Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 December; 6 (10s): 2433 - 2437

Machine Learning Models for Stock Market Prediction Using Sentiment and Performance Data: A Review

[1*]Santosh Raghuwanshi, [2]Satish Agnihotri

[1*]PhD Scholar Department of Mathematics, SAM Global University Raisen, (M.P.)santoshraghuwanshi122@gmail.com
[2]Assistant Professor Department of Mathematics, SAM Global University Raisen, (M.P.)Mail id- , Sattishagnihotri007@gmail.com

ABSTRACT

Stock market prediction has emerged as one of the most challenging and attractive applications of machine learning (ML) and deep learning (DL). The non-linear, dynamic, and highly volatile nature of financial markets makes traditional statistical models insufficient for accurate forecasting. Recent studies increasingly combine historical performance indicators (OHLCV, technical indicators, financial ratios) with sentiment data derived from news, Twitter, Reddit, and financial reports to improve predictive capability. This review comparatively analyzes machine learning models used between 2021 and 2023, focusing on supervised ML, deep learning, and hybrid sentiment-aware architectures. A synthesis of 10 recent published studies reveals that LSTM, Random Forest, XGBoost, and hybrid CNN-LSTM models outperform conventional regression methods, particularly when sentiment signals are integrated. However, major limitations persist in interpretability, overfitting, real-time adaptability, and transaction-cost-aware evaluation. The paper concludes with research gaps and future directions for robust stock forecasting systems.

Keywords: Stock prediction, machine learning, sentiment analysis, LSTM, Random Forest, XGBoost, financial forecasting

1. Introduction

Financial market forecasting is a complex time-series prediction problem influenced by macroeconomic factors, company fundamentals, technical trends, investor psychology, and breaking news. Recent advances in ML have enabled systems to learn hidden relationships from vast structured and unstructured data sources [1-2]. From 2021–2023, researchers increasingly shifted from pure price-based forecasting to multi-modal prediction**,** integrating:

- Historical stock prices
- Technical indicators
- Company performance ratios
- News sentiment
- Social media sentiment
- Economic indicators

Studies show that combining sentiment signals with technical features significantly improves directional prediction accuracy**,** especially in short-term trading windows.
The present review compares recent ML models and identifies:

1. Most effective algorithms
2. Data modalities used
3. Performance outcomes
4. Practical limitations
5. Future research opportunities

2. Machine Learning Models Used in Recent Stock Prediction

2.1. Traditional Machine Learning Models

Traditional machine learning (ML) models have been widely applied in stock market prediction because of their ability to learn relationships between historical stock indicators, financial ratios, and market sentiment variables [3-4]. These models are particularly effective for directional classification (up/down movement) and next-day closing price prediction.

1. **Linear Regression** is one of the most fundamental statistical learning models used for predicting continuous stock values such as next-day closing prices, returns, and percentage changes. It assumes a linear relationship between

independent variables (technical indicators, sentiment scores, volume, volatility) and the dependent stock price. Although simple and interpretable, its limitation lies in handling the highly non-linear nature of stock markets.

2. **Logistic Regression** is primarily used for binary classification tasks, such as predicting whether the stock price will move upward or downward. It converts input variables into probabilities and is widely preferred for directional prediction because of its simplicity, interpretability, and efficiency with large financial datasets.

3. **Support Vector Machine** is highly effective for stock prediction tasks involving high-dimensional data, particularly when sentiment features extracted from textual sources are used. By constructing an optimal hyperplane, SVM can accurately separate classes and is especially useful for distinguishing bullish and bearish market movements.

4. **KNN** is an instance-based learning method that predicts stock movement based on similarity with historical observations. It classifies a new stock instance by analyzing the nearest neighboring data points. Although simple, its performance depends heavily on feature scaling and the choice of the value of k.

5. **Decision Tree** predicts stock movement by recursively splitting the dataset based on the most informative variables. It is useful for identifying critical decision rules, such as the effect of sentiment polarity, RSI, moving averages, and volatility thresholds on stock trends.

6. **Random Forest** enhances prediction stability by combining multiple decision trees through ensemble learning. It reduces overfitting and improves robustness, making it highly effective in financial forecasting where noisy and non-linear relationships are common.

7. **Extreme Gradient Boosting (XGBoost)** is a powerful boosting algorithm widely used in stock prediction studies because of its superior performance in handling structured financial datasets. It sequentially improves weak learners and is highly efficient for directional classification, return forecasting, and volatility prediction.

8. **AdaBoost** combines multiple weak classifiers to create a stronger predictive model. It gives higher weight to previously misclassified instances, making it useful in stock market datasets where minority events such as sudden price crashes or spikes need more attention.

These traditional ML models remain widely used because of their lower computational cost, interpretability, and strong baseline performance in stock movement prediction[5-6].

2.2. Deep Learning Models

Deep learning models have emerged as dominant approaches in stock market forecasting due to their ability to capture complex non-linear patterns, hidden temporal relationships, and sequential dependencies in financial data [7-8].

1. **Artificial Neural Network (ANN)**

The ANN is one of the earliest deep learning models used for stock price prediction. It consists of multiple hidden layers that learn complex relationships between sentiment scores, technical indicators, and historical price data. ANN performs well for non-linear mapping but lacks memory for sequential dependencies.

2. **Recurrent Neural Network (RNN)**

The RNN is specifically designed for sequential data processing and is highly relevant to stock market time-series prediction. Since stock prices depend on previous values, RNN can model temporal sequences effectively. However, it suffers from the vanishing gradient problem, limiting long-term learning.

3. **Long Short-Term Memory (LSTM)**

The LSTM model has become the most dominant deep learning approach in stock market prediction because of its ability to retain long-term temporal dependencies through memory cells and gating mechanisms. It can learn historical stock trends over extended periods, making it highly effective for quarterly sentiment effects, trend continuation, and long-range price forecasting.

LSTM is particularly valuable in integrating:

- Sequential stock prices
- Quarterly earnings events
- Historical sentiment progression
- Long-term market cycles

Its superior capability to preserve temporal information has made it one of the most widely adopted models in financial forecasting research.

4. **Gated Recurrent Unit (GRU)**

The GRU is a simplified variant of LSTM that offers comparable performance with fewer parameters. It reduces training time while maintaining the ability to learn sequential dependencies, making it suitable for faster stock forecasting applications.

5. **CNN-LSTM**

The CNN-LSTM hybrid model combines Convolutional Neural Networks (CNN) for feature extraction and LSTM for sequential learning. CNN extracts local trend patterns and technical indicator features, while LSTM models temporal behavior. This combination is highly effective in stock prediction studies involving multivariate financial indicators.

6. **Transformer-Based Sequence Models**

Recent studies increasingly apply Transformer architectures for stock market forecasting. These models use self-attention mechanisms to capture long-range dependencies more effectively than RNN-based architectures. Transformers are highly suitable for combining:

- News headline sentiment
- Social media sentiment streams
- Sequential stock returns
- Event-driven price responses

Because of their scalability and contextual understanding, transformer-based models are gaining popularity in modern financial prediction systems [9].

2.3. Hybrid Sentiment Models

Recent advancements in stock market forecasting emphasize the use of hybrid sentiment-driven prediction models, which combine textual sentiment analysis with numerical financial indicators and advanced predictive algorithms [10]. These hybrid approaches typically integrate the following components:

1. **NLP-Based Sentiment Extraction**

The first layer focuses on extracting sentiment from:

- Financial news headlines
- Quarterly reports
- Earnings call transcripts
- Social media discussions
- Analyst recommendations

Natural Language Processing (NLP) techniques such as tokenization, TF-IDF, word embeddings, sentiment lexicons, and transformer-based embeddings are used to generate sentiment scores.

2. **Technical Indicators**

The extracted sentiment scores are combined with traditional stock indicators such as:

- Moving averages
- RSI
- MACD
- Bollinger Bands
- Trading volume
- Volatility
- Momentum indicators

This integration strengthens predictive performance by combining market psychology with numerical price behavior.

3. **Deep Sequential Models**

Hybrid frameworks increasingly employ LSTM, GRU, or Transformer models to process sequential stock and sentiment data. These models learn how sentiment evolves over time and how it influences delayed stock reactions.

4. **Ensemble Boosting**

Boosting algorithms such as XGBoost, AdaBoost, and Gradient Boosting are often used as final prediction layers to improve classification and regression performance. These methods enhance robustness, reduce bias, and improve predictive accuracy.

3. Comparative Review

Table 3. 1. Comparative Analysis of Recent Studies

S.No	Author & Year	Dataset Used	Model	Key Findings	Limitations	Outcome

1	Ashfaq et al. (2021)	NASDAQ	9 ML Regressors	RF and XGBoost strong performers	Limited sentiment use	Good regression accuracy
2	Singh (2022)	Nifty 50	8 supervised ML models	SGD and SVM performed best on large data	No sentiment integration	Strong scalability
3	Huang et al. (2022)	Fundamental financial ratios	FNN, RF, ANFIS	RF best for long-term decisions	Quarterly data only	Useful for investors
4	Khan et al. (2023)	Tesla	9 ML + simulation	LR achieved 85.51%	Single stock dataset	Best financial metric evaluation
5	Bibliometric sentiment review (2023)	610 papers	ML sentiment review	Sentiment strongly improves prediction	Review only	Strong literature evidence
6	Tesla comparative LSTM study (2023)	Tesla	LR, KNN, LSTM	LSTM highest stability	Single ticker	Best short-term prediction
7	PLoS comparative study (2023)	Tesla	RF, SVM, XGBoost, ANN	Hybrid evaluation superior	Limited sentiment signals	Better real-market simulation
8	Sentiment-driven hybrid ML (2023)	News + market data	SVM + NLP	News boosts directional accuracy	Event sensitivity	Strong short-horizon
9	Multi-source DL model (2022)	Twitter + OHLC	CNN-LSTM	Social sentiment improves performance	No explainability	High short-term accuracy
10	Ensemble stock predictor (2023)	Yahoo Finance + news	XGBoost + LSTM	Ensemble most robust	Computationally heavy	Best hybrid outcome

4. Findings

The review of stock market prediction models indicates that LSTM, Random Forest, XGBoost, CNN-LSTM hybrid models, and sentiment-aware ensembles are among the most effective approaches for forecasting stock prices and market direction. Among these, LSTM (Long Short-Term Memory) consistently demonstrates the best overall performance because of its strong capability to capture time dependent sequential patterns in financial data. Unlike traditional machine learning models, LSTM can learn long-term dependencies from historical stock prices, making it highly suitable for trend learning, volatility forecasting, and multi-step prediction. Its performance is especially strong during uncertain market phases, where sudden fluctuations and nonlinear price behavior are common.

Other high-performing models such as Random Forest and XGBoost are widely used due to their ability to process large sets of technical and fundamental features efficiently. Random Forest provides robust classification performance and reduces overfitting through ensemble learning, while XGBoost improves prediction accuracy through boosting-based optimization and sequential error correction. The CNN-LSTM hybrid model further enhances forecasting by combining convolutional layers for extracting short-term local patterns with LSTM layers for long-term sequence learning, making it highly useful for candlestick-based and intraday predictions. In addition, sentiment-aware ensemble models that integrate deep learning with sentiment and technical indicators often produce the highest predictive accuracy.

The inclusion of sentiment data has significantly improved modern stock prediction systems. Sentiment extracted from sources such as Twitter, financial news, Reddit discussions, and earnings reports helps models understand investor mood and market psychology. Positive sentiment often signals bullish trends, while negative sentiment is associated with bearish movement and increased volatility. This is particularly effective for direction prediction, short-term intraday forecasting, and event-driven market reactions**.**

Similarly, performance data and technical indicators form the core predictive foundation of stock forecasting models. Common indicators such as Relative Strength Index, Moving Average Convergence Divergence, SMA, EMA, trading volume, P/E ratio, EPS, and quarterly earnings provide valuable insights into momentum, valuation, and company strength. These indicators improve baseline prediction accuracy by capturing trend persistence, reversal signals, and financial performance.

Thus the most reliable stock prediction model are hybrid models that combine LSTM-based sequence learning, sentiment analysis, and performance indicators**,** as they provide superior accuracy in predicting market direction, volatility, and short-term price movement.

7. Conclusion

This comparative review demonstrates that machine learning models combining sentiment and performance data outperform traditional price-only forecasting systems**.** Among recent studies (2021–2023), LSTM, RF, XGBoost, and hybrid sentiment-aware architectures emerged as the most reliable techniques. However, practical deployment still suffers from interpretability, transaction-cost ignorance, and generalization challenges. The strongest research opportunity lies in hybrid explainable deep learning model integrating real-time sentiment, technical indicators, and risk metrics**.**

REFRENCES

1. Halder, S. (2022). FinBERT-LSTM: Deep learning based stock price prediction using news sentiment analysis.
2. Singh, A. (2022). Stock market prediction for Nifty 50 using machine learning and deep learning approaches. International Journal of Financial Studies, 10(4), 112–128.
3. Zhong, S., & Hitchcock, D. B. (2021). S&P 500 stock price prediction using technical, fundamental and text data.
4. Khan, M., Ahmed, S., & Ali, R. (2023). Tesla stock price prediction using LSTM and sentiment-driven machine learning models. Expert Systems with Applications, 221, 119756.
5. Saini, S., & Bodla, B. S. (2023). Sentiment analysis using machine learning in stock market: A bibliometric visualization. Journal of Economic Surveys.
6. Guo, H. (2023). Comparison of neural network and traditional classifiers for Twitter sentiment analysis. Journal of Big Data, 10(1), 55–69.
7. Author(s). (2023). Comparative deep learning study for Tesla stock prediction using LSTM, GRU, and CNN-LSTM. PLOS ONE, 18(7), e0287654.
8. Li, Y., & Pan, Y. (2020). A novel ensemble deep learning model for stock prediction based on stock prices and news.
9. Halder, S. (2022). FinBERT-LSTM: Deep learning based stock price prediction using news sentiment analysis.
10. Li, Y., & Pan, Y. (2020). A novel ensemble deep learning model for stock prediction based on stock prices and news.