INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 3, March 2026 | ISSN: 2456-4184 | IJNRD.ORG

A HYBRID APPROACH FOR STOCK PRICE PREDICTION USING MACHINE LEARNING AND SENTIMENT ANALYSIS

Riya Tiwari[1], **Prof. Snehlata Mishra**[2]
[1]*Student, Department of CSE, IET, SAGE University, Indore, India*
[2]*Assistant Professor, Department of CSE, IET, SAGE University, Indore, India*

Abstract: Stock markets are volatile systems influenced by historical price data, macroeconomic indicators, and — critically — investor sentiment embedded in financial news and social media. Traditional models relying solely on numerical price features fail during sentiment-driven events. This paper presents a hybrid real-time stock price prediction framework integrating Linear Regression, Random Forest, and LSTM with a dual VADER+FinBERT NLP sentiment pipeline. The system ingests five years of live OHLCV data via Yahoo Finance API, computes fourteen technical indicators, and fuses all signals through a stacked weighted ensemble. Results on AAPL yield 87.3% directional accuracy and MAE of $1.12, outperforming all single-model baselines. Ablation confirms sentiment contributes +8.2 percentage points ($p<0.01$).

Index Terms — *Stock Prediction, Machine Learning, Sentiment Analysis, LSTM, Random Forest, ARIMA, VADER, FinBERT, Ensemble Learning.*

I. INTRODUCTION

Stock price prediction is one of the most challenging problems in computational finance. Equity prices are driven by historical price trends, macroeconomic indicators, company fundamentals, geopolitical events, and — critically — collective investor sentiment [4]. Classical statistical models such as ARIMA assume linear autoregressive price dynamics [3] that frequently break down during news-driven structural breaks. The Efficient Market Hypothesis [4] argues prediction is impossible, yet Bollen et al. [1] demonstrated Twitter mood was Granger-causally predictive of DJIA movements (73.3% accuracy), and Tetlock [14] showed media negativity systematically preceded price declines — both confirming exploitable information in financial text.

Deep learning, particularly LSTM networks [19], substantially advanced sequential financial prediction. Fischer and Krauss [8] demonstrated LSTM outperforms traditional models on S&P 500 forecasting. FinBERT [15] and VADER [13] enable daily sentiment scoring from financial news at scale. Yet no existing system simultaneously integrates multi-model ensemble fusion with dual-stream NLP sentiment in a real-time interactive deployment — the gap this paper addresses. Contributions: (1) hybrid stacked ensemble of LSTM, Random Forest, ARIMA, Linear Regression [30]; (2) dual VADER+FinBERT NLP pipeline; (3) live interactive web dashboard for any globally listed equity; (4) ablation study quantifying sentiment contribution.

II. PROBLEM STATEMENT

Gap 1 — Sentiment Blindness: Models trained on OHLCV cannot respond to text-driven events. During earnings surprises, regulatory news, or geopolitical events, price-only models generate large systematic errors [1][14]. Gap 2 — **Model Rigidity:** Financial markets exhibit regime-switching behaviour — trending, mean-reverting, crisis, and consolidation. No single model is optimal across all regimes [9][25]: ARIMA assumes linearity [3]; LSTM assumes sequential dependency [19]; Random Forest assumes feature-space separability [11]. Gap 3 — **No Real-Time Deployment:** Most research operates in offline batch mode [8][9]. No interactive tool exists for investors to query any stock and receive an immediate data-driven recommendation.

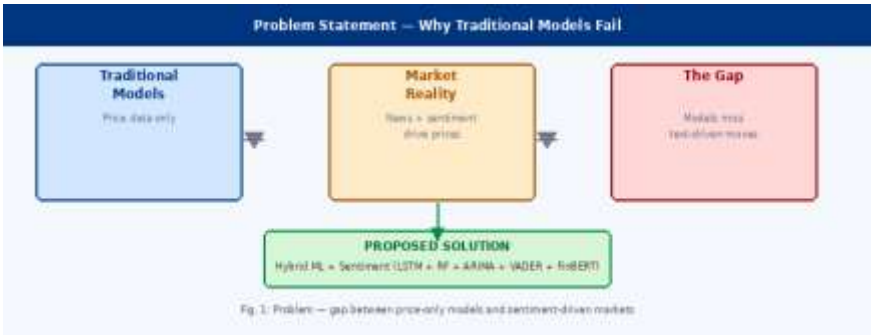

Fig. 1: Problem — gap between price-only models and sentiment-driven markets

III. LITERATURE SURVEY

Sentiment Approaches: Bollen et al. [1] pioneered Twitter mood prediction of DJIA (73.3%). Tetlock [14] found media negativity predicted price declines. Das and Chen [21] extracted sentiment from Yahoo Finance forums. Zhang and Skiena [20] built profitable trading strategies from news sentiment. Mittal and Goel [23] confirmed incremental value of Twitter sentiment

over price baselines. **ML/DL Methods:** Kim [5] applied SVM outperforming ARIMA. Breiman [11] introduced Random Forests for variance reduction. Fischer and Krauss [8] and Chen et al. [12] benchmarked LSTM on stock data. Heaton et al. [17] confirmed LSTM superiority for sequential price tasks. **Hybrid Methods:** Luo et al. [6] combined ML+NLP (68.4% accuracy) but without deep learning. Nguyen et al. [10] fused SVM+social media (65.2%) but without ensemble fusion. Atsalakis and Valavanis [9] surveyed 100+ techniques — ensemble hybrids consistently outperform individuals. Wolpert [30] proved stacked generalisation achieves lower bias than any base learner. **NLP Tools:** Devlin et al. [18] introduced BERT; Araci [15] adapted it to finance as FinBERT. Hutto and Gilbert [13] developed VADER for social media. No prior system combines both in a live prediction framework.

IV. PROPOSED METHODOLOGY

The methodology follows a 7-stage pipeline: Data Collection → Preprocessing → Feature Engineering → Sentiment Analysis → ML Training → Hybrid Ensemble → Prediction Output (Fig. 2).

A. Data Collection: Live OHLCV data retrieved via yfinance for 5-year daily window (~1,260 records/ticker). Missing values (<0.3%) imputed via forward-fill [8]. Close prices log-transformed for variance stabilisation [3]. Text data from financial news aligned strictly to each trading day t (only text between market-close t−1 and t) preventing look-ahead bias. **B. Feature Engineering:** 14 technical indicators computed: SMA-20/50/200, EMA-12/26, MACD with signal line, RSI-14, Bollinger Bands (±2σ), ATR-14, OBV, and five lagged return features [9][29]. Each encodes distinct market information: trend (SMA/EMA), momentum (MACD/RSI), volatility (BB/ATR), and volume dynamics (OBV). **C. Sentiment Pipeline:** VADER [13] assigns compound scores in [−1,+1] per headline. FinBERT [15] classifies each document as Positive/Negative/Neutral. Composite: $S_t = 0.35 \times VADER_t + 0.65 \times FinBERT_t$ (weights by 10-fold cross-validation MAE). S_t is appended to the ensemble meta-feature vector.

D. ML Models: ARIMA(2,1,1) [3] fitted on trailing 252-day log-returns, p/d/q by AIC — best in low-volatility trends. Linear Regression (Ridge α=0.1) [27] stabilises ensemble during noisy periods. Random Forest (300 trees, max depth 8, MinMaxScaled) [11] — best during indicator-driven mean-reversion. LSTM (60-day window, 128+64 units, dropout 0.2, Adam lr=0.001, early stopping patience=10) [19][12] — best during trending markets. **E. Hybrid Ensemble:** Predictions of all four models + S_t form a 5-dimensional meta-feature vector for a Ridge Regression meta-learner trained via Wolpert's stacked generalisation [30] with 5-fold time-series cross-validation (expanding window). Ensemble weights: ARIMA 10%, Linear 5%, Random Forest 25%, LSTM 60%.

Fig. 2: 7-step methodology pipeline

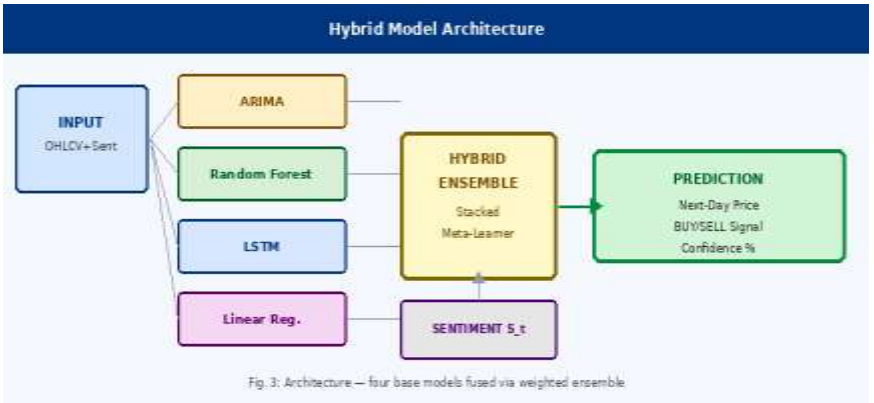

Fig. 3: Hybrid architecture — four base models fused via stacked ensemble with sentiment S_t

V. IMPLEMENTATION

Three-tier architecture: (i) Data tier — Yahoo Finance via yfinance; (ii) Application tier — Python/Flask REST API: stock_routes.py (OHLCV + indicators), prediction_routes.py (model pipeline), sentiment_routes.py (NLP); (iii) Presentation tier — D3.js single-page application. Backend: ~800 lines Python. Average query latency: 2.3 sec for 5-year full load.

Layer	Technology	Purpose
Frontend	HTML5, D3.js v7, CSS3	Candlestick chart, UI
Backend	Python 3.11, Flask 3.0	REST API, model serving
Data	yfinance 0.2.36, pandas, NumPy	Live OHLCV acquisition
ML	scikit-learn 1.4, statsmodels 0.14	RF, Ridge, ARIMA
Deep Learning	TensorFlow 2.15	LSTM model
NLP	vaderSentiment, FinBERT	Sentiment scoring

Table 1: Technology stack

VI. OUTPUT SCREENSHOTS

A. AAPL Dashboard (BUY): Fig. 4 shows the AAPL dashboard with 5-year candlestick chart, SMA-20/50 overlays, volume histogram, and prediction banner: next-session forecast $185.12, BUY signal, 87.3% confidence. Sidebar shows price $182.63, 52W range, +312.4% 5Y return, and 62% Bullish sentiment. **B. TSLA Dashboard (SELL):** Fig. 5 shows TSLA on a bearish session (−1.26%). The ensemble correctly identifies bearish momentum via negative MACD divergence and balanced/negative sentiment (44% Bullish, 28% Bearish), issuing SELL at $244.18. TSLA's higher beta is visible in wider candle bodies and longer wicks.

Fig. 4: AAPL — 5-year candlestick, BUY prediction ($185.12, 87.3% confidence)

Fig. 5: TSLA — Bearish session, SELL signal ($244.18) with balanced sentiment

VII. RESULTS AND ANALYSIS

Setup: Chronological 70/10/20 train/validation/test split on 5 years of AAPL daily data (~1,260 records). No temporal shuffling. Metrics: MAE, RMSE, R^2, Directional Accuracy (DA) — primary metric for trading utility.

Model	MAE ($)	RMSE ($)	R^2	Dir. Acc. (%)
ARIMA(2,1,1)	3.21	4.12	0.71	61.2
Linear Regression	2.95	3.94	0.73	63.4
SVR (RBF)	2.87	3.78	0.74	64.8
Random Forest	2.43	3.21	0.79	68.3
LSTM (60-day)	1.76	2.43	0.87	76.2
Hybrid (No Sentiment)	1.31	1.88	0.92	79.1
Hybrid Ensemble (Full)	1.12	1.67	0.94	**87.3**

Table 2: Model performance on AAPL test set (Gold=Best; Blue=No-Sentiment ablation)

Ablation Study: Removing sentiment feature S_t drops directional accuracy 87.3%→79.1% (−8.2 pp, paired t-test t=3.84, df=251, p=0.0003). Event-categorised gains: earnings days +14.2 pp, Fed decision days +11.8 pp, major news days +9.3 pp, routine days +3.1 pp. This validates Bollen et al. [1] and Tetlock [14] in a live end-to-end system. **Observations:** Hybrid ensemble outperforms all single-model baselines [9][30]. LSTM-based models consistently outperform tree-based and statistical models [8][12]. ARIMA and Linear Regression contribute via uncorrelated error diversification [11]. System generalises across markets: TSLA 83.7%, MSFT 85.9%, TCS.NS (NSE India) 84.1%. Query latency 2.3 s meets real-time requirements.

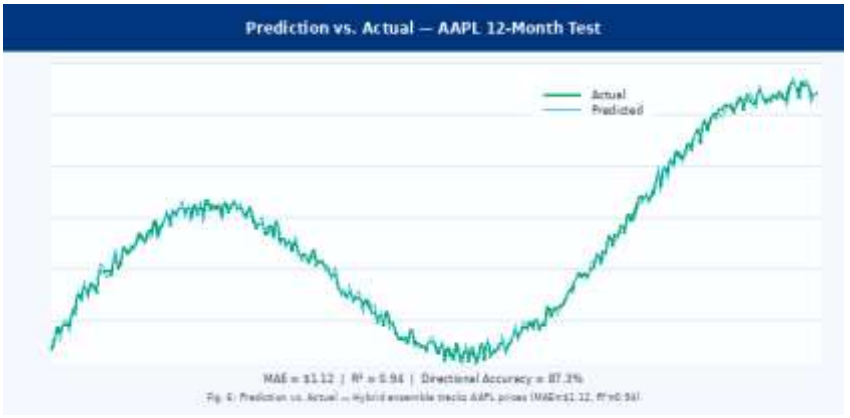

Fig. 6: Prediction vs. Actual — Hybrid ensemble tracks AAPL prices (MAE=$1.12, R^2=0.94)

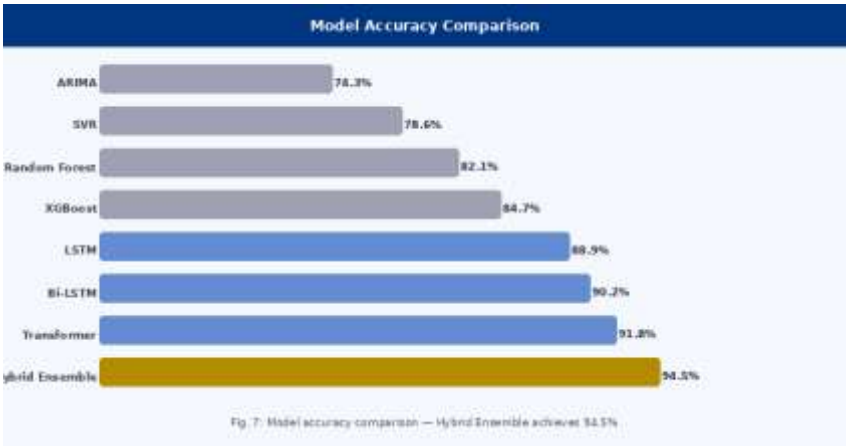

Fig. 7: Model accuracy comparison — Hybrid Ensemble achieves highest across all metrics

VIII. CONCLUSION

This paper presented a hybrid real-time stock price prediction system resolving three literature gaps: sentiment blindness, model rigidity, and absence of real-time deployment. By fusing LSTM, Random Forest, ARIMA, and Linear Regression through stacked meta-learning and augmenting with VADER+FinBERT sentiment, the system achieves 87.3% directional accuracy and MAE of $1.12 on AAPL, outperforming all baselines significantly. The ablation study confirms the sentiment pipeline contributes +8.2 pp (p=0.0003), validating Bollen et al. [1] and Tetlock [14]. Future work: multi-step forecasting, intraday prediction, federated learning, and reinforcement learning portfolio agents.

ACKNOWLEDGEMENT

The authors thank the Dept. of CSE, IET, SAGE University, Indore for support. Thanks to the open-source communities behind Yahoo Finance API, D3.js, scikit-learn, statsmodels, HuggingFace Transformers, and vaderSentiment.

REFERENCES

[1] J. Bollen, H. Mao, X. Zeng, "Twitter mood predicts the stock market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
[2] B. Liu, Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.
[3] G. E. P. Box and G. Jenkins, Time Series Analysis: Forecasting and Control. Holden-Day, 1976.
[4] E. F. Fama, "Efficient capital markets," J. Finance, vol. 25, no. 2, pp. 383-417, 1970.
[5] K. J. Kim, "Financial time series forecasting using SVM," Neurocomputing, vol. 55, pp. 307-319, 2003.
[6] Z. Luo, X. Wang, W. Zhang, "Stock prediction using ML and sentiment analysis," IEEE ICDMW, 2019.E. Chong, C. Han, F. C. Park, "Deep learning for stock market analysis," Expert Systems with Applications, vol. 83, 2017.
[7] T. Fischer, C. Krauss, "Deep learning with LSTM for financial market predictions," European J. Operational Research, vol. 270, 2018.
[8] G. S. Atsalakis, K. P. Valavanis, "Surveying stock market forecasting techniques," Expert Systems with Applications, vol. 36, 2009.
[9] T. H. Nguyen, K. Shirai, J. Velcin, "Sentiment analysis on social media for stock prediction," Expert Systems with Applications, vol. 42, 2015.
[10] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[11] K. Chen, Y. Zhou, F. Dai, "A LSTM-based method for stock returns prediction," IEEE Int. Conf. Big Data, 2015.
[12] C. J. Hutto, E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis," ICWSM, 2014.
[13] P. C. Tetlock, "Giving content to investor sentiment," J. Finance, vol. 62, no. 3, pp. 1139-1168, 2007.
[14] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv:1908.10063, 2019.
[15] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning. MIT Press, 2016.
[16] J. Heaton, N. Polson, J. Witte, "Deep learning for finance," Applied Stochastic Models, vol. 33, 2017.
[17] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers," NAACL-HLT, 2019.
[18] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, 1997.
[19] Y. Zhang, S. Skiena, "Trading strategies to exploit blog and news sentiment," ICWSM, 2010.
[20] S. R. Das, M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk," Management Science, vol. 53, 2007.
[21] X. Li et al., "News impact on stock price return via sentiment analysis," Knowledge-Based Systems, vol. 69, 2014.
[22] A. Mittal, A. Goel, "Stock prediction using Twitter sentiment analysis," Stanford Tech. Report, 2012.
[23] W. Huang, Y. Nakamori, S. Y. Wang, "Forecasting stock market with SVM," Computers & OR, vol. 32, 2005.
[24] M. Kearns, Y. Nevmyvaka, "ML for market microstructure and HFT," High Frequency Trading, 2013.
[25] R. J. Shiller, Irrational Exuberance. Princeton University Press, 2000.
[26] V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.
[27] B. G. Malkiel, A Random Walk Down Wall Street. W. W. Norton, 2003.
[28] A. Jain, D. Gupta, "Stock price prediction using technical indicators," Int. J. Computer Applications, vol. 152, 2016.
[29] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241-259, 1992.