SARF: Stock Market Prediction with Sentiment-Augmented Random Forest

Saber Talazadeh[1] and Dragan Perakovi ć[2(✉)]

[1] Computer Department, British Columbia Institute of Technology, Burnaby, Canada
saber_talazadeh@bcit.ca
[2] Faculty of Transport and Traffic Sciences, University of Zagreb, Zagreb, Croatia
dragan.perakovic@fpz.hr

Abstract. Stock trend forecasting, a challenging problem in the financial domain, involves extensive data and related indicators. Relying solely on empirical analysis often yields unsustainable and ineffective results. Machine learning researchers have demonstrated that the application of random forest algorithm can enhance predictions in this context, playing a crucial auxiliary role in forecasting stock trends. This study introduces a new approach to stock market prediction by integrating sentiment analysis using FinGPT generative AI model with the traditional Random Forest model. The proposed technique aims to optimize the accuracy of stock price forecasts by leveraging the nuanced understanding of financial sentiments provided by FinGPT. We present a new methodology called "Sentiment-Augmented Random Forest" (SARF), which incorporates sentiment features into the Random Forest framework. Our experiments demonstrate that SARF outperforms conventional Random Forest and LSTM models with an average accuracy improvement of 9.23% and lower prediction errors in predicting stock market movements.

Keywords: Machine Learning · Large Language Model · Random Forest · Sentiment Analysis · Natural Language Processing · Stock Price Prediction

1 Introduction

Predicting stock trends is a tough task because of the many factors involved. Despite the development of stock predictors based on statistical models, the dynamic, non-linear, and complex nature of the stock market makes effective trend prediction a persistently challenging task [1]. In the field of quantitative finance, the focus is on intelligent timing and stock selection. As quantitative investment and machine learning increasingly converge, understanding the rise and fall of stocks becomes pivotal. Diverse stock price forecasting methods exist, each with its own advantages and drawbacks. Machine learning models, in particular, showcase effectiveness and extensibility by learning relationships between predictor variables and stock movement directions in historical data [2, 3]. Unlike traditional statistics and econometric models, machine learning models demonstrate superior prediction performance and robustness. Researchers have explored various machine

D. Perakovic and L. Knapcikova (Eds.): FABULOUS 2024, LNICST 596, pp. 54–64, 2025.
https://doi.org/10.1007/978-3-031-72393-3_5

learning models, such as support vector machines and random forests, for stock trend prediction. Integrating these models presents challenges, especially in handling time series data, selecting technical indicators, and optimizing parameter combinations [4, 5]. This study contributes by systematically building a stock forecasting model that integrates technical indicators with sentiment analysis throughout the process and incorporating exponential smoothing to reprocess technical indicators. The primary contribution of this research is integration of sentiment analysis through the incorporation of sentiment scores and dynamic weight adjustments in the optimized Random Forest model with data sourced from Yahoo Finance. This integration enhances the model's ability to capture information reflecting stock movement and the impact of market sentiment on stock prices.

To extract textual sentiment information, we employ the FinGPT model, a transfer learning model pre-trained on massive finance textual content. This model demonstrates superior performance in finance sentiment analysis [6]. The study's goal is to evaluate the performance of the optimized random forest in medium- and long-term stock forecasting, aiming to improve overall forecasting accuracy. The paper concludes by comparing the prediction performance of SARF, RF, and LSTM based on relevant metrics.

2 Related Work

Researchers employ various technologies, including statistics and data mining, to classify and predict future stock values. Nti et al. [7] investigates the correlation between different sector stock prices and Macroeconomic Variables, aiming to predict a 30-day ahead stock price using Random Forest and Long Short-Term Memory Recurrent Neural Network. The proposed model, demonstrates high prediction accuracy and superior mean absolute error compared to traditional time-series techniques, highlighting its effectiveness in automatic identification and extraction of MVs influencing diverse sector stocks for accurate future price predictions.

Mehtab et al. [8] explores the challenges in predicting stock price movements, proposing a hybrid approach involving machine learning, deep learning, and natural language processing. The research employing various predictive models, classification techniques, regression models, and a sentiment analysis module on Twitter data to enhance prediction accuracy. Basak et al. [9] addresses the challenge of predicting stock market trends, emphasizing the direction (gains or losses) rather than precise prices. Utilizing random forests and gradient-boosted decision trees, the study introduces an experimental classification framework with selected technical indicators, demonstrating improved accuracy for medium to long-run stock price direction prediction. Random forest is shown to have more advantages than XGBoost overall.

Zhou et al., introduces a learning architecture, combining logistic regression and gradient boosted decision trees for forecasting stock indices. The proposed model, evaluated against various models on emerging and mature stock markets, demonstrates superior performance, offering both statistical and economic advantages for trading strategies, even when considering transaction costs [10]. Oriani et al. [11] assesses the impact of various lagging technical indicators, such as the Exponential Moving Average and Weighted Moving Average, on stock price prediction through artificial neural networks. The findings suggest that incorporating these indicators, both individually and in combination,

enhances the accuracy of stock forecasts, Fundamental and technical indicators, as studied by Beyaz et al., across 140 S&P 500 companies, revealing that models incorporating fundamental indicators outperform those relying on technical indicators. Additionally, utilizing combined indicators proves advantageous in over 95% of cases, yielding lower Root Mean Square Error compared to standalone fundamental or technical indicators [12].

3 Proposed Method

In this section, we propose a method that leverages the Random Forest model to integrate technical indicators with sentiment analysis using FinGPT. We extract sentiment scores (positive, negative, neutral) for each data point. By incorporating sentiment-based features extracted from FinGPT using financial news articles and using the sentiment scores as additional features, we enhance the Random Forest model's capacity to capture and incorporate market sentiment.

3.1 Random Forest Model

We utilized the Random Forest algorithm as the foundational model, leveraging its robustness and capacity to manage relationships within financial data through a robust ensemble learning approach. This algorithm proved effective in handling non-linear relationships and mitigating overfitting challenges present in financial datasets. The model's capability to furnish feature importance scores played important role in discerning the significant factors influencing stock movements [13, 14]. Among the critical hyperparameters applied in our research are number of trees, maximum tree depth, minimum samples required for splitting and the feature subset size. The tuning of these hyperparameters was used for optimizing the predictive performance of the model within the dynamic nature of stock markets [15].

3.2 Sentiment Analysis with FinGPT

In our study, we conducted experiments with FinGPT model for sentiment analysis. FinGPT demonstrated notable strengths, particularly in its capacity to comprehend context, generate coherent responses, and provide diverse financial insights beyond sentiment analysis. Due to its broader focus, adeptness in handling various financial queries, and proficiency in generating responses in natural language, we opted to utilize FinGPT in our research. FinGPT is utilized to perform sentiment analysis on financial news. The model's contextual understanding of financial language provides valuable sentiment scores. It is specifically designed to provide information and answer questions related to finance, banking, and investing. FinGPT uses natural language processing (NLP) technology to understand and respond to user queries, and it has been trained on a large dataset of financial information to ensure that its responses are accurate and relevant.

It also provides a built-in sentiment analysis API that can be used to analyze text data and extract sentiment scores. In this study we used FinGPT APIs to analyze text data and get a sentiment score ranging from -1 (negative) to 1 (positive).

3.3 SARF - Sentiment-Augmented Random Forest

We introduced the SARF model, which combines technical indicators and sentiment features with the Random Forest model. We added sentiment-based features from FinGPT as extra inputs to improve the Random Forest model's ability to grasp market sentiment. This hybrid approach aims to leverage the complementary strengths of both models. SARF builds upon the ensemble learning paradigm of Random Forests by introducing a mechanism to integrate sentiment analysis and technical indicators. The SARF model consists of an ensemble of decision trees, each trained on different subsets of the dataset, but with the inclusion of sentiment-based and technical indicator features.

SARF builds upon the ensemble learning paradigm of Random Forests by introducing a mechanism to integrate sentiment analysis and technical indicators. The SARF model consists of an ensemble of decision trees, each trained on different subsets of the dataset, but with the inclusion of sentiment-based and technical indicator features. We used TA-Lib (Technical Analysis Library) to calculate 15 technical indicators. During feature selection, variables with high correlation are eliminated to addresses multicollinearity issues, and helps prevent overfitting. This process enhances the model's interpretability, improves computational efficiency, and aids in dimensionality reduction, making the model more practical and resource-efficient. Subsequently, the selected indicators serve as input vectors for training the SARF model, and the model's performance is assessed on the test dataset (Fig. 1).

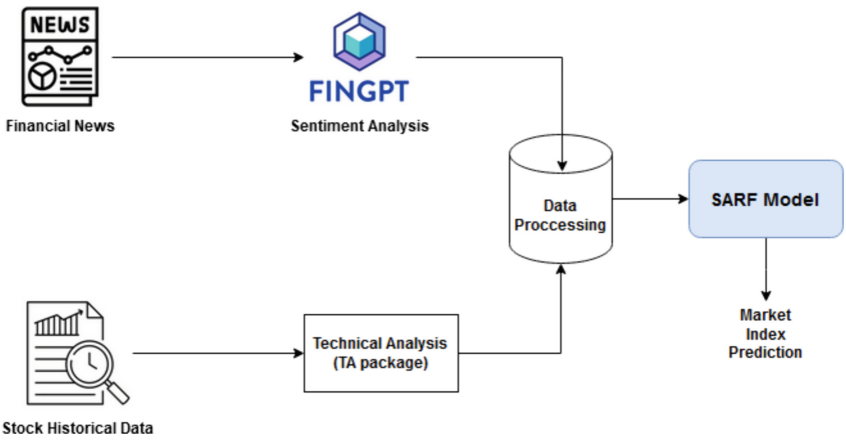

Fig. 1. Data Processing Diagram

4 Experimental Evaluation

We conduct extensive experiments using historical financial data, comparing the predictive performance of SARF against traditional Random Forest models. Evaluation metrics include accuracy, precision, recall, and F1 score. We utilized cross-validation to ensure robustness and minimize overfitting in the proposed model.

4.1 Data Collection

For data collection we used Yahoo Finance using yfinance Python package, which is freely available. Yahoo Finance is a financial news platform that provides various services related to finance, including stock quotes, financial reports, and market news. It offers information on stock market data, technical indicators, and historical prices. This study utilizes data on the price and volume1 of market indices for testing. NASDAQ, S&P 500, and Dow Jones are prominent stock market indices that respectively represent technology stocks, a broad market cross-section, and 30 major industrial companies, reflecting the overall performance of the U.S. stock market.

We chose to use US market indices such as NASDAQ, S&P 500, and Dow Jones for stock market predictions instead of individual stocks because these indices give a broader and more varied view compared to looking at single stock- of a company. They include a mix of different parts of the market, covering various industries and companies, making them a stronger reflection of overall market trends. Predicting the market based on these indices let us consider a wider range of factors, like big economic changes, and reduces the impact of events specific to one company that might affect its stock. Plus, these indices have a stable history, making them helpful for understanding overall stock market performance and trend.

The dataset was collected at daily intervals by querying Yahoo Finance APIs, capturing key metrics such as opening price, lowest price, highest price, closing price, and trading volume. The data spanned from January 2, 2015, to December 30, 2023.

In this study, we leveraged these technical indicators as independent variables to predict future stock market movements. Technical indicators are mathematical calculations derived from historical data, providing insights into trading patterns for financial assets. Throughout the study, we utilized several commonly used indicators, some of which have been previously explored by other researchers.

The learning algorithm used in our paper is random forest. The time series data is acquired, smoothed and technical indicators are extracted as shown in Table 1. Technical indicators are parameters which provide insights to the expected stock price behavior in future. These technical indicators are then used to train the random forest. The time series historical stock data is first exponentially smoothed. Exponential smoothing applies more weightage to the recent observation and exponentially decreasing weights to past observations [17, 18].

4.2 Feature Extraction

In the area of technical analysis, an important parameter is the technical index derived from stock data, serving as a predictor for stock market trends and a tool frequently utilized by investors. These indicators play a significant role in evaluating short-term stock price dynamics and can prove effective for medium- and long-term purposes, such as identifying entry and exit points. To validate the efficacy of the process-optimized stochastic forest model proposed in this study, we employ key technical indicators and sentiment indicators from FinGPT as input features for model training. The output variable is whether the market index moves up or down, predicting the dynamic trend of the

Table 1. Technical Indicators

Indicator Name	Description
Moving Averages (MA)	The average value of a security over a given time. Helps identify trends and potential reversals
Moving Average Convergence Divergence (MACD)	Measures the relationship between two moving averages. Signals trend strength and direction
Relative Strength Index (RSI)	Measures the speed and change of price movements. Indicates overbought or oversold conditions
Stochastic Oscillator	Compares a security's closing price to its price range over a specific period. Shows momentum
Williams %R	Measures overbought or oversold levels. Similar to the stochastic oscillator
Bollinger Bands	Consists of three lines: moving average, upper band, and lower band. Indicates volatility and trends
On-Balance Volume (OBV)	Measures positive and negative volume flow. Helps predict price movements
Accumulation/Distribution Line (ADL)	Tracks buying and selling pressure. Reflects accumulation or distribution of a security
Average True Range (ATR)	Measures market volatility. Indicates potential price movement
Ichimoku Cloud	Provide a comprehensive view of support, resistance, and trends
Parabolic SAR (Stop and Reverse)	Helps identify potential reversal points. Useful for setting stop-loss orders
Fibonacci Retracement	Uses Fibonacci ratios to predict potential retracement levels in price movements
Chaikin Money Flow (CMF)	Combines price and volume data to assess buying and selling pressure
Average Directional Index (ADX)	Measures trend strength. Helps determine whether a security is trending or ranging

stock market index. Initially, various indicators for technical analysis are selected based on the objective of medium- and long-term stock forecasting.

To establish an effective technical data system, we carefully select three indices as the dataset and conduct feature importance analysis by constructing three decision trees. The resulting technical feature importance data aids in the selection of appropriate indicators. The overall dataset of technical indicators is then divided into a training set (two-thirds) and a test set (one-third). From a combination of 15 technical indicators and 4 sentiment

indicators, we derive 14 predictor variables for forecasting. Through exploratory analysis and correlation coefficient calculations, it is evident that some predictor variables show correlations. The precise correlations or highly correlated relationships can lead to multicollinearity, impacting model stability or hindering accurate parameter estimation. To address this, the SARF algorithm is employed, introducing a penalized function to compress less important variables and eliminate multicollinearity.

Correlation analysis is performed to calculate the correlation coefficients between predictor variables, and highly correlated features (typically exceeding 0.8 correlation) are eliminated to prevent redundancy. This strategic pruning ensures our model remains efficient and does not process redundant information [18].

To further mitigate multicollinearity, principal component analysis (PCA) and ridge regression are employed. Ridge regression, a penalized regression method, effectively shrinks the coefficients of highly correlated features towards zero, mitigating the multicollinearity issue and enhancing the stability of the model [19].

4.3 Parameter Optimization

In our efforts to optimize the Random Forest model using parameter optimization, we applied various techniques. We began with Super parametric optimization, adjusting the number of trees (n_estimator) for the S&P index manually and evaluating precision, recall rate, F1 value, and accuracy rate. We then explored Grid Search, systematically searching for the best parameter values within a specified range. However, due to its inefficiency and resource consumption, we turned to Random Search, a more valuable and efficient alternative based on low effective dimension. We optimized parameters like the number of decision trees, maximum tree depth, and minimum samples for node subdivision and leaf nodes. Model performance was evaluated through 3-fold cross-validation, using the AUC score of the ROC curve as the primary index. We also considered a fixed random number of seeds to mitigate sampling error. We observed that more trees in the Random Forest model don't necessarily mean better performance, considering resources, stability, and the model's robustness to outliers [20, 21]. Optimal parameters, obtained through random sampling, were evaluated using the AUC score and demonstrated effectiveness by comparing results before and after optimization.

4.4 Evaluating Indicator

Our assessment goes beyond simple metrics like accuracy, delving deeper into various performance indicators specific to the chosen classification model. We delve into precision, representing the accuracy of positive predictions, and recall, revealing the model's ability to capture true positives. We further utilize the F1-score, a harmonic mean of precision and recall, providing a balanced measure of overall effectiveness. These metrics offer a nuanced understanding of the model's strengths and weaknesses. In our analysis, we focused on imbalanced classes were present, and we found that metrics such as AUC-ROC and precision-recall curves offered valuable insights.

Through a comprehensive evaluation, we gained a clear picture of the model's performance, enabling us to make informed decisions about its suitability for real-world deployment. We identified potential biases, areas for improvement, and opportunities

for further optimization, ultimately ensuring the model effectively addressed the target problem.

5 Experimental Result

Preliminary results demonstrate that SARF outperforms conventional Random Forest models, showcasing its efficacy in predicting stock market movements, particularly in volatile market conditions. Our study revealed evidence of the better performance of the SARF model compared to conventional Random Forest models [22]. Through parameter optimization, we identified an optimal combination of parameters, leveraging different time windows for predicting stock market movements. Specifically, the model's precision, recall rate, F1 value, and accuracy rate were evaluated at a 60-day time window, demonstrating its efficacy for medium and long-term predictions within the 62–82 days range. These preliminary findings underscore the potential of SARF, particularly in volatile market conditions.

During our experiment we found analyzing stock market indices offers advantages over individual stocks for prediction purposes. Unlike single stocks, which can be highly volatile due to company-specific events, indices tend to be more stable as they aggregate multiple stocks. Indices, like the S&P 500, provide a holistic view of market sentiment by encompassing a diverse array of companies across different sectors. Consequently, analyzing an index allows for broader exposure and a more comprehensive understanding of market dynamics.

To address the variability in financial news release frequencies across stocks, we focus on periods characterized by high activity across 3 indices datasets. Given the time series nature of stock price data, we used cross-validation to assess model performance. This approach maintains temporal dependencies crucial for accurate predictions. Utilizing a standard 75–25 data split, we allocate 2250 samples for model training and 1500 for evaluation, ensuring a robust assessment while preserving data integrity. The details are listed in Table 2.

Table 2. The training and testing set for major US index

Index	Training Set	Testing Set
S&P 500	2015/02/21–2019/12/02	2020/01/10–2022/12/29
Nasdaq	2015/01/08–2019/12/06	2020/02/28–2023/10/30
Dow Jones	2016/02/18–2020/12/04	2021/01/10–2022/11/30

In the comparative analysis with the LSTM model, our constructed optimized random forest stock prediction model exhibited advantages. ROC curves drawn from the training results showcased the performance of each model, with the comprehensive results presented in Table 3 indicating the optimized random forest model's superior performance over the original random forest and LSTM models. This emphasizes the generalization

ability and superior trend prediction capability achieved through the random forest algorithm's decision tree feature extraction and random parameter optimization [23, 24]. The utilization of Precision-Recall curves further deepened our understanding, highlighting the SARF model's stability in stock forecasting, especially when compared to the LSTM model across various stock market index. These findings substantiate the effectiveness of SARF in providing robust predictions in dynamic stock market environments. The hybrid nature of SARF offers some advantages over traditional Random Forest models such as inclusion of sentiment features and technical indicators provides additional insights into market sentiment and historical price movements, allowing for a more comprehensive understanding of the model's predictions (Table 4).

Table 3. Performance Comparison

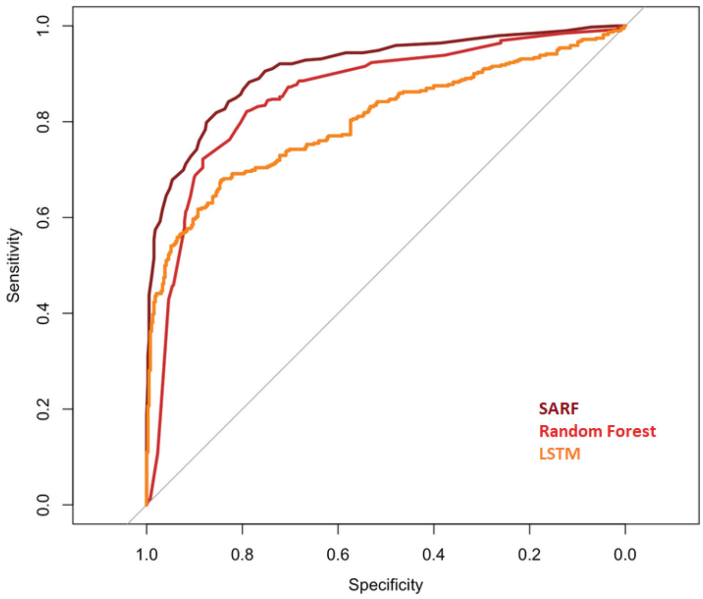

Table 4. Experiments Outcome evaluating the accuracy of models on stock indices

Index	Traditional Random Forest	LSTM	Optimized Random Forest (SARF)
S&P 500	0.67	0.58	0.78
Nasdaq	0.64	0.69	0.85
Dow Jones	0.59	0.61	0.82

6 Conclusion and Future Works

This study introduces a new method called SARF to enhance the precision of stock market predictions. By leveraging sentiment analysis in conjunction with FinGPT and an optimized Random Forest model, our research aims to achieve more accurate and reliable forecasts. The initial findings showcase the promising potential of SARF, positioning it as a valuable model for practical applications in financial forecasting. Through evaluated experiments, SARF shows better performance compared to traditional Random Forest and LSTM models, boasting an average accuracy enhancement of 9.23% and reduced prediction errors when predicting stock market movements.

Future research will explore the scalability of SARF to handle larger datasets, investigate additional sentiment features using other LLM in financial domain, and assess its performance in diverse market conditions. Additionally, the integration of real-time sentiment analysis could further enhance the model's responsiveness to dynamic market changes. It would be worth using ML optimization techniques to tune the hyperparameters and improve the model prediction accuracy further.

References

1. Chen, Y., et al.: Financial trading strategy system based on machine learning. Math. Probl. Eng. **2020**, 1–13 (2020)
2. Lv, J., Wang, C., Gao, W., Zhao, Q.: An economic forecasting method based on the LightGBM-optimized LSTM and time-series model. Comput. Intell. Neurosci. (2021)
3. NekoeiQachkanloo, H., et al.: Artificial counselor system for stock investment. Natl. Conf. Artif. Intell. **33**(1), 9558–9564 (2019)
4. Sharma, N., Juneja, A.: Combining of random forest estimates using LSboost for stock market index prediction. In: 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, pp. 1199–1202. IEEE (2017)
5. Tan, Z., Yan, Z., Zhu, G.: Stock selection with random forest: an exploitation of excess return in the Chinese stock market. Heliyon **58**, e02310 (2019)
6. Liu, X.Y., Wang, G., Zha, D.: FinGPT: democratizing internet-scale data for financial large language models. arXiv preprint arXiv:2307.10485, 19 July 2023
7. Nti, K.O., Adekoya, A., Benjamin, W.: Random forest-based feature selection of macroeconomic variables for stock market prediction. Am. J. Appl. Sci. **167**, 200–212 (2019)
8. Mehtab, S., Sen, J.: A robust predictive model for stock price prediction using deep learning and natural language processing. arXiv: Statistical Finance (2019). https://doi.org/10.1109/I2CT.2017.8226316
9. Basak, S., et al.: Predicting the direction of stock market prices using tree-based classifiers. N. Am. J. Econ. Finance **47**, 552–567 (2019)
10. Zhou, F., et al.: Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. Appl. Soft Comput. **84**, 105747 (2019)
11. Oriani, F.B., Coelho, G.P.: Evaluating the impact of technical indicators on stock forecasting. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, pp. 1–8. IEEE (2016). https://doi.org/10.1109/SSCI.2016.7850017
12. Beyaz, E., et al.: Comparing technical and fundamental indicators in stock price Forecasting. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, pp. 1607–1613. IEEE (2018). https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00262

13. Lohrmann, C., Luukka, P.: Classification of intraday S&P500 returns with a Random Forest. Int. J. Forecast. **35**(1), 390–407 (2019)
14. Loke, K.S.: Impact of financial ratios and technical analysis on stock price prediction using random forests. In: 2017 International Conference on Computer and Drone Applications (IConDA), Kuching, pp. 38–42. IEEE (2017). https://doi.org/10.1109/ICONDA.2017.8270396
15. Basak, S., Kar, S., Saha, S., Khaidem, L., Dey, S.R.: Predicting the direction of stock market prices using tree-based classifiers. N. Am. J. Econ. Finance **47**, 552–567 (2019)
16. Darapaneni, N., et al.: Stock price prediction using sentiment analysis and deep learning for Indian markets. arXiv preprint arXiv:2204.05783, 7 Apr 2022
17. Wang, Q., Xu, W., Zheng, H.: Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. Neurocomputing **299**, 51–61 (2018)
18. Oriani, F.B., Coelho, G.P.: Evaluating the impact of technical indicators on stock forecasting. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, pp. 1–8 (2016). https://doi.org/10.1109/SSCI.2016.7850017
19. Heyman, D., Lescrauwaet, M., Stieperaere, H.: Investor attention and short-term return reversals. Finance Res. Lett. **29**, 1–6 (2019)
20. Nisar, T.M., Yeung, M.: Twitter as a tool for forecasting stock market movements: a short-window event study. J. Finance Data Sci. **4**(2), 101–119 (2018)
21. Lee, T.K., Cho, J.H., Kwon, D.S., Sohn, S.Y.: Global stock market investment strategies based on financial network indicators using machine learning techniques. Expert Syst. Appl. **117**, 228–242 (2019)
22. Feuerriegel, S., Gordon: News-based forecasts of macroeconomic indicators: a semantic path model for interpretable predictions. Eur. J. Oper. Res. **272**(1), 162–175 (2019)
23. Pathak, A., Shetty, N.P.: Indian stock market prediction using machine learning and sentiment analysis. In: Behera, H., Nayak, J., Naik, B., Abraham, A. (eds.) Computational Intelligence in Data Mining. AISC, vol. 711, pp. 595–603. Springer, Singapore (2019). https://doi.org/10.1007/978-981-10-8055-5_53
24. Khan, W.: Stock market prediction using machine learning classifiers and social media, news. J. Ambient Intell. Hum. Comput. (2020)