International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

A Comprehensive Review of Stock Market Price Prediction through News Sentiment Analysis and Machine Learning Approaches

[1]Rakshit Gupta, [2]Mr. Saket Nigam

[1]Research Scholar, Department of Computer Science and Engineering, Shri Krishna University, Chhatarpur

[2]Supervisor, Department of Computer Science and Engineering, Shri Krishna University, Chhatarpur

Abstract

Stock market prediction has long been an area of interest for researchers, investors, and policymakers due to its potential to minimize risks and optimize decision-making. Traditional forecasting methods, relying on fundamental and technical analysis, have proven insufficient in capturing the dynamic, non-linear, and sentiment-driven nature of modern financial markets. With the increasing availability of unstructured data from financial news, press releases, and social media platforms, news sentiment analysis has emerged as a crucial tool for understanding investor psychology and its influence on price fluctuations. Simultaneously, advancements in machine learning (ML) and deep learning have enabled the development of models capable of processing high-dimensional datasets and identifying hidden patterns that conventional methods often overlook.

This review provides a comprehensive synthesis of studies that integrate sentiment analysis with machine learning techniques for stock price forecasting. It discusses the effectiveness of classical algorithms, ensemble methods, and deep learning architectures such as LSTM, CNN, and transformers in improving predictive accuracy. Furthermore, the review identifies key challenges—including data quality, scalability, overfitting, and model interpretability—while outlining promising directions such as multimodal data fusion, real-time sentiment streaming, explainable AI, and cross-market applicability. By critically evaluating existing literature, this study highlights the transformative potential of sentiment-driven machine learning models in shaping the future of financial forecasting.

Keywords: Stock Market Prediction, News Sentiment Analysis, Machine Learning, Deep Learning, Financial Forecasting

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Introduction

The prediction of stock market movements has long been a central focus of finance and economics, owing to its implications for investors, policymakers, and businesses. Traditionally, stock market forecasting relied on fundamental analysis, involving corporate earnings, macroeconomic indicators, and financial statements, or on technical analysis, which focuses on historical price and volume patterns. However, with the rise of digital media and the increasing influence of information flow on investor behavior, the role of textual data such as financial news, social media posts, and press releases has become pivotal. News sentiment, in particular, captures the collective tone and perception of market-related information, providing insights into investor psychology and market reactions. Sentiment analysis, a natural language processing (NLP) technique, has thus emerged as a powerful tool to extract meaningful indicators from unstructured text. The integration of sentiment data with stock market prediction not only enhances forecasting accuracy but also bridges the gap between quantitative modeling and behavioral finance, offering a more holistic view of market dynamics.

In parallel, machine learning (ML) techniques have revolutionized financial forecasting by enabling the handling of complex, non-linear, and high-dimensional datasets that traditional models often fail to capture. Models such as support vector machines (SVM), random forests, recurrent neural networks (RNN), and long short-term memory networks (LSTM) have been extensively explored to uncover hidden patterns in market data. Recent advances in deep learning and transformer-based architectures further expand these capabilities, allowing models to learn contextual sentiment from large-scale textual corpora. The convergence of sentiment analysis with machine learning thus represents a promising frontier in financial prediction, offering tools to model the dynamic interplay between information, investor behavior, and price fluctuations. This review seeks to synthesize existing literature on stock market prediction using news sentiment analysis and machine learning approaches, highlighting methodological advancements, comparative performance, challenges, and future research directions. By critically examining prior studies, the review aims to provide a comprehensive understanding of how sentiment-driven machine learning models are shaping modern financial forecasting.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

Background of Stock Market Forecasting

Stock market forecasting has been one of the most extensively studied areas in finance and economics due to its profound implications for investment strategies, risk management, and policymaking. Historically, forecasting approaches have revolved around fundamental analysis and technical analysis. Fundamental analysis evaluates the intrinsic value of securities by studying macroeconomic indicators, company financial statements, earnings reports, and market conditions. In contrast, technical analysis relies on identifying patterns in historical price and volume data to anticipate future movements. While these methods have been useful, they have limitations in dealing with the dynamic and unpredictable nature of modern financial markets. Stock prices are not only influenced by quantifiable economic indicators but also by qualitative factors such as news events, market rumors, and investor sentiment, which are often overlooked in conventional models. The advent of computational methods and access to vast volumes of unstructured data has encouraged a shift from purely statistical models to **data-**driven and AI-powered models. These modern approaches aim to integrate multiple dimensions of market influence, capturing both numerical and textual data. Thus, the background of stock market forecasting reflects a gradual transition from simplistic models based on price history or corporate fundamentals to sophisticated, hybrid models capable of handling high-frequency, high-dimensional, and real-time data for more accurate and adaptable predictions

Emergence of Sentiment Analysis in Financial Prediction

The emergence of sentiment analysis in financial prediction reflects the growing recognition of the role that investor psychology and market sentiment play in shaping stock price dynamics. Financial markets are not always efficient; they often respond disproportionately to news events, rumors, and opinions disseminated through various media outlets. The rise of digital journalism and social media platforms has further amplified this effect, making sentiment analysis an essential tool for capturing the emotional tone and bias embedded in financial information. Sentiment analysis, a subfield of natural language processing (NLP), focuses on determining whether a piece of text expresses a positive, negative, or neutral attitude. In finance, this means quantifying how investors and traders may react to earnings reports, market announcements, political developments, or global crises. Early approaches relied on lexicon-based methods, where predefined dictionaries assigned scores to words.

However, with advancements in machine learning and deep learning, more sophisticated techniques have emerged, enabling context-aware sentiment detection. For example, transformer-based models such as BERT and FinBERT allow the capture of nuanced sentiment from complex financial narratives. Empirical studies demonstrate that sentiment signals, when integrated with stock market data, improve predictive accuracy and provide early warning of market shifts. Hence, sentiment analysis has transitioned from being a supplementary tool to a core component in financial forecasting, bridging the gap between qualitative information and quantitative modeling.

Role of Machine Learning in Capturing Non-Linear Patterns

The role of machine learning (ML) in financial forecasting is particularly significant because stock market data is inherently non-linear, volatile, and influenced by multifaceted variables. Traditional econometric models, such as ARIMA or linear regression, often fall short in detecting the intricate dependencies between stock prices, trading volumes, and external factors like news or geopolitical events. Machine learning addresses these challenges by providing algorithms capable of uncovering hidden patterns, relationships, and anomalies within massive datasets. Classical ML models such as support vector machines, random forests, and gradient boosting methods can identify predictive signals even when variables interact in complex ways. Deep learning models, including recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), further enhance predictive power by effectively modeling temporal dependencies in sequential financial data. More recently, transformer architectures have enabled advanced contextual understanding of financial text, integrating sentiment with price data. The strength of machine learning lies in its adaptability—models can be continuously retrained as new data arrives, making them well-suited for fast-moving markets. Moreover, hybrid approaches that combine technical indicators, sentiment analysis, and macroeconomic variables benefit from ML's capacity to process heterogeneous data types. While challenges such as overfitting, interpretability, and computational complexity remain, machine learning has proven to be a powerful paradigm shift, enabling more robust, dynamic, and accurate modeling of financial markets compared to traditional techniques.

Objectives and Scope of the Review

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

The primary objective of this review is to provide a comprehensive synthesis of existing research on stock market prediction using news sentiment analysis and machine learning techniques. The review aims to critically evaluate how these two domains—sentiment analysis and machine learning—intersect to enhance the accuracy and reliability of financial forecasting. Specifically, the paper seeks to (1) trace the evolution of forecasting methods from traditional approaches to sentiment-driven machine learning models, (2) analyze the strengths and weaknesses of different sentiment extraction methods, (3) compare the performance of machine learning and deep learning techniques in stock prediction, and (4) identify gaps and opportunities for future research. The scope of the review extends across a wide range of studies, covering classical machine learning algorithms, deep learning architectures, and hybrid models that integrate textual sentiment with numerical market data. It also considers the role of different data sources, such as financial news portals, earnings reports, and social media, in shaping sentiment-driven predictions. Furthermore, the review highlights the challenges of data quality, real-time applicability, and model interpretability, while exploring solutions proposed in the literature. By consolidating insights from various studies, the review not only serves as a reference point for scholars and practitioners but also outlines pathways for innovation in predictive modeling. Ultimately, its scope is both academic and practical, aiming to inform future research while offering insights valuable to investors, financial analysts, and policymakers.

Literature Review

The prediction of stock prices has been a central challenge in computational finance, and numerous studies have proposed advanced machine learning and deep learning frameworks to improve accuracy. Raviraj, Pai, and Pai (2021) applied a deep learning approach to predict Indian stock prices using time series data. Their study demonstrated the effectiveness of recurrent neural networks in modeling sequential dependencies and highlighted how deep architectures outperform traditional econometric models in handling stock volatility. Similarly, Su and Yi (2022) explored Hidden Markov Models (HMMs) for efficient stock prediction, emphasizing probabilistic state transitions as a way to capture hidden patterns in market behavior. These studies collectively underline the importance of moving beyond linear models, instead leveraging architectures that can adapt to dynamic market signals.

However, while deep learning enhances predictive performance, issues such as overfitting and interpretability persist, especially in volatile markets.

Parallel to this, scholars have examined the integration of external information, particularly news sentiment, to enrich predictive frameworks. Duarte, Gonzalez, and Cruz (2021) analyzed the Brazilian market and demonstrated that incorporating news data improved the detection of potential stock price falls. Their study highlighted how sentiment-driven approaches can serve as early warning systems, complementing traditional price and volume-based predictors. Wu et al. (2020) further advanced this direction by proposing a labeling method for financial time series based on trends, allowing models to distinguish between upward and downward market shifts. Such innovations illustrate the growing consensus that stock market movements are influenced not only by numerical data but also by textual and event-driven factors, which machine learning algorithms are well-suited to capture.

Deep learning continues to dominate predictive modeling, with a variety of architectures being tested and compared. Staffini (2022) introduced a novel approach using deep convolutional generative adversarial networks (DCGANs) for stock price forecasting, which outperformed conventional neural networks by generating synthetic data to balance training sets. Ding and Qin (2020) applied an LSTM-based associated network model, demonstrating the ability of long short-term memory networks to capture long-range temporal dependencies in stock data. Similarly, Qi, Khushi, and Poon (2020) designed an event-driven LSTM for foreign exchange prediction, showing that sequential models can adapt effectively to high-frequency data environments. Collectively, these studies reinforce the versatility of deep learning architectures in modeling sequential and nonlinear financial patterns, though their computational complexity and reliance on large datasets remain challenges for practical adoption.

Hybrid and ensemble approaches have also gained traction in recent literature, offering solutions to some of the limitations of single-model frameworks. Liu, Pei, and Zou (2021) examined the volatility factors of Chinese listed companies using an integrated FA-ANN-MLP model, which combined factor analysis with artificial neural networks to better capture structural influences on stock fluctuations. Lv, Gao, and Yu (2021) developed a hybrid transfer learning framework for stock index forecasting, emphasizing the advantage of knowledge transfer across markets and datasets. These hybrid approaches illustrate how

combining feature engineering, dimensionality reduction, and machine learning architectures can lead to more robust and generalizable models. Similarly, De Pauli, Kleina, and Bonat (2020) compared different artificial neural network architectures for Brazilian market forecasting, concluding that the performance of ANN models depends heavily on architecture design and parameter tuning. Such studies emphasize that predictive accuracy can be enhanced through model diversity, hybridization, and systematic evaluation of architectures.

Overall, the reviewed studies demonstrate significant progress in stock market prediction through machine learning and sentiment-driven models, yet also highlight persistent challenges. While deep learning architectures such as LSTM, CNN, and GANs have shown superior accuracy, their black-box nature raises concerns regarding interpretability and practical use in finance, where explainability is crucial. Studies incorporating news sentiment (e.g., Duarte et al., 2021) reveal the critical role of behavioral finance, yet also face challenges of data quality, bias, and language processing in multilingual markets. Hybrid models and transfer learning frameworks present promising directions, offering improved generalization across different contexts. However, issues such as computational overhead, real-time applicability, and overfitting remain open research problems. In summary, the literature reflects a vibrant and rapidly evolving field, where the integration of sentiment analysis, deep learning, and hybrid approaches represents the frontier of financial prediction research.

Stock market prediction has increasingly embraced big data analytics and machine learning as powerful tools for financial forecasting. Peng (2019) highlighted the application of big data analytics in stock prediction, demonstrating how large-scale, high-frequency financial data can be processed to detect patterns that traditional methods often miss. Site, Birant, and Isik (2019) extended this direction by systematically comparing machine learning models such as decision trees, support vector machines, and ensemble approaches for forecasting stock prices, finding that model performance varied significantly based on dataset characteristics and feature selection. Similarly, Dingli and Fournier (2017) emphasized the adaptability of machine learning techniques in financial time series forecasting, noting that models like neural networks could capture non-linear dependencies more effectively than econometric approaches. Balaji, Ram, and Nair (2018), focusing on Bankex data, empirically tested deep learning models and confirmed their superiority in predictive accuracy compared

to shallow models. These studies collectively reinforce the shift from conventional econometrics to data-driven computational intelligence for more robust financial forecasting.

In parallel, machine learning models have been refined to handle the complexity of financial series. Vijh et al. (2020) applied various ML techniques for predicting stock closing prices and identified that ensemble models performed best in balancing bias and variance. Suzgun, Belinkov, and Shieber (2018) evaluated the generalization capabilities of LSTM models, stressing their potential and limitations when applied to structured sequences such as stock data. This emphasis on model generalization is crucial, as overfitting remains a recurring issue in financial prediction. Moreover, Charles, Simon, and Daniel (2008) investigated exchange rate volatility in the context of the Ghana Stock Exchange, demonstrating how macroeconomic variables significantly influence stock movements. By connecting currency fluctuations to stock performance, their work underscores the necessity of integrating external economic indicators into prediction frameworks. Together, these contributions highlight the growing complexity of financial forecasting, where diverse machine learning architectures and macroeconomic variables must be balanced to produce reliable outcomes.

The efficiency of financial markets has also been scrutinized in recent scholarship. Jayakumar and Sultan (2013) tested the weak form efficiency of the Indian stock market using NSE data and found that stock returns exhibited predictable patterns, contradicting the Efficient Market Hypothesis (EMH). Their findings provide theoretical justification for employing machine learning in Indian markets, where inefficiencies may be exploited for predictive advantage. Similarly, the pioneering work of Bollen, Mao, and Zeng (2011) introduced social media sentiment as a predictive feature, showing that Twitter mood metrics correlated strongly with Dow Jones Industrial Average movements. This research marked an important shift by demonstrating that investor psychology and collective mood, as expressed on digital platforms, can significantly influence stock markets. Chen, De, Hu, and Hwang (2016) expanded this line of inquiry by examining the “wisdom of crowds” in financial forecasting, analyzing how stock-related opinions on social media reflected valuable insights for market movements. These studies collectively indicate that sentiment analysis is not merely supplementary but often central to modern forecasting models.

A recurring theme across the literature is the integration of hybrid approaches that combine technical, fundamental, and sentiment-driven data. For instance, studies like Balaji et al.

(2018) and Vijh et al. (2020) highlight the success of deep learning models in capturing sequential dependencies in price data, while works such as Bollen et al. (2011) and Chen et al. (2016) emphasize the predictive power of textual data from social media. By uniting these two perspectives, hybrid models provide a more holistic understanding of market behavior. The role of transfer learning, as suggested in related literature, further supports cross-market applicability, allowing predictive models trained in one context to adapt effectively to another. However, challenges remain in data quality, noise reduction, and real-time processing, especially when integrating large volumes of unstructured sentiment data with numerical financial indicators. The heterogeneity of datasets and market conditions continues to complicate the development of universally applicable models.

Challenges and Limitations

One of the foremost challenges in stock market prediction using news sentiment and machine learning lies in data quality. Financial news and social media sources are prone to bias, noise, and misinformation, which can distort sentiment signals. News outlets may frame events with varying degrees of optimism or pessimism, and automated sentiment classifiers often struggle with sarcasm, ambiguity, or domain-specific financial terminology. Furthermore, data availability is uneven—while large markets like the U.S. provide abundant financial news and disclosures, emerging markets often suffer from limited datasets. Inconsistent labeling of sentiment data and the presence of duplicate or irrelevant content further complicate preprocessing. These issues not only reduce model accuracy but also raise concerns about the reliability of sentiment-driven predictions. Additionally, building models capable of real-time forecasting presents scalability difficulties. Processing massive streams of text, integrating them with market data, and generating predictions in seconds requires significant computational power and optimized algorithms, which are often expensive and resource-intensive.

Another persistent limitation relates to overfitting and generalizability. Machine learning models trained on historical data often capture noise rather than genuine predictive patterns, leading to poor performance when applied to unseen scenarios. Financial markets are highly dynamic, influenced by geopolitical crises, policy changes, and unexpected shocks, making generalization a constant challenge. Moreover, the interpretability of models is a critical concern in finance. Black-box models such as deep neural networks may achieve high

predictive accuracy, but they offer little transparency regarding how predictions are derived. For investors, regulators, and policymakers, explainability is as important as accuracy, since financial decisions involve substantial risk. The growing field of explainable AI (XAI) offers potential solutions, but balancing accuracy with interpretability remains unresolved. Together, these challenges underscore the need for robust, transparent, and adaptable prediction frameworks.

Future Research Directions

Future advancements in stock market prediction will depend heavily on the integration of multimodal data fusion, combining not just text-based sentiment from news and social media but also audio and video signals from earnings calls, press conferences, and interviews. Such multimodal approaches can capture subtle cues such as tone of voice, hesitation, or visual expressions that influence investor perceptions. Alongside this, the development of real-time sentiment streaming systems is critical for applications like high-frequency trading, where milliseconds can determine profitability. This requires building highly efficient pipelines that process vast amounts of unstructured textual and audiovisual data at scale, while maintaining low latency and high accuracy.

Another promising direction is the adoption of explainable AI (XAI) to address the interpretability gap in financial prediction models. XAI methods can help investors and regulators understand how sentiment features influence predictions, improving trust and usability. Furthermore, future studies should emphasize cross-market and cross-domain applicability, testing whether models trained in one market (e.g., U.S. equities) can be generalized to other contexts such as emerging markets or alternative asset classes like forex and cryptocurrencies. This would not only improve robustness but also reduce dependency on market-specific datasets, paving the way for more universal, adaptive, and transparent forecasting systems.

Conclusion

The prediction of stock market prices has evolved significantly, moving from traditional methods based on fundamental and technical analysis to modern, data-driven approaches that integrate sentiment analysis and machine learning techniques. This review has highlighted how financial forecasting increasingly relies on unstructured data sources such as news articles, press releases, and social media streams, which capture investor psychology and

collective sentiment. Machine learning models, ranging from classical algorithms like support vector machines to advanced deep learning architectures such as LSTM and transformers, have demonstrated strong potential in capturing the non-linear and dynamic behavior of markets. Moreover, hybrid approaches that combine sentiment features with technical and macroeconomic indicators offer more holistic frameworks, showing improved predictive accuracy. Despite these advancements, challenges related to data quality, scalability, overfitting, and model interpretability persist, limiting the seamless adoption of these models in real-world trading and policymaking contexts.

Looking forward, the future of stock market forecasting lies in multimodal integration, real-time sentiment processing, and explainable AI frameworks. Incorporating text, audio, and video data from sources like earnings calls can provide deeper insights into market reactions, while real-time streaming models can enable high-frequency trading strategies. At the same time, explainability will be crucial to build trust among investors, regulators, and financial institutions, ensuring that predictions are not only accurate but also transparent. Extending predictive models across multiple markets and domains will further enhance their robustness and adaptability. In essence, the convergence of sentiment analysis and machine learning represents a transformative direction in financial prediction, bridging the gap between behavioral finance and computational intelligence. While challenges remain, the growing body of literature indicates that sentiment-driven machine learning models will continue to redefine financial forecasting, offering actionable insights for researchers, practitioners, and policymakers alike.

References

1. Shravan Raviraj, Manohara Pai M. M., & Krithika M. Pai. (2021). Share price prediction of Indian stock markets using time series data - A deep learning approach. *IEEE Mysore Sub Section International Conference (MysuruCon), IEEE*.
2. Duarte, J. J., Gonzalez, S. M., & Cruz, J. C. (2021). Predicting stock price falls using news data: Evidence from the Brazilian market. *Computational Economics, 57*(1), 311–340.
3. Su, Z., & Yi, B. (2022). Research on HMM-based efficient stock price prediction. *Mobile Information Systems*.

4. Staffini, A. (2022). Stock price forecasting by a deep convolutional generative adversarial network. *Frontiers in Artificial Intelligence, 5*.
5. Liu, J., Pei, X., & Zou, J. (2021). Analysis and research on the stock volatility factors of Chinese listed companies based on the FA-ANN-MLP model. In *International Conference on Computer, Blockchain and Financial Development (CBFD), IEEE*.
6. Lv, C., Gao, B., & Yu, C. (2021). A hybrid transfer learning framework for stock price index forecasting. In *IEEE International Conference on Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, IEEE*.
7. Ding, G., & Qin, L. (2020). Study on the prediction of stock price based on the associated network model of LSTM. *International Journal of Machine Learning and Cybernetics, 11*(6), 1307–1317.
8. De Pauli, S. T. Z., Kleina, M., & Bonat, W. H. (2020). Comparing artificial neural network architectures for Brazilian stock market prediction. *Annals of Data Science, 7*(4), 613–628.
9. Qi, L., Khushi, M., & Poon, J. (2020). Event-driven LSTM for forex price prediction. In *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1–6). IEEE.
10. Wu, D., Wang, X., Su, J., Tang, B., & Wu, S. (2020). A labeling method for financial time series prediction based on trends. *Entropy, 22*(10), 1162.
11. Peng, Z. (2019). Stocks analysis and prediction using big data analytics. In *International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE* (pp. 569–572).
12. Site, A., Birant, D., & Isik, Z. (2019). Stock market forecasting using machine learning models. In *Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE* (pp. 1–6).
13. Dingli, A., & Fournier, K. S. (2017). Financial time series forecasting—A machine learning approach. *Machine Learning and Applications: An International Journal, 4*(1/2), 3–13.
14. Balaji, A. J., Ram, D. H., & Nair, B. B. (2018). Applicability of deep learning models for stock price forecasting: An empirical study on Bankex data. *Procedia Computer Science, 143*, 947–953.

15. Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia Computer Science, 167*, 599–606.
16. Suzgun, M., Belinkov, Y., & Shieber, S. M. (2018). On evaluating the generalization of LSTM models in formal languages. *arXiv preprint arXiv:1802.08770*.
17. Charles, A., Simon, K., & Daniel, A. (2008). Study on effect of exchange rate volatility with reference to Ghana Stock Exchange. *African Journal of Accounting, Economics, Finance and Banking Research, 3*(3), 28–47.
18. Jayakumar, D. S., & Sultan, A. (2013). Testing the weak form efficiency of Indian stock market with special reference to NSE. *Advances in Management, 6*(9), 18–26.
19. Bollen, J., Mao, H., & Zeng, X. (2011). *Twitter mood predicts the stock market*. Journal of Computational Science, 2(1), 1–8.
20. Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2016). *Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media*. The Review of Financial Studies, 27(5), 1367–1403.