International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Predicting Stock Market Movements Using News Sentiment and Machine Learning

Rakshit Gupta
Research Scholar, Department of Computer Science and Engineering, Shri Krishna University, Chhatarpur
Mr. Saket Nigam
Supervisor, Department of Computer Science and Engineering, Shri Krishna University, Chhatarpur

Abstract
Stock market prediction has always been a challenging task due to the volatility, complexity, and non-linear dynamics of financial markets. Traditional approaches based on fundamental and technical analysis often fail to capture the behavioral factors that drive market fluctuations. In recent years, the integration of news sentiment analysis with machine learning (ML) techniques has emerged as a promising solution for improving predictive accuracy. By extracting sentiment scores from financial news and combining them with historical stock data and technical indicators, researchers can capture both quantitative and qualitative aspects of market behavior. This allows models to reflect not only economic fundamentals but also investor psychology, which frequently influences short-term price movements.
In this study, a hybrid approach was applied to predict stock movements of selected Nifty50 companies by combining sentiment data with ML algorithms such as K-Nearest Neighbour (KNN), Random Forest (RF), XGBoost, and Long Short-Term Memory (LSTM). The results reveal that ensemble-based models, particularly Random Forest and XGBoost, outperform deep learning and instance-based methods, achieving the lowest error rates and highest R^2 scores. These findings confirm that sentiment-enhanced machine learning frameworks are effective in capturing market dynamics. The study highlights the importance of integrating behavioral signals with computational intelligence, paving the way for more robust, transparent, and practical forecasting models in modern financial markets.
Keywords: Stock Market Prediction, News Sentiment Analysis, Machine Learning, Deep Learning, Ensemble Methods

Introduction

Stock market prediction has traditionally been considered one of the most complex and uncertain tasks in finance due to the highly dynamic, volatile, and non-linear nature of market behavior. Conventional forecasting approaches relied primarily on fundamental analysis—evaluating company performance, earnings reports, and macroeconomic indicators—or technical analysis, which focused on historical price and volume patterns to predict future movements. While these methods provided useful insights, they often failed to capture the full range of factors influencing

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

stock prices, particularly the behavioral and psychological dimensions of investor decision-making. In today's interconnected global economy, stock prices are increasingly sensitive to external influences such as breaking news, political developments, global crises, and social media activity, which shape market sentiment and investor reactions. As a result, researchers and practitioners have turned to alternative approaches that integrate sentiment analysis and machine learning techniques to address the limitations of traditional models and improve predictive accuracy.

Sentiment analysis, a branch of natural language processing (NLP), has gained significant traction in financial research by quantifying opinions, emotions, and attitudes expressed in unstructured textual data. Sources such as financial news articles, company announcements, analyst reports, and even platforms like Twitter provide rich signals of collective investor mood that often precede market movements. For instance, positive earnings announcements or optimistic headlines can stimulate buying behavior, while negative news related to economic downturns or regulatory changes may trigger selling pressures. By extracting positive, negative, or neutral sentiment scores from textual content, sentiment analysis bridges the gap between qualitative narratives and quantitative forecasting. When combined with machine learning (ML) models, these sentiment indicators can significantly enhance the accuracy of predictions. Machine learning algorithms excel at uncovering complex relationships within high-dimensional datasets, making them well-suited for integrating textual sentiment with numerical stock data. This integration reflects a growing recognition that financial markets are not solely driven by rational fundamentals but also by investor psychology and perception, which can amplify volatility and drive short-term trends.

Machine learning further strengthens stock prediction by addressing the limitations of linear models in handling non-linear dependencies and large-scale datasets. Classical ML techniques such as support vector machines, random forests, and gradient boosting have been successfully applied to financial forecasting, while advanced deep learning models—particularly recurrent neural networks (RNNs), long short-term memory networks (LSTMs), convolutional neural networks (CNNs), and transformers—have demonstrated superior performance in capturing sequential patterns and contextual sentiment. The convergence of sentiment analysis and machine learning has therefore become a promising frontier in financial research, offering predictive frameworks that are not only more accurate but also adaptable to diverse market conditions. However, challenges such as data quality, model interpretability, real-time scalability, and generalizability remain pressing issues that researchers continue to address. This study focuses on predicting stock market movements through the integration of sentiment analysis and machine learning, aiming to evaluate the effectiveness of these methods, explore their limitations, and highlight their potential for advancing modern financial forecasting.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Role of News Sentiment in Shaping Market Movements

News sentiment plays a crucial role in shaping stock market dynamics because financial markets are not driven solely by fundamentals but also by investor psychology and perception. Headlines, press releases, policy announcements, and social media commentary often act as triggers for rapid market responses, amplifying volatility. Positive sentiment, such as optimistic earnings reports or favorable economic forecasts, can stimulate investor confidence and buying activity, pushing stock prices upward. Conversely, negative news, including political instability, corporate scandals, or global crises, can provoke fear and uncertainty, leading to sell-offs and price declines. Unlike traditional financial indicators that capture historical or numerical data, sentiment analysis taps into the emotional tone and behavioral aspects underlying market reactions. Studies such as Bollen et al. (2011) demonstrated that collective mood, as reflected in social media, strongly correlates with stock index movements, reinforcing the idea that sentiment serves as an early signal of market direction. In this sense, sentiment is not just supplementary but often predictive of short-term fluctuations, making it an essential input in modern forecasting frameworks.

Emergence of Machine Learning in Financial Prediction

The emergence of machine learning (ML) in financial prediction represents a paradigm shift from linear econometric models toward data-driven, adaptive, and non-linear approaches. Traditional models like ARIMA or regression struggle with the complex dependencies and volatility inherent in stock data. ML algorithms, however, excel at detecting hidden patterns in large, high-dimensional datasets, making them highly suitable for financial forecasting. Classical techniques such as support vector machines, random forests, and gradient boosting have been widely applied to capture structural dependencies in stock data. More recently, deep learning architectures—particularly long short-term memory networks (LSTMs), convolutional neural networks (CNNs), and transformers—have demonstrated strong performance in modeling sequential and contextual patterns, especially when combined with sentiment features. Machine learning's strength lies in its ability to learn adaptively from evolving datasets, thereby improving predictive accuracy in dynamic markets. At the same time, its integration with sentiment analysis enables hybrid frameworks that combine behavioral insights with technical indicators, reflecting the multifaceted drivers of stock price movements.

Research Methodology

The methodology adopted in this study integrates several supervised machine learning algorithms to predict stock index values and the daily direction of market movement. The approach begins with data collection and preprocessing, where historical stock prices and financial indicators are combined with relevant sentiment features extracted from news articles. Data cleaning, normalization, and feature selection are conducted to ensure the quality of input

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

variables. The processed dataset is then used to train multiple machine learning classifiers, each of which offers distinct advantages in handling complex, non-linear, and high-dimensional stock market data.

The K-Nearest Neighbour (KNN) algorithm is applied as a baseline classifier due to its simplicity and efficiency in capturing similarity-based patterns. As a non-parametric and lazy learning method, KNN identifies k-nearest data points to classify or predict unknown instances. It is particularly useful for short-term stock trend analysis, where proximity to recent patterns can provide reasonable predictions. However, due to its sensitivity to noisy data and computational intensity with large datasets, KNN is complemented with ensemble methods for robustness.

The Random Forest (RF) algorithm is employed as an ensemble learning approach that averages the predictions of multiple decision trees. By combining bagging and feature randomness, RF reduces variance and improves generalization performance. Each decision tree is trained on a bootstrapped dataset with randomly selected features, ensuring diversity among classifiers. RF is especially effective in handling complex datasets with mixed variables, making it a suitable choice for financial data where interdependencies are intricate.

Advanced boosting techniques such as the XGBoost classifier and deep learning models like Long Short-Term Memory (LSTM) networks are utilized. XGBoost constructs trees sequentially, correcting errors of previous trees, thus providing higher accuracy in stock movement prediction. Meanwhile, LSTM networks, with their ability to capture long-term dependencies in sequential data, are particularly suited for time series forecasting. Their gated memory mechanisms allow effective learning of temporal patterns in stock prices. Together, these models provide a hybrid methodology that leverages both classical ML efficiency and deep learning adaptability, ensuring comprehensive predictive performance.

Results and Discussion

Dataset and Data Pre-processing

To investigate the impact of public sentiment on fluctuations in stock market index prices, this study utilized historical stock data of ADANIPORTS, ASIANPAINT, and AXISBANK from the Nifty50 index, sourced from Kaggle. The dataset for ADANIPORTS spans the period from November 27, 2007, to April 30, 2021, while both ASIANPAINT and AXISBANK cover the period from January 3, 2000, to April 30, 2021. Alongside price and index data, textual information was collected and subjected to sentiment analysis using machine learning (ML) and deep learning (DL) models. Sentiment scores derived from this textual data were then integrated with the stock dataset to evaluate the relationship between market sentiment and stock price movements.

The integrated dataset, which combines interpolated historical stock prices, daily average sentiment scores, and five technical indicators, was used to perform single-step forecasting

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

experiments. The primary objective was to assess how enhanced public sentiment and technical indicators influence fluctuations in stock market prices. The performance of different models was evaluated using error metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

For the ADANIPORTS dataset, the LSTM model achieved an MSE of 402.43, RMSE of 20.06, MAE of 13.12, and an R^2 score of 0.965. The KNN classifier performed significantly better, with an MSE of 103.26, RMSE of 10.16, MAE of 6.07, and an R^2 score of 0.997. The XGBoost model further improved the results, yielding an MSE of 43.10, RMSE of 6.56, MAE of 2.20, and an R^2 score of 0.998. Among all models, the Random Forest (RF) classifier delivered the most accurate predictions, producing an MSE of 8.39, RMSE of 2.89, MAE of 1.22, and an exceptionally high R^2 score of 0.999. These findings demonstrate that ensemble methods, particularly Random Forest, outperform deep learning and other machine learning approaches in capturing the relationship between sentiment, technical indicators, and stock price movements in the ADANIPORTS dataset.

Table 1 Simulation of ADANIPORT

Model	MSE	RMSE	MAE	R^2-score
LSTM	402.43	20.06	13.12	0.965
KNN	103.26	10.16	6.07	0.997
XGBoost	43.10	6.56	2.20	0.998
RF	8.39	2.89	1.22	0.999

The performance comparison across four models shows a clear ranking in predictive accuracy for stock price forecasting. The LSTM model, designed for sequential data, recorded the highest error values, with an MSE of 402.43, RMSE of 20.06, and MAE of 13.12, alongside an R^2 score of 0.965. While LSTM effectively captured temporal patterns, it struggled with the dataset's complexity, resulting in less accurate predictions than the other approaches.

The KNN model improved considerably, achieving an MSE of 103.26, RMSE of 10.16, MAE of 6.07, and an R^2 of 0.997. This indicates strong predictive capacity, though it remains sensitive to noise and less efficient for larger datasets. The XGBoost classifier performed even better, reducing errors significantly with an MSE of 43.10, RMSE of 6.56, MAE of 2.20, and an R^2 of 0.998, reflecting its strength in handling non-linear dependencies. However, the Random Forest (RF) classifier delivered the best results, with the lowest error values (MSE 8.39, RMSE 2.89, MAE 1.22) and the highest R^2 score of 0.999. This confirms that RF is the most effective model in this context, offering superior precision, stability, and robustness for stock price prediction.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Table 2: Simulation of ASIANPAINT

Model	MSE	RMSE	MAE	R^2-score
LSTM	307.8	53.91	41.14	0.984
KNN	289.29	26.25	16.36	0.999
XGBoost	188.32	13.72	5.15	0.999
RF	47.90	6.92	3.27	0.999

The performance comparison of the four models demonstrates clear differences in their predictive capabilities. The LSTM model, widely known for handling sequential and time-series data, produced an MSE of 307.8, RMSE of 53.91, MAE of 41.14, and an R^2 score of 0.984. Although LSTM effectively captured temporal dependencies, its relatively high error values indicate limitations in adapting to the dataset's complexity when compared with other models.
The K-Nearest Neighbour (KNN) model showed substantial improvement, with an MSE of 289.29, RMSE of 26.25, MAE of 16.36, and an excellent R^2 score of 0.999, suggesting strong predictive accuracy and consistency. The XGBoost classifier further reduced errors, achieving an MSE of 188.32, RMSE of 13.72, and MAE of 5.15, while also maintaining a near-perfect R^2 of 0.999, confirming its robustness in identifying complex non-linear patterns. Finally, the Random Forest (RF) classifier delivered the best results, with the lowest MSE (47.90), RMSE (6.92), and MAE (3.27), alongside an R^2 of 0.999. These results highlight that ensemble-based methods, especially RF, outperform deep learning and instance-based methods for this dataset, offering higher precision and stability in stock movement prediction.

Table 3 : Simulation of AXISBANK

Model	MSE	RMSE	MAE	R^2-score
LSTM	367.78	33.90	28.17	0.911
KNN	155.05	12.45	8.13	0.999
XGBoost	17.35	4.16	2.21	0.999
RF	7.08	2.66	1.60	0.999

The simulation results for the AXISBANK dataset reveal distinct variations in model performance. The LSTM model, which is well-suited for sequential data, recorded an MSE of 367.78, RMSE of 33.90, MAE of 28.17, and an R^2 score of 0.911. While it managed to capture temporal dependencies in stock price trends, the relatively high error values indicate that LSTM struggled to provide accurate forecasts for this dataset, leading to weaker performance compared with other models.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

The KNN model showed substantial improvement, with an MSE of 155.05, RMSE of 12.45, MAE of 8.13, and an R^2 score of 0.999, demonstrating strong predictive reliability. However, more advanced ensemble techniques achieved even better accuracy. The XGBoost classifier significantly reduced errors, achieving an MSE of 17.35, RMSE of 4.16, and MAE of 2.21, while maintaining a near-perfect R^2 of 0.999. The Random Forest (RF) classifier emerged as the best-performing model, with the lowest MSE (7.08), RMSE (2.66), and MAE (1.60), coupled with an R^2 of 0.999. These findings suggest that ensemble-based methods, particularly RF, are highly effective in capturing complex relationships in the AXISBANK dataset, outperforming both deep learning and instance-based approaches in terms of precision and consistency.

Conclusion

This study examined the effectiveness of various machine learning models—LSTM, KNN, XGBoost, and Random Forest—in predicting stock price movements using integrated datasets of historical stock data, sentiment scores, and technical indicators. The results demonstrated that while deep learning models such as LSTM are well-suited for handling sequential data and capturing long-term dependencies, their performance in this case was less accurate compared with ensemble-based approaches. KNN achieved better results than LSTM, indicating that instance-based learning can capture short-term similarities effectively. However, its sensitivity to noisy data and reliance on dataset structure limited its overall efficiency. On the other hand, XGBoost significantly reduced prediction errors, highlighting the advantages of boosting in modeling complex, non-linear financial patterns.

Among all models, the Random Forest classifier consistently outperformed the others, producing the lowest error rates and the highest R^2 scores across datasets. This underscores the strength of ensemble learning in capturing diverse features and avoiding overfitting through bagging and feature randomness. The findings emphasize that integrating sentiment analysis with machine learning not only improves forecasting accuracy but also bridges behavioral insights with technical indicators. the study concludes that ensemble-based approaches, particularly Random Forest, provide the most robust and reliable framework for stock price prediction. Future work should explore multimodal data sources, real-time sentiment streaming, and explainable AI methods to enhance transparency, scalability, and adaptability in financial forecasting, making these models more applicable for investors, traders, and policymakers in dynamic market environments.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

References

1. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
2. Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2016). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5), 1367–1403.
3. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669.
4. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273.
5. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 66(5), 740–755.
6. Nelson, D. M., Pereira, A. C. M., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. *International Joint Conference on Neural Networks (IJCNN)*, 1419–1426.
7. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
8. Li, X., Xie, H., Wang, R., & Cai, Y. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69, 14–23.
9. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65.
10. Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653–7670.
11. Li, X., Xie, H., Wang, R., & Cai, Y. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14–23.
12. Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems, 55(3), 685–697.