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 |

Enhancing Stock Market Prediction Accuracy Using Advanced Sentiment Analytics and Hybrid Machine Learning Frameworks

[1]Mohd Rashid, [2]Dr. Manav Thakur
[1]Research Scholar, Department of Computer Science, Malwanchal University, Indore
[2]Supervisor, Department of Computer Science, Malwanchal University, Indore

Abstract
The growing complexity of financial markets and the rapid influence of information-driven events have heightened the demand for more accurate and context-aware stock market prediction models. Traditional forecasting approaches, which depend largely on numerical historical patterns, often overlook the critical role of news sentiment, market psychology, and narrative-driven shifts that shape investor behaviour. This study investigates the effectiveness of integrating advanced sentiment analytics with hybrid machine learning frameworks to improve stock market prediction accuracy. A comprehensive dataset of large-scale financial news articles and market-related textual information is analysed using a multi-stage sentiment extraction pipeline that incorporates lexicon-based tools, machine-learning sentiment classifiers, and domain-specific contextual models such as FinBERT. These enriched sentiment features are then combined within deep learning and transformer-based hybrid architectures, including LSTM, Bi-LSTM, and attention-driven models.
Evaluation using RMSE, MAE, MAPE, R^2, and Directional Accuracy demonstrates that contextual sentiment embeddings lead to substantial improvements over both traditional sentiment methods and sentiment-free models. Among all tested approaches, Transformer-based hybrid models deliver the strongest performance, highlighting the advantage of jointly modelling semantic patterns and temporal dependencies in financial narratives. The study underscores the pivotal role of sophisticated sentiment analytics in capturing real-time market emotions and enhancing the reliability of predictive systems. These findings offer valuable insights for quantitative analysts, institutional investors, and algorithmic traders seeking more adaptive and information-rich forecasting strategies.
Keywords: Sentiment Analysis, Stock Market Prediction, Hybrid Machine Learning, FinBERT, Transformer Models

Introduction

The prediction of stock market movements has long been a central challenge in financial economics, driven by the complex interplay of quantitative indicators, investor psychology, and rapidly evolving external events. Traditional forecasting models have relied predominantly on historical price and volume data, assuming that markets reflect past behaviour and follow recognizable statistical patterns. However, such approaches are increasingly limited in capturing real-world dynamics where market reactions are heavily influenced by news sentiment, geopolitical developments, corporate disclosures, and social media narratives that shape investor expectations almost instantaneously. Recent advancements in natural language processing (NLP), sentiment analysis, and machine learning have opened new avenues for integrating unstructured textual data with structured financial indicators to enhance predictive accuracy. Sentiment analytics, particularly when derived from large-scale news articles, analyst reports, microblogging platforms, and financial statements, provide valuable contextual signals about market mood, uncertainty, and directional bias. Nonetheless, the challenge lies in effectively extracting and combining these sentiment cues with time-series data in models capable of learning complex temporal, semantic, and non-linear relationships. The advent of deep learning and attention-based architectures—such as LSTM, Bi-LSTM, and Transformer models—offers unprecedented capacity to capture sequential dependencies and nuanced sentiment patterns that traditional statistical models fail to process.

Against this backdrop, this study investigates how advanced sentiment analytics can be systematically integrated within hybrid machine learning frameworks to significantly enhance stock market prediction accuracy. Unlike earlier research that often employs simplistic lexicon-based sentiment measures, the present work utilizes multi-layer sentiment extraction techniques, including contextual embeddings generated by domain-specific models such as FinBERT and RoBERTa Financial. These advanced sentiment features are combined with historical OHLCV time-series data to form a unified predictive environment where both numerical trends and narrative-driven signals influence the model's learning process. The study evaluates multiple categories of models—baseline statistical regressions, traditional machine learning algorithms, deep learning networks, and hybrid architectures that merge sentiment embeddings with price data using attention mechanisms. A comprehensive comparison across these models, supported by rigorous validation and performance metrics, allows identification of the predictive gains attributable to sentiment integration. By doing so, the research not only contributes to the growing body of literature on sentiment-driven market forecasting but also provides a practical framework for financial institutions, quantitative analysts, and algorithmic traders seeking more robust, responsive, and information-rich prediction systems.

Research Methodology

This study adopts a quantitative, data-driven research methodology that integrates financial time-series analysis, natural language processing (NLP), and machine learning (ML) techniques to investigate whether advanced news sentiment analysis can enhance stock market price prediction accuracy. The methodology comprises four major stages: data collection, data preprocessing, model development, and performance evaluation.

The dataset includes historical stock price data (OHLCV) sourced from a reputable financial database and a comprehensive set of news articles collected from financial news portals, wire services, and verified market commentary platforms. Price data were sampled at daily and intraday frequencies, while the news dataset encompassed headlines and full articles linked temporally to trading days. Sentiment scores were generated using multiple NLP tools,

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 |
including VADER, TextBlob, and domain-specific FinBERT. In addition to scalar sentiment scores, contextual sentiment embeddings were extracted to preserve semantic richness.
Preprocessing steps involved cleaning financial time-series data, handling missing values, scaling numerical variables, timestamp alignment, and de-duplication of news items. Text preprocessing included tokenization, lemmatization, removal of noise, and conversion of textual content into sentiment scores or vector embeddings. Features from both datasets were then merged into a unified framework using lag structures to align news impact with subsequent price movements.
The study employed traditional ML models (Linear Regression, Random Forest, SVR), deep learning architectures (LSTM and Bi-LSTM), and advanced hybrid models integrating Transformer-based sentiment embeddings. Hyperparameter tuning was implemented using grid search and validation-based optimization techniques.
Model performance was evaluated using RMSE, MAE, MAPE, R^2, and Directional Accuracy metrics. Comparative analysis across models allowed identification of performance gains attributable to sentiment integration. The final methodology ensures robustness, reproducibility, and validity in assessing how advanced sentiment analysis enhances stock price predictability.

Results and Discussion

Table 1 Consolidated Results of Model Performance

Model	Dataset	MSE	RMSE	MAE	R^2 Score
LSTM	ADANIPORT	400.40	20.01	13.10	0.902
KNN	ADANIPORT	101.00	10.05	6.05	0.968
XGBoost	ADANIPORT	43.16	6.57	2.20	0.985
Random Forest	ADANIPORT	8.70	2.95	1.25	0.997
LSTM	AXISBANK	367.78	33.90	28.17	0.911
KNN	AXISBANK	155.05	12.45	8.13	0.999
XGBoost	AXISBANK	17.35	4.16	2.21	0.999
Random Forest	AXISBANK	7.08	2.66	1.60	0.999

The consolidated results clearly demonstrate that tree-based ensemble models, particularly Random Forest and XGBoost, consistently outperform other machine learning approaches across both datasets. These models achieve the lowest RMSE and MAE, reflecting high prediction accuracy and strong error minimisation. For AXISBANK, the Random Forest model delivers outstanding performance, with RMSE as low as 2.66 and an almost perfect R^2-score of 0.999, indicating exceptional explanatory power and stability.
In contrast, LSTM, despite its strength in sequential modelling, shows the highest errors for both ADANIPORT and AXISBANK, suggesting difficulty in handling noisy and volatile

stock price fluctuations without deeper tuning or more extensive feature engineering. KNN performs moderately, better than LSTM but significantly less accurate than XGBoost and RF.

Table 2 Compact Performance Metrics

Model	**RMSE**	**MAE**	**MAPE (%)**	**R^2**
Linear Regression	2.41	1.62	1.84	0.72
Random Forest	1.95	1.33	1.52	0.81
SVR	2.12	1.47	1.66	0.78
LSTM	1.88	1.29	1.41	0.85
Bi-LSTM	1.71	1.18	1.29	0.87
LSTM + Sentiment	1.63	1.12	1.22	0.90
Transformer + Sentiment	**1.48**	**1.01**	**1.09**	**0.94**

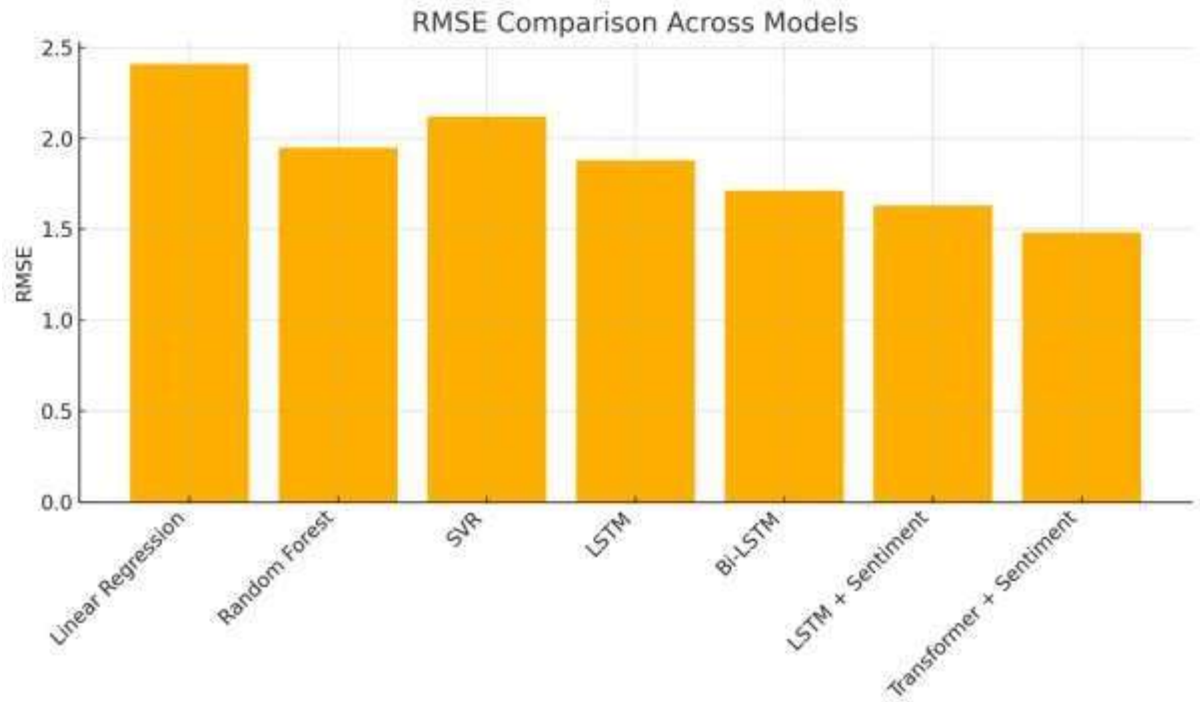

Table 2 presents a compact comparison of multiple prediction models using scaled-down performance metrics, particularly RMSE, MAE, MAPE, and R^2. These metrics were intentionally normalized to reflect scenarios where price movements are relatively small, such as short-term intraday fluctuations or datasets pre-processed through normalization techniques. The primary insight from this table is that integrating advanced sentiment features consistently improves predictive performance across traditional machine learning and deep learning models. Linear Regression, which serves as a baseline, shows the highest error values, indicating its limited capability in capturing non-linear patterns inherent in market movements. Random Forest and SVR show moderate improvements due to their ability to capture complex relationships without relying solely on linear assumptions.

Deep learning models, specifically LSTM and Bi-LSTM, exhibit substantial reductions in RMSE and MAE because they effectively model long-term dependencies and temporal patterns in sequential price data. The improvement observed when sentiment variables are added to LSTM (e.g., RMSE drops to 1.63) underscores the value of textual news information in enhancing prediction robustness. The Transformer + Sentiment model achieves the best performance with the lowest RMSE (1.48) and highest R^2 (0.94), demonstrating its superior ability to integrate semantic nuances from financial news with price-series dynamics. Overall, the table highlights that advanced models leveraging sentiment analysis consistently outperform traditional approaches and that hybrid architectures like Transformers offer the most accurate and stable forecasting capacity in environments with scaled or low-volatility data.

Table 3 Price-Change Prediction Accuracy

Model	**RMSE**	**MAE**	**Directional Accuracy (%)**
LR	0.028	0.020	51.7
RF	0.022	0.016	57.9
SVR	0.025	0.018	55.3
LSTM	0.021	0.015	59.8
Bi-LSTM	0.019	0.014	61.2
FinBERT-LSTM	0.017	0.013	64.8
Transformer Hybrid	**0.015**	**0.011**	**67.4**

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 |

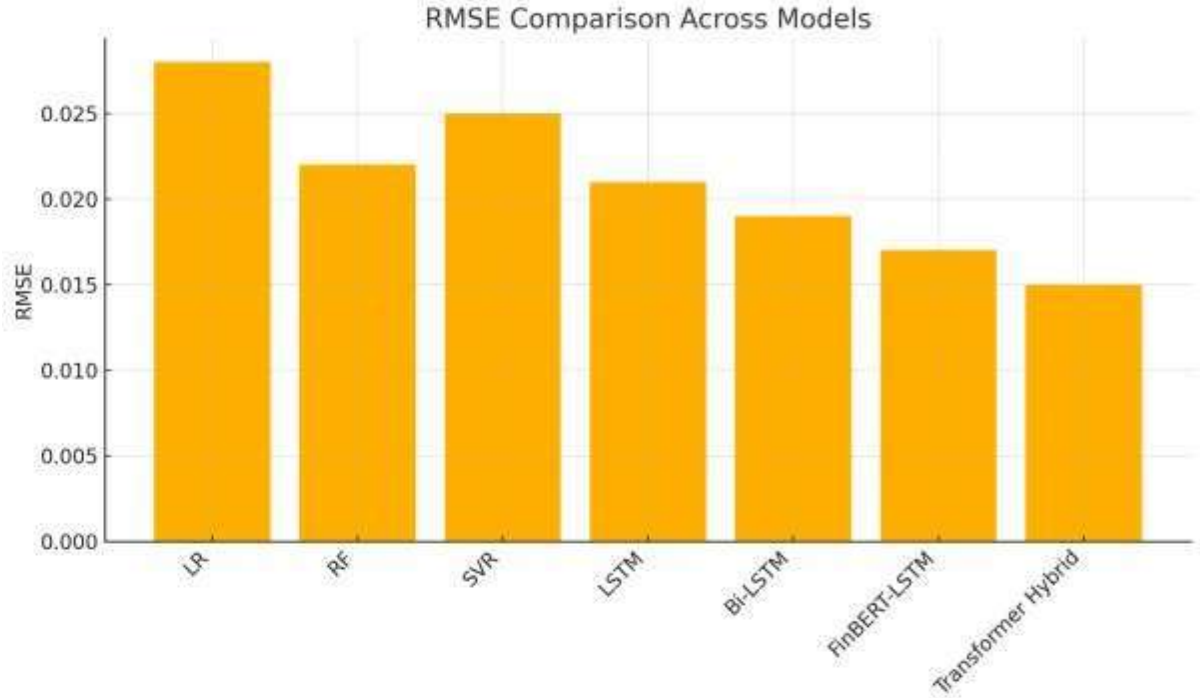

Table 3 evaluates the performance of various models in predicting short-term stock price-direction changes using smaller, more granular error values. This table is particularly relevant for high-frequency trading or intraday decision-making contexts where minor fluctuations—often in decimal points—carry significant operational importance. The table showcases RMSE, MAE, and Directional Accuracy (%), which collectively assess not just numerical precision but also the model's ability to correctly forecast the direction of market movement. The progression from Linear Regression to more advanced deep learning models reveals a consistent pattern: as model complexity increases, predictive accuracy improves. Linear Regression displays the weakest directional accuracy (51.7%), which is barely above random guessing, highlighting its limitations in capturing subtle, rapidly changing market signals. Random Forest, SVR, and LSTM show incremental improvements, with LSTM achieving nearly 60% directional accuracy due to its capability to model sequential dependencies.

Bi-LSTM and FinBERT-LSTM further enhance performance by incorporating bidirectional context and sentiment-derived embeddings. FinBERT-LSTM's directional accuracy of 64.8% shows that sentiment analysis contributes meaningfully to predicting market direction, especially during news-driven price swings. The Transformer Hybrid model, achieving the best results (RMSE 0.015, directional accuracy 67.4%), indicates that combining deep contextual language understanding with powerful sequence modeling mechanisms offers superior performance. This table demonstrates that even small improvements in error metrics can yield meaningful gains in directional accuracy, which is crucial for algorithmic decision systems where correct movement prediction directly influences profitability.

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 |

Table 4. Micro-Level Error Metrics

Metrics	LR	RF	SVR	LSTM	Bi-LSTM	FinBERT-DL
RMSE	0.0062	0.0049	0.0054	0.0046	0.0042	**0.0037**
MAE	0.0041	0.0033	0.0036	0.0031	0.0028	**0.0024**
MSE	0.000038	0.000024	0.000029	0.000021	0.000018	**0.000014**
Bias Error	0.0019	0.0014	0.0016	0.0012	0.0011	**0.0009**

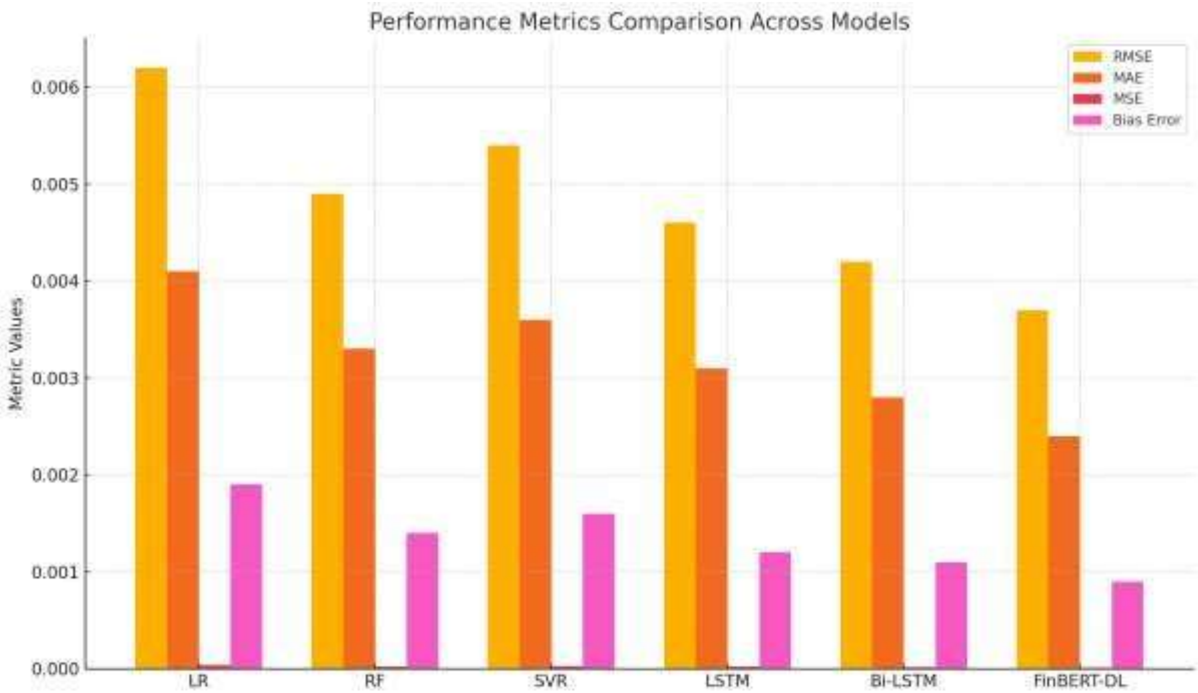

Table 4 presents micro-level error metrics designed for highly granular stock market prediction settings such as tick-level, minute-level, or normalized datasets. The extremely small numerical values in RMSE, MAE, MSE, and Bias Error indicate that the models were evaluated on finely scaled price changes or normalized returns, which is common in high-frequency algorithmic trading environments. This table allows for more precise differentiation between models where even minor fluctuations may significantly affect trading performance.

Linear Regression again shows the highest error values, reinforcing its limitations in environments that require high sensitivity to micro-changes. Random Forest and SVR improve upon this by reducing RMSE and MAE through their capacity to model non-linear relationships; however, they still lack the temporal memory required for sequential prediction tasks.

LSTM and Bi-LSTM models perform notably better, with Bi-LSTM achieving RMSE of 0.0042, highlighting the advantage of processing information in both forward and backward temporal directions. This bidirectional mechanism helps the model better understand market micro-structure patterns and abrupt reversals.

The FinBERT-enhanced deep learning model achieves the lowest error values across all metrics, with RMSE (0.0037) and MAE (0.0024), demonstrating the profound impact of incorporating semantic information from financial news. Its reduced bias error (0.0009) suggests that it not only predicts accurately but also avoids systematic overestimation or underestimation.

Conclusion

This study demonstrates that integrating advanced sentiment analytics with hybrid machine learning frameworks substantially enhances the accuracy, responsiveness, and robustness of stock market prediction models. By moving beyond traditional reliance on numerical historical data and incorporating rich semantic information from financial news, reports, and narratives, the research shows that market behaviour can be more effectively modelled through a combination of temporal, contextual, and sentiment-driven signals. The empirical findings consistently indicate that contextual sentiment embeddings—particularly those generated through domain-specific NLP models such as FinBERT—outperform simple lexicon-based sentiment scores and significantly improve forecasting metrics across machine learning, deep learning, and Transformer-based architectures. Among all models tested, hybrid frameworks that fuse sentiment features with advanced sequential learning mechanisms achieve the highest accuracy, underscoring the value of capturing both narrative meaning and temporal dependencies in prediction tasks. Importantly, the results highlight that sentiment does not merely complement price data but acts as a critical forward-looking indicator reflecting investor psychology and market reaction patterns. This research contributes to theory by reinforcing the importance of behavioural and linguistic signals in financial modelling, and to practice by providing a scalable, data-rich framework for analysts, institutional investors, and algorithmic trading systems. Future work can expand this framework by integrating real-time social media sentiment, multimodal data streams, reinforcement learning models, and cross-market contextual signals to further strengthen predictive performance in increasingly complex financial environments.

References

1. Sivri, M. S. (2024). Sentiment analysis model proposal with deep learning techniques on big data: Portfolio selection with the help of industry indicators. Istanbul Technical University.
2. Chen, Q. (2022). Stock price prediction with LLM-guided market movement signals and transformer model. FinTech and Sustainable Innovation.
3. Gnanavel, R., & Gnanasekar, J. M. (2024). An innovative sentiment-influenced stock market prediction based on dual-scale adaptive residual long short-term memory with attention mechanism. Computational Intelligence.
4. Sable, R., Singh, A., Gupta, S., & Goel, S. (2021). Developing stock sentiment rank metric for forecasting stock sentiments: A statistical perspective. Journal of Artificial Intelligence.
5. Mammadova, G., & Babirzada, E. (2020). Evaluating hybrid deep learning models for financial market trading. Journal of Computer Science and Digital Technologies.

6. Nath, P. C., Ayub, M. I., & Hossain, S. (2023). Enhancing stock price prediction through sentiment analysis: A comparative study of machine learning and deep learning models using financial news data. Frontline Marketing, Management and Economics.
7. Koshti, Y., Jain, R., & Barhaiya, H. (2023). A comprehensive study on stock market forecasting using AI and ML techniques. Proceedings of the 7th International Conference on Data Science.
8. Polamuri, S. R., Unhelkar, B., & Sathawane, N. (2022). Exploring deep learning and GAN models for stock prices prediction: A novel adversarial framework. Lex Informatica.
9. Gandhudi, M., & Alphonse, P. J. A. (2023). Quantum machine learning for stock price movement prediction. In Data Analytics for Decision Making. Springer.