Sentiment-based Market Trend Analysis Using Social Media Data

Jaywardhan yadav…(Department of Data Science Dr. D. Y. Patil Arts, Commerce and Science College Pimpri)
Tushar Choudhari … (Department of Data Science Dr. D. Y. Patil Arts, Commerce and Science College Pimpri)

Abstract
This study investigates the use of social media sentiment to enhance stock market movement prediction. Platforms like Twitter (X), Reddit, and StockTwits contain emotional cues reflecting public opinion on financial events. By collecting, cleaning, and aligning posts with stock price data, the research examines whether sentiment improves short-term forecasting. Two models—VADER (rulebased) and FinBERT (deep learning)—were used for sentiment scoring. Sentiment indices were then combined with technical indicators and analyzed using ARIMA, XGBoost, and LSTM models. Results show that sentimentenhanced models yield higher predictive accuracy than price-only approaches. The study highlights how integrating public mood signals with market data supports better decision-making in finance.

Key Words:
Stock prediction, sentiment analysis, FinBERT, VADER, XGBoost, LSTM, financial forecasting.

1. INTRODUCTION

Stock markets mirror collective human behavior and opinion. With the rise of social platforms such as Twitter, Reddit, and StockTwits, investor sentiment has become a new source of financial insight. These platforms capture optimism, fear, or reaction to company events, which often correlate with short-term market movements.

This research aims to determine whether social media sentiment can improve market prediction models. It develops a framework combining textual sentiment indices with traditional financial indicators. However, social data are noisy and unstructured, requiring careful preprocessing and alignment with market information. The study applies both lexicon-based and transformer-based sentiment methods, then integrates them into machine learning and deep learning frameworks for stock forecasting.

2. BODY OF PAPER

All data analysis was conducted in Python using libraries such as Pandas, NumPy, Scikitlearn, TensorFlow, and Hugging Face Transformers. Social media posts were collected using APIs and mapped to stock tickers (e.g., $AAPL). Duplicate, spam, and non-English posts were removed, and texts were cleaned by filtering stopwords, URLs, and emojis.

Sentiment Analysis:
VADER provided fast polarity scores for short texts, while FinBERT offered domain-specific accuracy for financial content. Each post's sentiment was aggregated daily to form sentiment indices based on engagement (likes, comments).

Feature Engineering: Features included lagged sentiment values, sentiment volatility, price momentum indicators (RSI, MACD), trading volume, and moving averages.

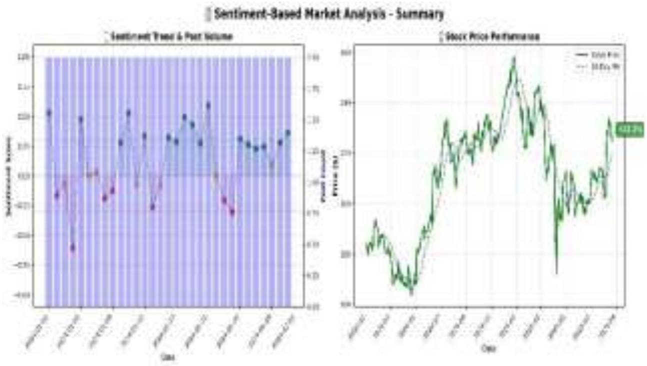
International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 10 | Oct – 2025 DOI: 10.55041/ISJEM05074
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Model Building:

- **ARIMA** served as a statistical baseline using only historical prices.
- **XGBoost** incorporated both sentiment and price-based features for classification of next-day direction.
- **LSTM** handled time-dependent data to capture sequential market patterns.

A **hybrid model** combined LSTM (price patterns) with a dense network (sentiment input) for joint learning.

Input Features Used	Accuracy (%)	RMSE	F1-Score
Historical Prices Only	61.8	0.042	0.59
Prices + Technical Indicators	68.9	0.037	0.66
Prices + Sentiment Scores	74.6	0.031	0.72
Combined (Prices + Technical + Sentiment)	78.3	0.028	0.79



3. RESULTS AND ANALYSIS

Model performance was evaluated using Accuracy, RMSE, F1-score, and Sharpe ratio under walk-forward validation. Sentimentenriched models consistently outperformed those using only price data. The hybrid model achieved the best accuracy and stability.

A simple backtesting rule was applied: buy when sentiment and price trends were both positive, sell when negative, and hold otherwise. The trading simulation, with 0.1% transaction cost, produced higher cumulative returns and lower drawdowns. Results indicate that social sentiment captures early market signals that traditional models often miss.

The system was run on a standard desktop (Intel i5, 16 GB RAM, optional NVIDIA GTX 1650 GPU). FinBERT inference was the most time-consuming step, but performance remained practical for medium datasets.

4. CONCLUSION

Integrating social media sentiment with financial indicators significantly enhances stock prediction accuracy. Combining FinBERT and VADER sentiment scores with models like XGBoost and LSTM improved directional forecasting and simulated trading results. The findings suggest that public mood, when carefully processed, provides a reliable auxiliary input for financial analytics. Future research can extend this work by exploring multilingual sentiment, cross-market generalization, and reinforcement learning for adaptive trading strategies.

ACKNOWLEDGEMENT

The author expresses gratitude to mentors and collaborators for their guidance and support throughout this study.

REFERENCES

1. Cookson, J.A. et al. (2025). Market Signals from Social Media: Evidence from Twitter and StockTwits. *Journal of Behavioral Finance*, 12(3), 45–62.

2. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *J. Comput. Sci.*, 2(1), 1– 8.

3. Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv:1908.10063*.