A Hybrid Stock Price Forecasting System Using ARIMA, LSTM, And Sentiment Analysis

Reshma RB[1], Swathika P[2], Sackcini M[3], Saadana R[4]
[1, 2, 3, 4] Dept of Computer Science and Engineering
[1, 2, 3, 4] Vivekanandha College of Technology for Women, (Autonomous), Tiruchengode

Abstract- *The accurate forecasting of stock prices remains one of the most persistent and challenging problems in quantitative finance, owing to the inherently noisy, nonstationary, and nonlinear nature of financial time series. Traditional forecasting methods struggle with the dual challenges of capturing both linear periodicities and nonlinear behavioral dynamics within a unified framework. This paper presents a combined approach to enhance stock price forecasting accuracy. The proposed model integrates three key components: Autoregressive Integrated Moving Average (ARIMA) for effectively modeling linear patterns and repeating cycles in stock price data; Long ShortTerm Memory (LSTM) networks for identifying complex, longterm dependencies in stock price movements; and market news sentiment analysis to capture investor behavior and emotional influences. By synergistically combining statistical timeseries modeling, deep learning, and sentimentdriven insights, the approach aims to overcome the limitations of individual methods. Experimental evaluation on three largecap stocks (AAPL, JPM, TSLA) over a 12month testing period demonstrates that the hybrid model achieves a 41.9% reduction in RMSE compared to standalone ARIMA and a 30.1% reduction compared to standalone LSTM, with directional accuracy improving from 52.9% and 59.9% to69.5%. The hybrid framework provides more robust and precise stock price predictions, supporting better informed financial decision making.*

Keywords: Stock Price Forecasting, ARIMA, LSTM, Sentiment Analysis, Hybrid Model, Time Series Prediction

I. INTRODUCTION

The accurate forecasting of stock prices remains one of the most persistent and challenging problems in quantitative finance, owing to the inherently noisy, nonstationary, and nonlinear nature of financial time series. Stock markets are influenced by a complex interplay of historical price patterns, macroeconomic indicators, corporate fundamentals, and—critically—investor sentiment driven by news and social media. According to recent estimates, over $90 trillion is traded annually across global equity markets, with even a 1% improvement in forecasting accuracy potentially translating into billions of dollars in optimized trading strategies and risk mitigation. Yet, traditional forecasting methods continue to struggle with the dual challenges of capturing both linear periodicities and nonlinear behavioral dynamics within a unified framework.

Conventional stock price forecasting approaches typically fall into two categories: statistical timeseries models and standalone neural networks. Statistical models such as Autoregressive Integrated Moving Average (ARIMA) excel at capturing linear patterns, trends, and repeating seasonal cycles in historical price data. ARIMA has been widely adopted for its mathematical tractability and interpretability, achieving reasonable baseline performance under stable market conditions. However, ARIMA inherently assumes linearity and stationarity after differencing, making it illsuited for detecting complex, longrange dependencies or abrupt regime shifts caused by breaking news or earnings surprises. In such scenarios, ARIMA forecasts often lag actual price movements by several time steps, leading to delayed trading signals and missed opportunities.

Conversely, deep learning architectures—particularly Long ShortTerm Memory (LSTM) networks—have demonstrated superior capability in identifying complex, nonlinear patterns and longterm dependencies in sequential data. Unlike ARIMA, LSTM can retain information over 100 or more time steps through its gated memory cell structure, making it effective at modeling volatility clustering, momentum effects, and mean reversion. Nevertheless, LSTM alone suffers from two critical limitations: it is purely datadriven, lacking any mechanism to incorporate exogenous qualitative information, and it tends to overfit during volatile periods when price movements are driven primarily by investor emotion rather than historical patterns.

A third dimension—often overlooked in purely pricebased models—is the role of market news sentiment in shaping investor behavior and emotions. Financial markets are not efficient in the strict rational sense; instead, fear, greed, overreaction, and herding behavior drive 30–40% of shortterm price fluctuations. Sentiment extracted from news headlines, earnings call transcripts, and social media platforms provides a

behavioral lens that purely numerical models cannot access. However, integrating sentiment analysis with timeseries forecasting remains nontrivial: sentiment must be temporally aligned with price movements, aggregated across multiple sources, and fused with historical features without introducing latency or noise.

To bridge these gaps, this paper proposes a hybrid approach that integrates ARIMA, LSTM, and market news sentiment analysis into a unified stock price forecasting framework. The proposed system operates as follows: ARIMA models the linear trend and seasonal components of historical price series; LSTM captures residual nonlinear patterns and longterm dependencies; and a sentiment scoring module quantifies investor emotions from realtime news streams. The three components are fused through a sequential residual correction mechanism.

II. LITERATURE REVIEW

Early efforts in stock price forecasting relied heavily on classical statistical methods that assume linearity and stationarity. Box and Jenkins laid the foundation for ARIMA modeling, which remains widely used due to its mathematical interpretability and strong performance on trended and seasonal data. Adebiyi et al. compared ARIMA with neural networks on Nigerian Stock Exchange data, demonstrating that ARIMA effectively captured linear autocorrelations and repeating weekly patterns, achieving a mean absolute percentage error (MAPE) of approximately 8–12% under stable market conditions. However, ARIMA's forecasting accuracy degrades sharply during earnings seasons or macroeconomic announcements when price movements become nonlinear and sentimentdriven.

As computational resources grew, researchers turned to recurrent neural networks (RNNs) to model sequential dependencies without linearity assumptions. Hochreiter and Schmidhuber proposed the Long ShortTerm Memory (LSTM) network, specifically designed to overcome the vanishing gradient problem in standard RNNs, enabling the model to retain information over 100+ time steps. Chen et al. applied LSTM to S&P 500 index prediction, reporting a 15–22% reduction in root mean square error (RMSE) compared to ARIMA, with particular improvement during highvolatility periods. Baek and Kim reported that standalone LSTM models exhibit 30–40% higher prediction error during eventdriven selloffs compared to calm periods, indicating a need for supplementary behavioral signals.

Recognizing that financial markets are not perfectly rational, researchers began integrating textual sentiment to capture the emotional drivers of price movements. Bollen et al. famously demonstrated that Twitter sentiment could predict DJIA movements with 87.6% accuracy, highlighting the predictive power of collective investor emotions. Ma et al. compared CNNbased sentiment classifiers with lexicon methods on Bloomberg headlines, showing that contextual embeddings (BERT) improved F1 scores by 12–18%. However, sentiment integration remains nontrivial. Multiple studies identified three persistent challenges: sentiment aggregation across heterogeneous sources requires careful normalization; sentiment effects often lag price movements by minutes to hours; and sentimentbased models are prone to noise during periods of low news volume.

A limited number of studies have attempted to combine all three paradigms. The most relevant work by Liu and Wu introduced a threestage model: ARIMA for trend decomposition, LSTM for nonlinear residuals, and a sentiment adjustment factor computed from Reuters headlines. They reported a 14% reduction in MAPE compared to LSTM alone. Despite this progress, several gaps remain. Most hybrid models treat sentiment as a static additive factor rather than a timevarying signal with decay effects. Existing frameworks rarely evaluate performance across distinct market regimes (bull, bear, sideways). Finally, the interpretability of hybrid predictions—essential for risk management—remains largely unaddressed.

III. PROBLEM STATEMENT

Despite the availability of various stock price forecasting techniques, existing systems still face several critical challenges. Traditional statistical models like ARIMA are heavily dependent on linearity assumptions, which limits their ability to capture nonlinear dependencies and abrupt regime shifts caused by breaking news or earnings surprises. As financial markets become increasingly complex and sentimentdriven, this limitation becomes increasingly significant.

Another major issue is the inability of standalone LSTM models to incorporate exogenous qualitative information such as news sentiment, analyst reports, or social media trends. This leads to poor performance during eventdriven volatility where price movements are driven primarily by investor emotion rather than historical patterns. Additionally, pure sentimentbased models lack temporal price dynamics and often produce noisy signals during periods of contradictory news.

Handling the fusion of heterogeneous data streams—price data at different frequencies, technical indicators, and

unstructured text—is also a challenge, as conventional systems struggle with temporal alignment and feature integration. Therefore, there is a need for an intelligent, hybrid, and sentimentaware forecasting system that can accurately predict stock prices by jointly modeling linear patterns, nonlinear dependencies, and behavioral context.

IV. PROPOSED SYSTEM

The proposed system is designed to provide an efficient and intelligent solution for stock price forecasting using a hybrid machine learning approach. It focuses on integrating multiple data sources and modeling paradigms to improve prediction accuracy and robustness.

The architecture of the system consists of multiple modules that work together to ensure accurate forecasting. The data acquisition layer gathers historical price data from financial APIs, realtime market feeds, and news sentiment sources. The preprocessing and feature engineering layer handles data cleaning, missing value imputation, outlier detection, and temporal alignment across different data streams. The feature engineering module computes technical indicators such as moving averages, RSI, and MACD to augment the input feature set.

The hybrid forecasting engine hosts three core modules. The ARIMA module fits the linear trend and seasonal components of the closing price series. The LSTM module, implemented as a stacked network with dropout regularization, learns residual nonlinear patterns and longterm dependencies. The sentiment integration module processes news headlines and social media content using a FinBERT model to generate timedecayed sentiment scores. The three components are fused through a sequential residual correction mechanism where ARIMA predictions are first generated, LSTM corrects the residuals, and sentiment serves as an adaptive adjustment factor.

The decision support layer provides final price predictions, confidence intervals, and trading signals to endusers. The system is designed to operate in a closed loop: predictions are compared with actual prices, and prediction errors are fed back to retrain the LSTM periodically.

V. METHODOLOGY

The methodology adopted in this system involves several important steps to ensure accurate and efficient stock price forecasting.

5.1 Data Preprocessing

Data preprocessing is a crucial step in the forecasting pipeline. Raw price data often contains missing values, outliers, and nonstationarity that can affect model performance. Missing values are handled using forwardfill for gaps of up to 3 days, followed by linear interpolation. Outlier detection is performed using the Zscore method, and values exceeding three standard deviations are capped. Stationarity is tested using the Augmented DickeyFuller test, and firstorder differencing is applied when necessary. For text data, raw news headlines undergo tokenization, stop word removal, and lemmatization before sentiment scoring.

5.2 Feature Engineering

Feature engineering helps in creating informative attributes that improve model performance. Technical indicators including Simple Moving Average (SMA), Exponential Moving Average (EMA), Relative Strength Index (RSI), and MACD are computed from raw price data. Sentiment features include the aggregate sentiment score for the current day, lagged sentiment scores for the previous two days, and the sentiment change (delta) between consecutive days. Dayofweek dummy variables are also added to capture weekly seasonality. All features are normalized using MinMax scaling to the range [0,1].

5.3 Hybrid Forecasting Models

The system uses three complementary modeling techniques. ARIMA(p,d,q) is used to model linear trends and seasonal components, with optimal parameters selected via grid search using the Akaike Information Criterion. A stacked LSTM network with three LSTM layers (128, 64, and 32 units) and dropout regularization is used to capture nonlinear dependencies and longterm patterns. Sentiment scoring is performed using a FinBERT model for contextaware classification, with VADER as a lightweight fallback. The final prediction is computed as: $\hat{F}r = (\hat{A}r + \hat{L}r) \times (1 + \gamma \cdot Mr)$, where Mr is a sentiment multiplier derived from current sentiment and sentiment change.

5.4 Model Evaluation

To evaluate the performance of the system, several metrics are used. Root Mean Square Error (RMSE) measures the magnitude of prediction errors. Mean Absolute Percentage Error (MAPE) provides error as a percentage of actual values. Directional Accuracy measures the percentage of times the model correctly predicts the direction of price movement (up or down). These metrics are computed across the entire test period and separately for different market regimes.

VI. IMPLEMENTATION

The implementation of the proposed system is carried out using modern tools and technologies. The backend is developed using Python 3.9, which provides powerful libraries for time series analysis, deep learning, and natural language processing. Pandas and NumPy are used for data manipulation, statsmodels for ARIMA implementation, and TensorFlow 2.12 with Keras for LSTM network construction. Transformers library by Hugging Face is used for FinBERT sentiment analysis.

Flask is used to create APIs that enable communication between the backend and frontend. The frontend is developed using Flutter and Dart, providing a userfriendly interface for data input, visualization of historical prices, sentiment trends, and forecasted values. All experiments were conducted on an NVIDIA T4 GPU with 16 GB RAM to accelerate LSTM training.

The modular design of the system ensures flexibility and scalability, allowing it to be extended to additional stocks, data sources, or forecasting horizons based on future requirements.

VII. RESULTS AND DISCUSSION

The experimental results demonstrate that the proposed hybrid system performs effectively in forecasting stock prices. The use of complementary modeling techniques significantly improves forecasting accuracy compared to individual methods. The hybrid model achieved a 41.9% reduction in RMSE compared to standalone ARIMA (10.07 to 5.85) and a 30.1% reduction compared to standalone LSTM (8.37 to 5.85) across three test stocks.

Directional accuracy improved from 52.9% for ARIMA and 59.9% for LSTM to 69.5% for the proposed hybrid model. The ablation study revealed that LSTM removal caused the largest performance degradation (42.2% higher RMSE), followed by sentiment removal (21.0% higher RMSE), confirming that nonlinear dependency modeling is the most critical component while sentiment provides substantial added value. Regimespecific analysis indicated that sentiment contributed most significantly during earnings seasons (28% RMSE improvement) and least during sideways markets (8% improvement). Among the three stocks, the hybrid model performed best on AAPL (RMSE 1.98) and most challenged on TSLA (RMSE 10.82), consistent with Tesla's higher inherent volatility.

The results indicate that the system is suitable for daily to hourly trading applications, with a computational profile of 48 minutes for weekly training and 58 milliseconds for inference per prediction.

VIII. ADVANTAGES

Integrates linear (ARIMA), nonlinear (LSTM), and behavioral (sentiment) modeling within a single framework

Achieves 41.9% lower RMSE than standalone ARIMA and 30.1% lower than standalone LSTM Improves directional accuracy from ~53–60% to nearly 70%, enabling better trading signals

Incorporates realtime news sentiment with timedecayed aggregation for timely adjustments

Provides regimespecific robustness, performing well in bull, bear, sideways, and earnings volatility periods

Reduces reliance on manual feature engineering through automated LSTM pattern learning

Offers interpretable contributions from each component via ablation analysis

Scalable architecture suitable for multistock portfolios and daily retraining cycles

IX. CONCLUSION

This paper presents a hybrid stock price forecasting system that integrates ARIMA, Long ShortTerm Memory (LSTM) networks, and market news sentiment analysis into a unified predictive framework. The proposed system overcomes the limitations of individual methods by jointly modeling linear trends, nonlinear dependencies, and behavioral context. Experimental evaluation on three largecap stocks (AAPL, JPM, TSLA) over a 12month testing period demonstrates that the hybrid model achieves a 41.9% reduction in RMSE compared to standalone ARIMA and a 30.1% reduction compared to standalone LSTM, with directional accuracy improving from 52.9% and 59.9% to 69.5%. The results demonstrate that the system is effective, scalable, and suitable for daily to hourly trading applications. It provides a strong foundation for developing advanced financial forecasting solutions in modern equity markets.

X. FUTURE WORK

Future work can focus on extending the model to multiasset forecasting using graph neural networks that capture crossstock correlations and sector influences. Adaptive source weighting for sentiment aggregation can be implemented using reinforcement learning to dynamically

adjust the reliability of different news sources. Model compression techniques such as knowledge distillation can be explored to reduce inference latency for highfrequency trading applications. Additionally, macroeconomic indicators (interest rates, inflation, GDP) can be integrated as exogenous variables, and automated trading strategies can be developed to translate predictions into realworld portfolio decisions with risk management constraints.

REFERENCES

[1] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco, CA: HoldenDay, 1976.
[2] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction," *Journal of Applied Mathematics*, vol. 2014, pp. 1–7, 2014.
[3] P. F. Pai and C. S. Lin, "A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.
[4] R. F. Engle, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.
[5] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," in *Proc. UKSimAMSS Int. Conf. Comput. Modelling Simul.*, 2014, pp. 105–112.
[6] S. Hochreiter and J. Schmidhuber, "Long ShortTerm Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[7] K. Chen, Y. Zhou, and F. Dai, "A LSTMBased Method for Stock Price Prediction," in *Proc. IEEE Int. Conf. Softw. Eng. Service Sci.*, 2015, pp. 49–53.
[8] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8,2011.
[9] Y. Ma, S. Zhang, and L. Chen, "Sentiment Analysis for Stock Price Prediction Using BERT and CNN," *IEEE Access*, vol. 9, pp. 112345112358, 2021.
[10] T. Liu and J. Wu, "A Hybrid ARIMALSTMSentiment Model for Stock Price Forecasting," *Expert Systems with Applications*, vol. 185, pp. 115128, 2021.

This reformatted version preserves all technical content, quantitative results, and intellectual contributions of your original stock forecasting paper while matching the exact structural template, section numbering style, author block format, keyword placement, and reference numbering of your intrusion detection system paper.