FROM FINANCIAL NEWS TO PRICE FORECASTS: ASSESSING THE ROLE OF NLP-BASED SENTIMENT ANALYSIS AND VARIABLE SELECTION IN STOCK PRICE PREDICTION

Erdem Korhan Akçay, İsmail Yenilmez*

Department of Statistics, Eskisehir Technical University, Eskişehir, Turkey

Abstract

This study investigates the impact of NLP-based sentiment analysis and variable selection techniques on stock price prediction. Sentiment indicators are extracted using Natural Language Processing (NLP) methods, including TextBlob-based sentiment polarity scores, VADER compound scores, and domain-specific lexicon-based sentiment scores. To address the complexity and dimensionality of financial data, variable selection techniques (PCA, LASSO, Elastic Net, PCA + LASSO, and PCA + Elastic Net) are employed. These methods help in constructing a more efficient feature set by reducing noise and multicollinearity. The selected features, combined with sentiment variables, are utilized in predictive models including ARIMAX, ANN, LSTM, and GRU. The models are tested on eight publicly traded stocks (AAPL, AMZN, GOOG, META, MSFT, NFLX, NVDA, TSLA) over a four-year period. The results indicate that the inclusion of sentiment variables improves forecasting performance, particularly when combined with dimensionality reduction and regularization techniques. Among these approaches, combinations of PCA with regularization techniques often lead to more stable and competitive forecasting performance. The findings highlight the value of integrating unstructured textual data from financial news into time series forecasting models, contributing to improved predictive performance in stock market applications.

Keywords: *Stock Price Prediction, Sentiment Analysis, Variable Selection, Dimensionality Reduction, Deep Learning Models*

1. INTRODUCTION

Stock price forecasting remains one of the most challenging yet essential topics in financial markets due to the high level of uncertainty and the multidimensional dynamics of financial time series. Classical statistical approaches such as the Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) and volatility-focused models like the Autoregressive Conditional Heteroscedasticity (ARCH) and its generalized version GARCH (Engle, 1982) have long been utilized for financial forecasting. While these linear time series models perform reasonably well for short-term predictions, they often fall short in capturing the intricate, non-linear patterns inherent in market dynamics (Brockwell and Davis, 1991; Khan and Alghulaiakh, 2020).

To address these limitations, extended models such as ARIMAX have been developed by incorporating exogenous variables. Furthermore, the evolution of artificial intelligence has led to the growing popularity of machine learning (ML) and deep learning (DL) techniques in time series forecasting (Vuong et al., 2024; Shobayo et al., 2024). Techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, and ensemble learning methods have shown promising results in capturing non-linear relationships within stock markets (Htun et al., 2023; Chaudhary, 2020; Kumar and Haider, 2021). Among deep learning approaches, Long Short-Term Memory (LSTM) models have emerged as particularly effective due to their ability to capture both short- and long-term dependencies in sequential data (Sezer et al., 2020; Jiawei and Murata, 2019).

More recently, researchers have highlighted the importance of Recurrent Neural Network (RNN) architectures such as LSTM and Gated Recurrent Units (GRU) in modelling complex temporal patterns in financial datasets (Çalık et al., 2022; Mugenzi and Yenilmez, 2023; Atmaca and Yenilmez, 2024). These approaches often outperform classical models in a variety of market settings. Nevertheless, advanced models are often hindered by the lack of integration with non-price data, particularly investor sentiment and news-based information.

Sentiment analysis through Natural Language Processing (NLP) has become a critical component in financial prediction, especially with the growing availability of text data from news articles, analyst reports, and social media platforms such as Twitter (now X). Studies have demonstrated that sentiment extracted from financial texts significantly influences stock price movements and volatility (Tetlock, 2007; Bollen et al., 2011; Antweiler and Frank, 2004). For instance, Bollen et al. (2011) successfully predicted the direction of the Dow Jones Industrial Average (DJIA) with 87.6% accuracy using Twitter sentiment scores, while Mo et al. (2016) found that news sentiment can have a lagged impact on market returns.

Hybrid models that combine structured (price, volume) and unstructured (sentiment) data have shown superior predictive performance. Integration of NLP features into deep learning models such as LSTM, GRU, and Convolutional Neural Networks (CNN) has been especially effective (Mohan et al., 2022; Koukaras et al., 2022; Ouf et al., 2024). Recent works have adopted pretrained language models such as FinBERT and GPT-4 for sentiment classification and financial forecasting (Gu et al., 2024; Shobayo et al., 2024), underscoring their enhanced contextual understanding compared to traditional lexicon-based approaches.

Another critical component of model performance lies in feature selection. High-dimensional datasets necessitate effective dimension reduction or regularization techniques to reduce model complexity and enhance generalization. Established methods like Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and Principal Component Analysis (PCA) have been widely applied for this purpose (Niu et al., 2020; Pabuccu and Barbu, 2024; Sarıkoç and Çelik, 2024). The combination of sentiment scores with optimized feature selection has further boosted forecasting accuracy (Sudhakar and Naganjaneyulu, 2020; Yang et al., 2022; Vallarino, 2025).

In this study, we examine the effectiveness of NLP-based sentiment variables in stock price forecasting. We adopt a hybrid framework that integrates both financial and textual features using classical and deep learning-based time series forecasting models, including ARIMAX, ANN, LSTM, and GRU. Feature selection techniques such as PCA, LASSO, and Elastic Net are applied to improve model efficiency and accuracy. Using data from eight different stocks, the performance effects of sentiment integration and variable selection were analyzed. The remainder of this paper is structured as follows: Section 2 outlines the methodology, Section 3 presents the results, Section 4 discusses the findings, and Section 5 concludes the study.

2. MATERIALS AND METHODS

2.1. Materials

This study utilizes both traditional and advanced deep learning models for time series forecasting, aiming to evaluate the contribution of Natural Language Processing (NLP)-based sentiment variables and feature selection techniques in enhancing predictive accuracy. The data comprises daily stock price information and financial news texts related to eight selected companies listed on major stock exchanges. Stock market data includes historical closing prices, trading volume, and a range of technical indicators (e.g., RSI, MACD), while textual data was extracted from online financial news platforms and social media sources.

Sentiment variables were derived using lexicon-based and rule-based NLP tools. The following four sentiment-based features were constructed for each observation:

Sentiment Polarity: Basic classification of textual content as positive, negative, or neutral.

Sentiment Score: A numerical score reflecting the average polarity intensity.

VADER Compound Score: A normalized, rule-based sentiment intensity score using the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon (Hutto and Gilbert, 2014).

Lexicon Score: Aggregate sentiment score based on domain-specific sentiment dictionaries (Taboada et al., 2011).

These features were integrated with technical indicators and historical price data for model input.

The list of models, feature selection techniques, and NLP variables used in the analysis are summarized in Table 1.

Time Series Models	Feature Selection Techniques	NLP Sentiment Variables
ARIMAX	PCA	Sentiment Score
ANN	LASSO	Sentiment Polarity
LSTM	Elastic Net	VADER Compound
GRU	PCA + LASSO / Elastic Net	Lexicon Score

Table 1. Summary of models, feature selection techniques, and NLP variables

2.2. Methods

This study employs both traditional and deep learning-based forecasting models to predict stock prices using enriched feature sets. The time series models selected include:

Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX): A traditional linear model that integrates external variables into the autoregressive framework (Brockwell and Davis, 1991). The ARIMAX model can be expressed as an extension of the ARIMA framework by incorporating a vector of exogenous variables. In this formulation, the dependent variable is modeled as a function of its own past values, past forecast errors, and a set of external predictors.

$$\left(1-\sum_{i=1}^{p}\psi_i L^i\right)(1-L)^d(X_t-m_t)=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\epsilon_t \tag{2.1}$$

$$L^k X_t = X_{t-k} \tag{2.2}$$

$$m_t = c+\sum_{i=0}^{b} n_i y_{t,i} \tag{2.3}$$

Here, p, d, and q denote the orders of the autoregressive, differencing, and moving average components, respectively. ψ_i and θ_i represent the autoregressive and moving average parameters, respectively. ϵ_t is a normally distributed error term with zero mean. L is a lag operator as defined in Equations 2.1 and 2.2. m_t is defined in Equation 2.3. Here, y_t denotes a set of predictors. n_i represents the weights of these predictors, and b indicates the number of time series predictors.

Artificial Neural Networks (ANN): A feedforward neural network trained on historical price and sentiment data (Al Aradi and Hewahi, 2020).

Long Short-Term Memory (LSTM): A recurrent neural network variant capable of modeling long-term dependencies in time series (Hochreiter and Schmidhuber, 1997; Sezer et al., 2020).

Gated Recurrent Units (GRU): A simplified and computationally efficient version of LSTM with similar capabilities (Yenilmez and Mugenzi, 2023).

To enhance model performance and reduce overfitting, various feature selection and dimensionality reduction techniques were applied:

Principal Component Analysis (PCA): Used to reduce feature dimensionality by transforming correlated variables into orthogonal components (Sarıkoç and Çelik, 2024).

LASSO (Least Absolute Shrinkage and Selection Operator): A regularization method that performs variable selection and coefficient shrinkage simultaneously (Pabuccu and Barbu, 2024).

Elastic Net: A regularization method combining both LASSO and Ridge regression penalties to handle correlated predictors (Xing et al., 2022).

Hybrid methods such as PCA + LASSO and PCA + Elastic Net were implemented to benefit from both dimensionality reduction and sparsity.

Model inputs were standardized and normalized prior to training. Each model was evaluated using standard regression performance metrics, including Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), on both training and test datasets. The comparative design allowed for assessment of how different combinations of model architecture, feature selection, and sentiment variables influence prediction accuracy.

All deep learning models were implemented using TensorFlow and Keras libraries, while ARIMAX and feature selection procedures were carried out using the statsmodels and scikit-learn packages in Python.

2.3. Model Evaluation Criteria

The performance of the models was evaluated using the following metrics:

MSE (Mean Squared Error) measures the average squared difference between predicted and actual values. Lower MSE indicates better model performance. The formula is given in Equation 2.4:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (2.4)$$

where N: Number of observations, y_i: Actual value, $\hat{y}_i$: Predicted value.

RMSE (Root Mean Squared Error) is the square root of MSE, used to assess model performance. The formula is in Equation 2.5:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (2.5)$$

Although both MSE and RMSE were calculated for model evaluation, only RMSE values are reported here for the sake of brevity and interpretability.

3. RESULTS

This section presents the integrated analysis of financial time series and news sentiment data for stock price forecasting. The dataset includes trading days from September 2020 to December 2024, encompassing both numerical financial indicators and textual news content. A total of eight stocks were selected from major sectors, technology, communication, consumer goods, and consumer services, to analyze model performance under diverse market conditions.

3.1. Data Overview

Financial data were obtained via the Yahoo Finance API (Yahoo, 2020) using a custom-built Python data collection algorithm. Each stock is represented by key indicators such as Open, High, Low, Close, Volume, 5-day Moving Average, RSI, MACD, 20-day Standard Deviation, and the respective Sector Index. Meanwhile, news text data were collected from Business Insider (Insider, 2019), a reputable financial news outlet, aligned to the same time range.

From the news texts, the following Natural Language Processing (NLP) sentiment features were extracted using different algorithms:

Sentiment Score (via Textblob), Sentiment Polarity (via TextBlob (Loria, 2018)), VADER Compound Score (via VADER), Lexicon Score (financial-domain lexicon)

These sentiment metrics were incorporated alongside financial indicators to investigate their joint influence on stock price movements.

For robustness, only stocks with ≤20% missing news coverage were included in the final sentiment analysis phase. This threshold ensures reliability in measuring the impact of news sentiment on financial outcomes. Stocks with excessive gaps between news publication dates and trading days were excluded to avoid distortions in temporal alignment. To ensure the reliability of sentiment analysis, only stocks with a missing news data rate of 20% or less were included in the forecasting models. Table 2 presents the list of stocks that satisfied this data completeness criterion.

Ticker	Missing Rate (%)	Missing Days / Total Days
NFLX	19.7%	227 / 1151
NVDA	12.8%	149 / 1162
GOOG	7.9%	92 / 1162
META	7.1%	83 / 1162
MSFT	6.7%	78 / 1162
AAPL	5.2%	60 / 1162
AMZN	3.1%	36 / 1151
TSLA	2.3%	26 / 1151

Table 2. Stocks with ≤20% Missing News Data

3.2. Results for Selected Stocks

The results for the eight selected stocks are summarized in Table 3, which reports the top three performing models for each stock based on the RMSE metric. While both MSE and RMSE were computed, only RMSE values are presented for clarity and conciseness.

Table 3 presents the top three performing models for each of the eight selected stocks based on RMSE values, highlighting the effectiveness of different model-feature-sentiment combinations. Overall, recurrent neural network models, particularly LSTM and GRU, exhibited strong predictive performance relative to traditional approaches for several stocks, including NVDA, MSFT, AAPL, and NFLX. Feature selection techniques combining PCA and Elastic Net or LASSO were commonly used. In addition, NLP-based sentiment variables, such as Lexicon Score, VADER Compound, and combinations of Sentiment Score and Sentiment Polarity, played a significant role in improving forecasting performance. GOOG achieved the lowest RMSE (1.0189) with an LSTM model using PCA + Elastic Net and VADER Compound sentiment, while TSLA showed the highest RMSE values, suggesting greater prediction difficulty. These results underscore the importance of asset-specific model design and the careful integration of sentiment information in stock price forecasting.

4. DISCUSSION

The forecasting models developed in this study were applied to eight technology- and consumer-focused stocks (AAPL, AMZN, GOOG, META, MSFT, NFLX, NVDA, and TSLA), chosen based on their news data completeness and sectoral diversity. The evaluation was primarily based on the RMSE metric on test sets, and the results were analyzed across three dimensions: (1) model performance, (2) feature selection efficacy, and (3) the impact of sentiment-based NLP variables.

The performance of the four different modeling approaches (ARIMAX, ANN, LSTM, GRU) varied notably across the selected stocks. ARIMAX, while effective for AMZN, showed limitations for other stocks, likely due to its reduced ability to model complex and high-frequency dynamics. In contrast, ANN models delivered competitive performance for stocks characterized by high volatility or non-linear behavior (e.g., TSLA and META). However, traditional and shallow learning models lack the explicit sequential learning mechanisms of LSTM and GRU, which are specifically designed to capture temporal dependencies in time series data.

Among recurrent neural network models, LSTM and GRU often yielded the lowest RMSE values for several stocks, with only marginal performance differences observed between the two architectures. In certain cases, such as NVDA, GRU outperformed LSTM, potentially due to its lower parameter complexity and enhanced generalization capability. These results are consistent with existing literature, which suggests that GRU can offer comparable or superior performance to LSTM while requiring fewer computational resources when modeling complex temporal dependencies.

Ticker	Model	Feature Selection	Sentiment Variables	RMSE
AAPL	LSTM	Elastic Net only	L.S., S.S., V.C.	1.2318
AAPL	GRU	Elastic Net only	S.P., S.S., V.C.	1.2447
AAPL	GRU	Elastic Net only	S.P., S.S.	1.2468
AMZN	ARIMAX	PCA + LASSO	Lexicon Score	1.2083
AMZN	ARIMAX	PCA only	Lexicon Score	1.2083
AMZN	ARIMAX	PCA + LASSO	Sentiment Polarity	1.2085
GOOG	GRU	PCA + Elastic Net	VADER Compound	1.0189
GOOG	GRU	PCA + LASSO	None	1.0306
GOOG	LSTM	PCA + Elastic Net	Lexicon Score	1.0331
META	LSTM	PCA only	Lexicon Score	3.7118
META	ANN	PCA + Elastic Net	L.S., S.P., S.S., V.C.	3.7428
META	ANN	PCA + Elastic Net	L.S., S.P., V.C.	3.7431
MSFT	LSTM	PCA + Elastic Net	VADER Compound	2.0351
MSFT	GRU	PCA + LASSO	L.S., V.C.	2.0395
MSFT	LSTM	PCA + Elastic Net	L.S., S.S., V.C.	2.0398
NFLX	LSTM	PCA + Elastic Net	L.S., S.S., V.C.	4.4395
NFLX	GRU	PCA + Elastic Net	L.S., S.S., V.C.	4.5055
NFLX	LSTM	PCA + LASSO	L.S., S.P., S.S., V.C.	4.5118
NVDA	GRU	PCA + Elastic Net	S.S., V.C.	1.5043
NVDA	GRU	PCA + Elastic Net	S.P., V.C.	1.5278
NVDA	GRU	PCA + Elastic Net	Sentiment Score	1.5317
TSLA	ANN	PCA + LASSO	S.P., S.S.	4.3020
TSLA	LSTM	PCA + LASSO	Sentiment Score	4.3357
TSLA	ANN	PCA + LASSO	Lexicon Score	4.3383

Table 3. Results for Selected Stocks (Top 3 RMSE per Stock)

*Repeated model entries correspond to different sentiment variable configurations or feature selection specifications. Although the underlying model architecture is the same, these configurations yield distinct predictive performances.

In terms of feature selection, hybrid methods combining PCA with regularization techniques (Elastic Net or LASSO) yielded the most accurate results. The PCA+Elastic Net combination stood out for multiple stocks (GOOG, META, MSFT, NFLX, NVDA), whereas PCA+LASSO also performed exceptionally well for AMZN and TSLA. The incorporation of NLP-based sentiment variables proved highly influential, as 23 of the top 24 models (95.8%) included at least one sentiment-related feature. This provides strong empirical support for the hypothesis that investor sentiment and news flow contain valuable information for stock price prediction.

Interestingly, for AMZN and META, models using PCA alone achieved lower RMSE values than hybrid feature selection approaches in this study, suggesting that dimensionality reduction and decorrelation played a more influential role than additional feature elimination for these assets. AAPL was the only stock for which the top-performing models relied exclusively on Elastic Net without PCA, indicating that dimensionality reduction may obscure certain informative structures in its feature space, while Elastic Net's ability to handle correlated predictors proved particularly beneficial. For these stocks, multiple configurations of the same model architecture appear among the top-performing results, highlighting that forecasting performance is sensitive to sentiment variable specification and feature selection choices, in addition to model structure.

The incorporation of NLP sentiment variables proved highly impactful: 23 of the top 24 models (95.8%) included at least one sentiment feature. This strongly supports the hypothesis that investor sentiment and news flow have measurable, modelable effects on stock prices. Moreover, in many cases, combinations of multiple sentiment scores (e.g., Sentiment Score, VADER Compound, Lexicon Score) led to superior model performance, implying that different sentiment measures capture complementary aspects of market perception.

That said, the GOOG stock presented a notable exception: its second-best result was achieved without any NLP variables, using only PCA and LASSO. This suggests that in certain contexts, financial indicators alone may outperform sentiment-enhanced models, possibly due to lower sensitivity to news, overlapping signal content, or limitations in the NLP methods for specific stock-related news.

Overall, the findings demonstrate that the added value of sentiment analysis depends on stock-specific characteristics, news flow patterns, and modeling techniques. Therefore, the optimal forecasting strategy should be fitted to the asset in question, with careful consideration of how sentiment data is sourced, processed, and integrated.

5. CONCLUSIONS

This study proposed a hybrid forecasting framework that integrates financial time series data with NLP-based sentiment indicators to enhance stock price prediction. The analysis was conducted on eight publicly traded stocks selected based on their low rate of missing news data and sectoral relevance, allowing for a comprehensive evaluation of model performance under varying market conditions.

The empirical results indicate that recurrent neural network models, particularly LSTM and GRU, can provide competitive predictive performance in modeling the temporal structure of financial time series, especially when integrated with sentiment-based features. However, the findings also reveal that no single model architecture consistently dominates across all stocks. For certain assets, such as AMZN and TSLA, traditional or shallow learning approaches (e.g., ARIMAX or ANN) produced competitive or superior results, highlighting the importance of stock-specific model selection.

Feature selection played a critical role in forecasting accuracy. In many cases, hybrid approaches combining PCA with regularization techniques such as Elastic Net or LASSO led to improved prediction performance by reducing dimensionality and mitigating multicollinearity. Nevertheless, the effectiveness of these methods varied across stocks. For example, some assets benefited more from PCA alone or from regularization without dimensionality reduction, indicating that the optimal feature selection strategy is inherently data-dependent.

The integration of NLP-based sentiment variables proved to be highly influential. The vast majority of top-performing models incorporated at least one sentiment metric, confirming that news-derived sentiment contains valuable information beyond traditional financial indicators. Moreover, combinations of multiple sentiment measures, such as Sentiment Score, Sentiment Polarity, VADER Compound, and Lexicon Score, often yielded enhanced performance, suggesting that different sentiment representations capture complementary aspects of market perception. At the same time, isolated cases showed that financial indicators alone can occasionally match or outperform sentiment-enhanced models, emphasizing the need for careful validation.

Overall, the results indicate that effective stock price forecasting requires a flexible, asset-specific, and data-driven modeling strategy, rather than reliance on a single universally optimal model or feature selection technique. By integrating structured financial data with unstructured textual sentiment information and employing appropriate variable selection methods, the proposed framework offers a robust and adaptable approach to financial forecasting.

Future research may further enhance model robustness by exploring adaptive regularization techniques, such as adaptive LASSO (Yenilmez (2025), and by investigating the effects of alternative imputation (Yenilmez and Atmaca (2025) and sentiment aggregation strategies in settings with incomplete or irregular news data. Such extensions would contribute to improving the stability, scalability, and real-world applicability of sentiment-integrated forecasting systems.

ACKNOWLEDGMENTS

This study was supported by Eskişehir Technical University Scientific Research Project Commission under grant no: 24ADP185. This study was produced from Erdem Korhan Akçay's Master's Thesis, which was supervised by Dr. İsmail Yenilmez. An earlier version of this study was presented at 24th International Conference Economy & Business (ICE&B) (For details see [Akçay & Yenilmez 2025]).

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process:

During the preparation of this manuscript, the authors used ChatGPT to assist in improving the readability and language of the text. After employing this tool, the authors thoroughly reviewed and edited the content as needed and took full responsibility for the final version. The AI assisted technology was used only to refine the manuscript’s clarity and presentation, following ethical guidelines.

REFERENCES

1. Akçay, E. K., & Yenilmez, İ. (2025). The impact of NLP-based sentiment analysis and variable selection techniques on stock price prediction. *Book of Abstracts of the 24th International Conference on Economy & Business (ICE&B)*, 17–20 August, Burgas, Bulgaria.
2. Al Aradi, M., & Hewahi, N. (2020, October). Prediction of stock price and direction using neural networks: Datasets hybrid modeling approach. In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)* (pp. 1-6). IEEE.
3. Antweiler, W. & Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance, 59(3), pp.1259–1294.
4. Atmaca, K. & Yenilmez, İ., 2024. RNNs and Transformer model in case of incomplete time series. In Proceedings of the V. International Applied Statistics Congress (UYIK - 2024), İstanbul, Türkiye, pp.800–805. ISBN: 978-975-7328-96-4.
5. Bollen, J., Mao, H. & Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), pp.1–8.
6. Brockwell, P.J. & Davis, R.A., 1991. Time Series: Theory and Methods. Springer.

7. Chaudhary, P., 2020. A deep learning-based approach for stock price prediction. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 11(3), pp.1983–1989.
8. Çalık, A.S., Yenilmez, İ. & Sevinçtekin, Ö., 2022. Comparison of Box-Jenkins and artificial neural networks methods: Demand forecast for ceramic sanitary ware. In EGE 7th International Conference on Applied Sciences, İzmir, Türkiye, pp.1024–1029. ISBN: 978-605-72197-9-4.
9. Engle, R.F., 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, 50(4), pp.987–1007.
10. Gu, W. et al., 2024. Predicting stock prices with FinBERT-LSTM: Integrating news sentiment analysis. In Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing, pp.67–72.
11. Hochreiter, S. & Schmidhuber, J., 1997. Long short-term memory. Neural Computation, 9(8), pp.1735–1780.
12. Htun, H.H., Biehl, M. & Petkov, N., 2023. Survey of feature selection and extraction techniques for stock market prediction. Financial Innovation, 9, 26.
13. Hutto, C.J. and Gilbert, E., 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, 8(1), pp.216–225.
14. Insider, 2019. Business Insider. Available at: https://markets.businessinsider.com/news/ [Accessed 31 May. 2025].
15. Jiawei, X. & Murata, T., 2019. Stock market trend prediction with sentiment analysis based on LSTM neural network. In International Multiconference of Engineers and Computer Scientists, pp.475–479.
16. Khan, S. & Alghulaiakh, H., 2020. ARIMA model for accurate time series stocks forecasting. International Journal of Advanced Computer Science and Applications, 11, Article 0110765.
17. Koukaras, P., Nousi, C. & Tjortjis, C., 2022. Stock market prediction using microblogging sentiment analysis and machine learning. Telecom, 3, pp.358–378.
18. Kumar, K. & Haider, M.T., 2021. Enhanced prediction of intra-day stock market using metaheuristic optimization on RNN–LSTM network. New Generation Computing, 39(1), pp.231–272.
19. Loria, S., 2018. textblob Documentation. Release 0.15. Available at: https://textblob.readthedocs.io/en/dev/# [Accessed 31 May. 2025].
20. Mo, S.Y., Liu, A. & Yang, S.Y., 2016. News sentiment to market impact and its feedback effect. Environmental Systems and Decisions, 36, pp.158–166.
21. Mohan, S. et al., 2022. Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. Computers in Biology and Medicine, 144.
22. Mugenzi, F. & Yenilmez, İ., 2023. Forecasting for GDP per capita using Multiplier Perceptron and Gated Recurrent Unit. In Proceedings of the Akdeniz 10th International Conference on Applied Sciences, Kyrenia, Cyprus, pp.318–327. ISBN: 978-625-6830-49-3.
23. Niu, T. et al., 2020. Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting. Expert Systems with Applications, 148, 113237.
24. Ouf, S. et al., 2024. A deep learning-based LSTM for stock price prediction using Twitter sentiment analysis. International Journal of Advanced Computer Science & Applications, 15(12).
25. Pabuccu, H. & Barbu, A., 2024. Feature selection with annealing for forecasting financial time series. Financial Innovation, 10, 87.

26. Sarıkoç, M. & Celik, M., 2024. PCA-ICA-LSTM: A hybrid deep learning model based on dimension reduction methods to predict S&P 500 index price. Computational Economics, 65, pp.2249–2315.
27. Sezer, O.B., Gudelek, M.U. & Ozbayoglu, A.M., 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90, 106181.
28. Shobayo, O. et al., 2024. Innovative sentiment analysis and prediction of stock price using FinBERT, GPT-4 and logistic regression: A data-driven approach. Big Data and Cognitive Computing, 8(11), 143.
29. Sudhakar, K. & Naganjaneyulu, S., 2020. Stock price prediction based on finance related news using NLP, Lasso and ARIMAX. i-manager's Journal on Software Engineering, 14(4), pp.11–19.
30. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), pp.267–307.
31. Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), pp.1139–1168.
32. Vallarino, D., 2025. An AI-enhanced forecasting framework: Integrating LSTM and transformer-based sentiment for stock price prediction. Journal of Economic Analysis, 4(3), pp.1–15.
33. Vuong, P.H. et al., 2024. A bibliometric literature review of stock price forecasting: From statistical model to deep learning approach. Science Progress, 107(1).
34. Xing, Y., Li, D. & Li, C., 2022. Time series prediction via elastic net regularization integrating partial autocorrelation. Applied Soft Computing, 129, 109640.
35. Yahoo, 2020. Yahoo Finance. Available at: https://finance.yahoo.com/quote/ [Accessed 31 May. 2025].
36. Yang, J., Wang, Y. & Li, X., 2022. Prediction of stock price direction using the Lasso-LSTM model combines technical indicators and financial sentiment analysis. PeerJ Computer Science, 8, e1148.
37. Yenilmez, İ., 2025. A novel robust adaptive Lasso estimator: Simulation studies and performance evaluation. In Proceedings of the 4th International Conference on Mathematics and Engineering, Bursa, Türkiye, pp.291–298. ISBN: 978-625-6283-75-6.
38. Yenilmez, İ. & Atmaca, K., 2025. Performance of deep learning models on imputed time series data: A simulation study and application to leading airline companies' stock price. International Journal of Advances in Engineering and Pure Sciences, 37(UYIK 2024 Special Issue), pp.30–39.
39. Yenilmez, İ. & Mugenzi, F., 2023. Estimation of conventional and innovative models for Rwanda’s GDP per capita: A comparative analysis of artificial neural networks and Box-Jenkins methodologies. Scientific African, 22, e01902.