in the Analysis and Forecasting of Financial
Time Series

Edited by Jaydip Sen and Sidra Mehtab

This book first published 2022

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-8324-4
ISBN (13): 978-1-5275-8324-5

Dedicated to my sister Nabanita who left us on July 27, 2021.
—Jaydip

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The financial sector including financial services, banking, and insurance has witnessed the maximum applications and use cases of machine learning, deep learning, and artificial intelligence. While the financial organizations have only skimmed the surface of the rapidly evolving areas such as deep neural networks and reinforcement learning, the possibility of applying such techniques in many applications in finance and econometrics vastly remains unexplored yet.

The chapters in the volume present several techniques of financial time series analysis and forecasting financial series using statistical, econometric, machine learning, and deep learning approaches. The historical data of the daily and monthly index values of various sectors and important stocks listed in the National Stock Exchange (NSE) of India and the Bombay Stock Exchange (BSE) are used in building predictive models which are later on used to predict the future values of the index and their movement patterns. The time series decomposition results of the financial sectors provide the readers with several useful insights into the behavioral characteristics of different sectors which can prove important for understanding the sectors in a better way. A deeper understanding of the sectoral behavioral patterns will enable the investors to take more effective investment decisions and gain higher profits. The statistical and econometric modeling approaches discussed in the chapters include exponential smoothing methods, Holt and Winter trend and seasonality method, autoregressive (AR) and moving average (MA) method, autoregressive integrated moving average (ARIMA) method, Granger causality analysis, univariate linear regression, multivariate linear regression, and multivariate adaptive regression spline (MARS). The machine learning-based predictive models include several classification and regression approaches including logistic regression, *k*-nearest neighbors, decision trees, bagging, adaptive boosting, extreme gradient boosting, random forest, support vector machine, and artificial neural network. In the deep learning category, the applications of convolutional neural network (CNN) and long-and-short-term memory (LSTM) network architectures are demonstrated in designing predictive models for financial time series data. The use of text mining and natural language processing (NLP) in building precise financial models is also presented in one of the chapters in the volume.

In Chapter 1, *Stock Price Prediction using Deep Learning and Natural Language Processing*, Sen & Mehtab propose several machine learning and deep learning-based models for predicting NIFTY 50 index values and their movement patterns on the National Stock Exchange (NSE) of India. The models are trained on historical NIFTY index values from January 4, 2010, to December 31, 2018. The evaluation of the models is done on test data from January 1, 2019, to December 31, 2019. The authors present an additional text analysis module to improve the accuracy of the predictive models. This is done by incorporating a sentiment analysis module that analyses public sentiment on Twitter on the NIFTY 50 stocks. The output of the sentiment analysis module is used as the second input to the predictive model in addition to the past NIFTY 50 index values. To model the nonlinear relationship between Twitter sentiment and the NIFTY index values at different lags, the authors design a five-layer self-organizing fuzzy neural network (SOFNN) that uses the ellipsoidal function as the basis function. The results indicate that the use of the sentiment analysis module further increases the prediction accuracy of the deep learning-based predictive model.

In Chapter 2, *Machine Learning and Deep Learning in Stock Price Prediction*, Sen presents several machine learning-based models for predicting stock price movement in a short-term time frame using classification techniques. The stock prices are also predicted using a short forecast horizon using several regression models as well. A deep learning-based regression built on long-and-short-term memory (LSTM) network architecture is also presented. The classification and regression models are trained and tested on the historical prices of two important stocks listed on the NSE of India. Extensive results are provided on the performances of the models. The results clearly show that the LSTM is far superior to its machine learning counterparts in predicting future stock prices and their movement patterns for stock price records with rich features and high frequencies.

In Chapter 3, *Stock Price Prediction using Convolutional Neural Networks*, Sen & Mehtab use daily index values of NIFTY 50 from January 29, 2014, to July 31, 2020, for training and testing several machine learning-based classification and regression models. The models are trained on data from December 29, 2014, to December 28, 2018, while evaluation of the models was done on data from December 31, 2018 to July 31, 2020. The classification models are used for predicting the movement patterns of the daily NIFTY index during the test period. On the other hand, the daily NIFTY index values are predicted using the regression models. Four deep learning-based regression models built on the convolutional neural network (CNN) architecture are also proposed. The authors have presented detailed results on the performances of the predictive models. The results have indicated that the CNN models are superior to the machine learning-based regression models in terms of their prediction accuracies.

In Chapter 4, *Robust Predictive Models for the Indian IT Sector using Machine Learning and Deep Learning*, Dutta & Sen present several machine learning models of classification and regression for accurately predicting the future daily index values and their movement patterns of the Indian information technology (IT) sector listed on the NSE. The Indian IT sector daily index values on the NSE from January 2008 to December 2018 are used for training the models. The models are tested on the data from January 2019 to December 2019. The authors present an extensive set of results that indicates the deep learning model built on LSTM architecture is more accurate than its machine learning counterparts in its future prediction of the daily IT index.

In Chapter 5, *A Causality Analysis between the Indian Information Technology Sector Index and the DJIA Index*, Paul & Sen investigate causality between the information technology sector index on the NSE of India and the Dow Jones Industrial Average (DJIA) index of the USA. In their work, the authors use the historical index values from January 5, 2009, to December 27, 2019. The historical index values from January 5, 2009, to December 28, 2018, are used for training the models while testing is done on the data from December 31, 2018, to December 27, 2019. Two regression models are built for predicting the index values of the Indian IT sector on the NSE. The first one is an autoregressive model that uses the lag values of the series as the predictors, while the other one is multivariate. The multivariate model uses the lag values of the DJIA index in addition to the lag values of the Indian IT sector index as the predictors. For both models, the target variable is the Indian IT sector index. It is found that the multivariate model is more accurate indicating a causal effect of the DJIA index series on the Indian IT sector index series.

In Chapter 6, *Stock Price Prediction using Machine Learning and Deep Learning Algorithms and Models*, Mehtab & Sen demonstrate how stock prices and stock returns and their movement patterns can be predicted on a short-term horizon with a fairly high level of accuracy. The authors propose a composite framework consisting of several statistical, econometric, machine learning, and deep learning architectures and models for predicting the future prices and the future price movements of an important stock listed on the NSE of India. The stock price data consist of records collected at five minutes intervals. The records are aggregated into three slots in a day, and the models are used to predict the price and its movement for successive slots. A deep learning regression model is also designed on the LSTM architecture and its prediction accuracies are compared with those of the machine learning models. The results show that while all the models are reasonably accurate in their prediction, the LSTM model outperforms the machine learning model yielding a very high level of precision in forecasting.

In Chapter 7, *Analysis of Different Sectors of the Indian Economy for Robust Portfolio Construction*, Sen analyzes the behavioral patterns of the daily index values of fifteen sectors of the Indian stock market listed on the Bombay Stock Exchange (BSE). The author uses the time series decomposition approach to decompose the daily index series of the sectors from January 2010 to December 2020 into their trend, seasonality, and random components. Based on the decomposition results, the author illustrates how the sectors exhibit a difference in their behavior in trend, seasonality, and randomness, and the way these differences can be utilized by the investors to gain profit from the stock market. Furthermore, four methods of forecasting are proposed for predicting monthly sectoral index, and their prediction accuracies are compared. These forecasting models will help the analysts and investors in predicting the future index values accurately.

While the chapters in this volume do not primarily deal with the basic theories on the topics involved, all the relevant principles and fundamentals are discussed in brief in the chapters for the sake of completeness. Hence, even if some background knowledge of statistics, econometrics, and machine learning may be useful, the book does not presume it for the readers. We believe that the volume can be a valuable resource to anybody interested in gaining knowledge in financial time series analysis. However, the primary target audience for the book is the advanced postgraduate and doctoral students of finance, management, data science, computer science, information technology, and econometrics. In addition, faculty members of graduate schools and universities, practitioners in the industry working in the area of financial analytics, risk management, security analysis, and portfolio management, will also find the subject matters discussed in the book quite useful.

We express our sincere thanks to all the contributors to the chapters in the volume. Without their valuable contributions, it would have been impossible to make this project a success. We also express our sincere thanks to Cambridge Scholars Publishing for providing us with the opportunity to publish our work with their prestigious publishing house. Special thanks are due to Adam Rummens and Amanda Miller of Cambridge Scholars Publishing for their patience, cooperation, and support during different phases of the long publishing process. The members of our respective families and our colleagues have been always the sources of our inspiration, and motivation during such a long, painstaking, and complex scholastic project. Without their support, the publication of this volume would not have been possible. Many thanks to all of them!

Jaydip Sen & Sidra Mehtab
Editors

CHAPTER 1

STOCK PRICE PREDICTION USING DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

JAYDIP SEN & SIDRA MEHTAB

Introduction

Prediction of the future movement of stock prices has been the subject of many studies. On the one hand, there are proponents of the *efficient market hypothesis* (EMH) who claim that stock prices cannot be predicted. Alternatively, some studies have shown that, if correctly modeled, stock prices can be predicted with a reasonable degree of accuracy. The latter focuses on the choice of variables, appropriate functional forms, and forecasting techniques. In this regard, Sen and Datta Chaudhuri proposed a novel approach to stock price forecasting based on a time series decomposition of the stock price time series (Sen & Datta Chaudhuri, 2016a; Sen & Datta Chaudhuri, 2016b; Sen & Datta Chaudhuri, 2016c; Sen & Datta Chaudhuri, 2016d; Sen & Datta Chaudhuri, 2017a; Sen & Datta Chaudhuri, 2017b; Sen & Datta Chaudhuri, 2018; Sen, 2017a; Sen, 2017b). Propositions also exist that use a granular approach to stock price prediction in a short-term time frame using machine learning- and deep learning-based models (Sen & Datta Chaudhuri, 2017c; Sen, 2018).

There are propositions in the literature on the technical analysis of stock prices where the objective is to identify patterns in stock movements and derive profits from them. Various indicators such as *Bollinger Bands*, *moving average convergence divergence* (MACD), *relative strength index* (RSI), *moving average* (MA), *momentum stochastics* (MS), and *meta sine wave* (MSW) have been devised towards this end. For gaining profits, traders extensively use patterns such as *head and shoulders*, *triangle*, *flag*, *Fibonacci fan*, and *Andrew's pitchfork*. These approaches provide the user with visual manifestations of the indicators, which help ordinary investors understand how stock prices may move.

This chapter proposes several machine learning and deep learning-based predictive models for predicting the NIFTY 50 index movement in the National Stock Exchange (NSE) of India. We use the daily stock price values for the period January 4, 2010, to December 31, 2018, as the training dataset for building the models, and apply the models to predict the daily stock price movement and actual closing value of the stock for the period January 1, 2019, to December 31, 2019. We further augment the predictive model by incorporating a sentiment analysis module that analyses public sentiment on Twitter on the NIFTY 50 stocks. The sentiment analysis module's output is used as the second input to the model with the historical NIFTY 50 data to predict future stock price movements. Following the approach proposed by Mittal and Goel, we classify the public sentiment on Twitter into four classes and study the causal effect of these sentiment classes on the NIFTY 50 stock price movement using the Granger Causality Test (Mittal & Goel, 2012).

We organize the chapter as follows. In the section titled *Problem Statement*, we explicitly define the problem at hand. The section titled *Related Work* provides a brief review of related work on stock price movement prediction. In the section titled *Methodology*, we describe our research methodology. Extensive results on the performance of the predictive models are presented in the section titled *Performance Results*. This section also describes all the predictive models built in this study and the results they have produced. Finally, in the section titled *Conclusion*, we conclude the paper.

Problem Statement

Our proposed method is based on collecting the NIIFTY 50 index's historical values from India's NSE and developing a robust forecasting framework for future stock index movement. We contend that index values' daily movement patterns can be learned using powerful machine learning and deep learning-based approaches. The knowledge gained can be applied to predict the future price movements of stocks. In this study, we choose the prediction horizon as one week. We hypothesize that the learning-based approaches can be further augmented by sentiment analysis of social media data on Twitter so that stock price movements can be predicted with even higher accuracy. Here, we do not address the forecasting of the short-term movement of the stock price for intra-day traders. Instead, our analysis is more relevant to long-term investors.

At any point in time in the Indian economy, given the appetite of financial market players, including individuals, domestic institutions, and foreign financial institutions, there is a finite amount of funds deployed in the stock market. This fund is distributed among various stocks. Thus, if some stock prices rise, other stock prices should fall. Using our proposed approach, an investor will be able to predict the movement pattern of the NIFTY 50 index, which depicts the stock market sentiment in India.

Related Work

We can classify the existing propositions in the literature on stock price movements and stock price predictions into four broad categories based on the choice of variables, approaches, and techniques adopted in modeling. The first category includes approaches that use *simple regression techniques* on cross-sectional data (Asghar et al., 2019; Dutta et al., 2012; Roy et al., 2015; Zhong & Enke, 2017; Jaffe et al., 1989). These models do not yield accurate results because stock price movement is a highly nonlinear process. The propositions in the second category exploit time series models and techniques using *statistical and econometric methods* such as *exponential smoothing*, *autoregressive integrated moving average* (ARIMA), *Granger causality test*, *autoregressive distributed lag* (ARDL), and *generalized autoregressive conditional heteroscedasticity* (GARCH) to forecast stock prices (Du, 2012; Khandelwal et al., 2015; Ning et al., 2019; Pai & Lin, 2005; Wang et al., 2020). The third strand includes propositions using *machine learning*, *deep learning*, and *natural language processing* to predict stock returns (Mehtab & Sen, 2020a; Mehtab et al., 2020d; Mehtab & Sen, 2022; Sen & Mehtab, 2021a; Sen & Mehtab, 2021b; Mehtab & Sen, 2022; Sen et al., 2020; Yang et al., 2017; Wu et al., 2012; Ballings et al., 2015; Lv et al., 2019). The propositions of the fourth category are based on *hybrid models* built on machine learning and deep learning with inputs of historical stock prices and sentiments in news articles on the social web (Attigeri et al., 2015; Mittal & Goel, 2012; Medhat et al., 2014; Nam & Seong, 2019).

First, we briefly discuss some of the regression-based propositions existing in the literature.

Zhong and Enke propose methods for the future stock price and price movement prediction using *auto-regressive moving average* (ARMA), *auto-regressive integrated moving average* (ARIMA), *generalized autoregressive conditional heteroscedastic* (GARCH), and *smooth transition autoregressive* (STAR) models (Zhong & Enke, 2017). The authors further present another set of statistical methods involving multiple input variables. The models of this category are *linear discriminant analysis* (LDA), *quadratic discriminant analysis* (QDA), and *multivariate regression analysis*. The authors forecast the daily direction of the S&P500 index return based on sixty predictor variables. Three methods of dimensionality reduction are first applied to the dataset. The methods used for dimensionality reduction are (i) *principal component analysis* (PCA), (ii) *fuzzy robust principal component analysis* (FRPCA), and (iii) *kernel-based principal component analysis* (KPCA). A neural network-based classification model is applied to the transformed data to forecast the direction of movement of future returns. It is observed that the PCA-ANN combination yields the highest level of classification accuracy among all the dimensionality reduction methods.

Dutta et al. propose a logistic regression model using financial ratios as the predictors to predict future stock returns (Dutta et al., 2012). This study attempts to classify the performance of companies as *good* or *bad* based on their one-year performance.

Roy et al. propose a penalty-based regression method – *least absolute shrinkage and selection operator* (LASSO) – to predict future stock prices (Roy et al., 2015). LASSO is an efficient linear regression model that avoids model overfitting and enables feature selection. The regression model is tested on the historical stock prices of the Goldman Sachs Group Inc. The performance of the model is found to be superior to that of another penalty-based regression method, the Ridge regression model.

In the following, we provide a brief description of some of the works in the literature based on econometric methods, such as ARIMA, Granger causality, and GARCH.

Du argues that because stock prices are complex nonlinear functions of many economic factors, it is necessary to design nonlinear models to achieve higher forecasting accuracy (Du, 2018). The author proposes a composite model integrating an ARIMA and a *backpropagation* (BP) *neural network* to forecast the Shanghai Securities Composition stock index. The integrated model's forecasting performance is then compared with those of the individual ARIMA and BP models. It is found that the prediction accuracy of the ARIMA-BP model is superior to that of the BP model, which in turn, is observed to be better than that of the linear ARIMA model.

Wang et al. propose a method of combining asymmetry and extreme volatility in stock prices to forecast volatility in future stock prices (Wang et al., 2020). The authors demonstrate that a shock has a highly significant effect on stock price volatility. It is also shown that volatility is affected by the asymmetry effect both in the long- and short-term. The model, when tested on out-of-sample data, yields an improved accuracy compared to the GARCH-MIDAS model.

Khandelwal et al. present a scheme based on the *discrete wavelet transform* (DWT) for time series forecasting (Khandelwal et al., 2015). The proposed method separates the linear and nonlinear components in a given time series by applying the DWT. After the time series is decomposed, an ARIMA and an *artificial neural network* (ANN) model are separately used on the time series to reconstruct the original time series from its decomposed components. Once the reconstruction is achieved with a sufficiently high accuracy level, the model is tested on four real-world financial time series. The performance of the model on the out-of-sample data is found to be superior to those of the standalone ARIMA, ANN, and the hybrid ARIMA and ANN models proposed by Zhang (Zhang, 2003).

In the following, we discuss the salient features of some machine learning and deep learning-based propositions in the literature for stock price prediction.

Yang et al. present an ensemble of deep learning models for forecasting the Shanghai Composite Index and the SZSE Component Index of the Chinese stock market (Yang et al., 2017). Using the *backpropagation algorithm* to minimize the error and the *Adam* optimizer to optimize the *gradient descent algorithm*, a set of component networks is built. An ensemble of all these component networks is constructed using *bagging* to create a generalized model. The ensemble model is then tested on the out-of-sample data, and trend predictions are computed based on the predicted index values. The accuracy of the trend predictions of the daily *high* and *low* values for the Shanghai Composite Index is 74.15%. The corresponding values for the SZSE Component Index are 73.95% and 72.34%, respectively.

Wu et al. present a novel approach for predicting stock price trends using a combination of *sequential chart patterns*, *l-means*, and the *AprioriAll algorithm* (Wu et al., 2012). The time series of stock prices is divided into a set of subsequences. For each subsequence, appropriate charts are constructed using a *sliding window* method. The charts are then clustered using the *k*-means algorithm (Wu, 2012). After clustering, the charts form the chart pattern sequences, which are mined further to identify frequently occurring patterns using the *AprioriAll* algorithm (Yu & Li, 2018). The existence of such frequent patterns indicates that some market behaviors are exhibited in a correlated fashion. Detection of the correlations of such hidden market behaviors enables accurate prediction of the trend in stock price. The authors validate their propositions by experimental results.

Ballings et al. present a scheme that compares the performances of ensemble methods such as *random forest*, *AdaBoost,* and *kernel factory*, with single classifier models such as *neural networks*, *logistic regression*, *support vector machines*, and *k-nearest neighbors* (Ballings et al., 2015). Using the time series of historical stock prices of 5767 public companies and the *area under the curve* (AUC) for the *receiver operating characteristic curve* (ROC) as the metric, the models' performances are compared over a forecast horizon of one year. The models, listed in the decreasing order of their performance, are *random forest*, *support vector machine*, *kernel factory*, *AdaBoost*, *neural network*, *k-nearest neighbors*, and *logistic regression*.

Mehtab and Sen present a series of works on the design of predictive models for predicting future stock prices and index values and movements using innovative machine learning and deep learning architectures (Mehtab & Sen, 2020a; Mehtab & Sen, 2020b; Mehtab & Sen, 2020c; Mehtab et al., 2020d; Mehtab & Sen, 2021a; Mehtab & Sen, 2021b; Mehtab & Sen, 2022). The authors use daily historical stock prices and stock index values at an interval of 5 minutes. Exploiting the power of *convolutional neural networks* (CNNs) and *long- and short-term memory* (LSTM) networks, the models are found to have achieved a high level of accuracy on the out-of-sample of data. The authors propose four CNN models and six LSTM models with different architectures and input data shapes (i.e., univariate or multivariate time series data). The models are compared on their *root mean square error* (RMSE) values. The results elicit two exciting observations: (i) the performances of the CNN models are superior to those of the LSTM models, and (ii) the models based on the univariate data yield more accurate results than the models using the multivariate data. In another set of works, Mehtab et al. propose further variants of CNN and LSTM-based models for predicting future stock price values and stock price movements (Mehtab et al., 2020d; Mehtab et al., 2021a; Sen & Mehtab, 2021b). The authors report extensive results of the performances of the models.

Finally, we discuss some of the hybrid models that use sentiments in the news and other relevant textual information for stock price prediction in the following.

Bollen et al. contend that emotions profoundly affect an individual's buy or sell decisions (Bollen et al., 2011). The authors propose a mechanism that computes the collective *mood* states of the public from many Twitter feeds and investigates whether the collective moods exhibit any correlation with the Dow Jones Industrial Average (DJIA) index. The public *moods* from the Twitter feed are computed using two mood tracking tools – (i) *OpinionFinder,* which classifies the *opinions* as either positive or negative, and (ii) the *Google Profile of Mood States* (GPOMS) which measures six dimensions of a *mood*. The six dimensions are – *calm*, *alert*, *sure*, *vital*, *kind*, and *happy*. The results show that DJIA index values can be more accurately predicted by including specific public mood dimensions in the model. The model is further augmented using a *Granger causality* analysis and a *self-organizing fuzzy neural network* (SOFNN) so that the public moods can be more efficiently used as a predictor in forecasting the future DJIA index values. An accuracy level of 86% is found in predicting the daily up and down changes in the *DJIA index's close values* using the composite mode with the Granger causality and the SOFNN module.

Checkley et al. study the use of sentiment metrics extracted from microblogs in predicting stock market index values (Checkley et al., 2017). Using the bearish and bullish sentiments from micro-blogging sites at different time intervals, the authors model their forecasting ability in future stock price movements, price volatility, and trade volume. The study reveals a significant causal link between the sentiments on the micro-blogging sites to the price movements, volatility, and the volume traded at short-term time intervals. The study concludes that investors' behavior in stock markets is more like that of a hasty mob than a group of wise people.

Chen et al. study the impact of sentiments from the news sites on the stock market in Taiwan (Chen et al., 2019). After the news articles are collected from the web, they are preprocessed, and the text corpus is transformed into a word feature set using the *Word2Vec* approach (Jatnika et al., 2019). After the *Word2Vec* model is ready, an LSTM-based deep learning model predicts future stock prices using the news articles from the news sites.

Dang and Duong highlight the influence of online news articles on the movement patterns of stock prices (Dang & Duong, 2016). Based on their hypothesis, the authors propose using a time series analysis that integrates a gamut of text mining and text analysis mechanisms. The proposed model is tested on real-world time series of historical stock prices. The model is found to achieve an accuracy of around 73%.

Galvez and Gravano emphasize the impact of online sentiments on the news articles on the stock prices in Argentina (Galvez & Gravano, 2017). The authors investigate two critical issues. The first question that the authors pose is whether there is any useful information in the stock exchange boards for predicting stock returns. The second question is whether the information is novel or is carried in the time series of the historical stock prices. The authors train, validate, and test a series of predictive models using machine learning and topic modeling algorithms in text analysis to address these questions. The models are built using various combinations of features and predictor variables. The study reveals that the information extracted from the online exchange boards complements the information in the time series of the historical stock prices. Hence, the use of information from the online exchange boards improves the performance of the predictive models.

Hu et al. argue that the quality, veracity, and comprehensiveness of online textual information related to the stock market and stock price movement vary widely, and in many cases, they are low-quality news (Hu et al., 2018). To make learning models robust against such chaotic information, the authors propose three principles of learning: (i) *sequential content dependency*, (ii) *diverse influence*, and (iii) *efficient and effective learning*. The proposition implements the first two characteristics of learning by including a *hybrid attention network* (HAN). The HAN attempts to predict the trend of the stock price time series using a sequence of recent and related online news. A controlled learning mechanism implements the third aspect of learning. Simulations show that trading strategies formulated based on the model yield a significantly improved annualized return on investment compared to the returns produced by the models without the text analytics module.

Jeong et al. propose a predictive model for forecasting stock prices and stock price movement that incorporates an *opinion mining* module (Jeong et al., 2018). The model is capable of performing three major tasks: (i) filtering fake information on the web, (ii) credit risk assessment, and (iii) detection of critical signals and using them in stock price prediction. These three operations are executed iteratively by the model on real-world time series of stock prices and news items available on the web. The results show that the proposed model helps an investor make investment decisions and increases the returns of investments.

Li et al. present a novel architecture of a predictive model for forecasting stock prices that uses an LSTM-based module and a text mining module that incorporates the sentiment of the investors and other market factors (Li et al., 2017). The model extracts the sentiment from the online news and opinions using a Naïve Bayes algorithm that filters out irrational and fake opinions. The proposed model is tested on the CSI300 index and produced an improved forecasting accuracy compared to the benchmark models that do not have any text processing components.

Li et al. discuss the limitation of the *bag-of-words* approach of online news processing in stock price prediction (Li et al., 2014). The authors argue that news sentiment should be considered a vital ring in the chain of sequence from the word patterns in the news to the final price movement of stocks. The authors use the Harvard psychological dictionary and the Loughran-McDonald financial sentiment dictionary to construct the vocabulary for the text processing module. After the vocabulary space is built, news articles are assigned quantitative measures and then placed in