The Limits of Learning: A Critical Study on the Infeasibility of Stock Prediction Using Machine Learning

Guitar Poem
University of Github

Abstract

This paper critically examines the fundamental limitations of machine learning (ML) approaches for stock market prediction. Despite significant technological advancements and research interest, we argue that reliable stock prediction using ML remains inherently infeasible. We first analyze methodological flaws in existing research, particularly focusing on the StockNet paper, which we demonstrate suffers from selective reporting and reproducibility issues. Through a series of experiments using LSTM networks with historical price data and social media sentiment, we show that ML models consistently fail to outperform naive baselines. Our regression models achieve seemingly impressive R^2 values (0.62-0.95) but underperform compared to simple persistence models, revealing their inability to learn meaningful patterns beyond autocorrelation. Furthermore, LLM-generated sentiment labels from social media show no statistically significant correlation with next-day price movements. These findings align with the Efficient Market Hypothesis, suggesting that publicly available information is rapidly incorporated into stock prices, eliminating exploitable patterns. We conclude that while ML excels at pattern recognition in structured data, it fundamentally struggles with the inherently stochastic, noisy, and adversarial nature of financial markets, making reliable short-term stock prediction infeasible using conventional ML approaches.

1 Introduction

The application of machine learning (ML) to stock market prediction has garnered significant attention, driven by its potential for financial gain and market insights. However, despite technological advancements, this paper argues that reliable stock prediction using ML remains fundamentally infeasible.

The Nature of Financial Markets

Financial markets are inherently stochastic, noisy, and adversarial systems that pose significant challenges for prediction. The Efficient Market Hypothesis (EMH) provides a foundational theoretical framework for understanding these limitations, arguing that stock prices quickly incorporate all available information, leaving little room for systematic profit through prediction. According to Lo [1], this efficient price adjustment mechanism severely constrains the potential of machine learning approaches, which aim to detect hidden patterns but consistently face significant barriers to reliable forecasting. Recent empirical work by Caporale and Plastun [2] further confirms that even advanced ML methods struggle to consistently extract meaningful signals, particularly when accounting for practical considerations like transaction costs and the adaptive nature of market responses. Unlike traditional technical and fundamental analyses, ML-based forecasting must contend with these fundamental limitations while attempting to navigate the complex, dynamic nature of financial markets.

Infeasibility Claims in Empirical Studies

Empirical comparisons frequently show that ML models fail to outperform naive baselines. Sundqvist [3] evaluated multiple ML techniques against a random walk model and found no consistent superiority in predictive accuracy. Notably, many deep learning models excel only in-sample but degrade sharply in out-of-sample tests, raising concerns of data leakage and overfitting. Nguyen et al. [4] applied LSTM models to the Vietnamese stock exchange and found strong training accuracy but negligible predictive power under realistic, out-of-sample evaluations.

Issues with Current Research

Elangovan and Prasad [5] conducted a comprehensive review of machine learning models used in stock prediction, concluding that many studies suffer from overfitting, lack of rigorous validation, and poor generalization. Similarly, Ghosh et al. [6] identified methodological flaws in a wide range of ML applications to market forecasting and emphasized the need for reproducibility and robustness in empirical design. Furthermore, Bailey et al. [7] demonstrated the high probability of backtest overfitting in financial models, showing that many apparently successful strategies fail to generalize to new data. Similarly, Taleb [8] emphasized the inherent unpredictability of financial markets due to black swan events and extreme value theory. These works collectively underscore the challenges of applying machine learning to stock prediction, particularly in the face of market complexity and non-stationarity.

Only Effective in High-Frequency Trading
Despite infeasibility for daily stock prediction, high-frequency approaches remain effective. The Jane Street Kaggle Competition (2021–2023) [9] demonstrated successful ML models optimized for extremely short-term, high-frequency trading scenarios. Nonetheless, these methods are difficult to extend to broader market applications due to their limited timeframe and highly constrained, noisy environments.

These findings collectively support our central thesis: machine learning's effectiveness in stock prediction remains fundamentally constrained by market complexity, data quality issues, and poor model generalizability.

2 Questionable Results in StockNet

In this section, we critically analyze the StockNet paper [10], published in the Association for Computational Linguistics (ACL) conference, which proposes leveraging social media data to enhance stock movement prediction. While the authors report promising results, our analysis reveals significant **issues with selective reporting and reproducibility**. We conduct a series of experiments and analyses to critically evaluate the claims made in the paper and to demonstrate the limitations of their approach.

The author claims a balanced dataset with 49.78% falling and 50.22% rising classes. **However, the authors do not report that the actual test set is skewed:** 45.39% falls and 54.61% rises. This imbalance means that a naive majority-class predictor would already achieve 54.61% accuracy—very close to the reported 58.23% accuracy of StockNet. The marginal gain raises concerns about the practical significance of the results. Moreover, No statistical tests are reported to confirm that the performance gain is significant.

Furthermore, the authors treat 50% as the random baseline, which is misleading given the class skew. Most baseline models perform worse than the naive majority-class predictor (54.61%): RAND achieves 50.89%, ARIMA 51.39%, Random Forest 53.08%, and TSLDA 54.07%. Only a sentiment classification model HAN [11] achieves 57.64%, slightly above the naive baseline.

We also attempted to reproduce StockNet's results using the same dataset but with improved sentiment labels generated by LLM. Our models failed to outperform the naive baseline, and in most cases, adding sentiment degraded performance. Also, our analysis revealed no significant correlation between sentiment labels and next-day stock movements. This suggests that the claimed benefit of social media sentiment is questionable.

3 Experimental Method

We evaluate stock prediction feasibility using LSTM networks with two data sources: historical prices and social media sentiment. For each source, we implement binary classification (up/down) and price regression tasks. This design tests whether models learn meaningful patterns beyond baselines while assessing StockNet's claims.

3.1 LSTM with Historical Data

We evaluate LSTM networks trained solely on historical price data from a subset of the StockNet dataset [10], containing 88 U.S. stocks over two years. The data is preprocessed into 5-day sequences of normalized price vectors (adjusted closing, high, and low prices). The LSTM is tested on two tasks: binary classification of next-day stock movement (up/down) using accuracy, and regression of next-day closing price using MAE and R2 metrics, where R2 measures the proportion of variance explained.

This control experiment assesses technical signals' predictive power and historical data's generalization capability in stochastic markets.

3.2 LSTM with Social Media Data

This experiment investigates whether social sentiment from Twitter contains meaningful signals for forecasting stock movements, we train a LSTM network using sentiment labels derived from tweets in the StockNet dataset [10]. The dataset contains stock-specific tweets collected over a two-year period.

For the sentiment labels, we employ a specialized LLM agent to process stock-related tweets and extract sentiment. The agent follows a structured analysis process: identifying the target stock, extracting positive and negative factors from each tweet, assessing individual tweet sentiment, and aggregating these into an overall sentiment score ([Positive], [Neutral], or [Negative]). For implementation, we use the DeepSeek-R1-Distill-Qwen-1.5B model, which balances computational efficiency with strong reasoning capabilities for financial contexts.

We guide the LLM with a structured prompt:

```
<Task>
Analyze tweets to determine their likely
impact on future stock price. Answer with:
"[Positive]", "[Neutral]", or "[Negative]".

<Solving Process>
1. Identify Target Stock
2. Tweet-by-Tweet Analysis:
    a) Identify Positive/Negative factors
    b) Assess sentiment impact
3. Conclude overall sentiment
```

```
<Tweets>
{tweet_text}

<Output>
Conclude with:
[Positive]/[Neutral]/[Negative]
```

4 Experimental Results

4.1 Infeasibility of Predicting Stock Movements Direction

As shown in Table 1, the classification task of predicting stock movements direction achieves accuracy equal to or lower than simply predicting the dominant class in the test set. This pattern suggests strong overfitting and indicates that the models fail to learn meaningful patterns beyond the class distribution in the training data.

4.2 Infeasibility of Predicting Stock Prices

Table 2 shows R^2 values for price-based models ranging from 0.6230 to 0.9514 across six stocks. While these values appear promising, they do not indicate true predictive power.

Table 3 reveals a critical insight: despite the LSTM models' seemingly good R^2 values (ranging from 0.6230 to 0.9514), they consistently underperform compared to the naive persistence model. This pattern persists across all six stocks, with the persistence model achieving higher R^2 values in every case. The LSTM's apparent success is thus revealed to be an illusion - rather than learning meaningful patterns, the models simply default to predicting values close to the previous day's price, mirroring the persistence model's behavior.

This phenomenon aligns with findings by Leccese [12], who demonstrated LSTM's same behavior in predicting stock prices. Leccese's analysis of U.S. market data from 1950-2018 showed an R^2 of 0.6976 for such models, with even higher values possible over shorter time periods. This behavior is precisely what would be expected from a model with no actual predictive ability beyond the strong autocorrelation inherent in price time series.

The models' failure to learn meaningful patterns is further evidenced by the overfitting observed in the classification task (Table 1). Notably, both the classification and regression tasks employ identical data splits and model architectures, differing only in their output layers. Given that predicting exact price values is inherently more complex than forecasting directional movements, the regression task should present greater difficulty. Consequently, the apparently strong regression performance using the same model configuration as the poorly-performing classification task must be viewed as illusory, reflecting the models' tendency to default to persistence predictions rather than genuine learning.

4.3 Sentiment Label Effectiveness

Moreover, the sentiment generated from social media data doesn't provide meaningful signal to improve the results, as shown in Table 2. For five of the six stocks (AAPL, AMZN, BAC, D, and GOOG), adding sentiment data resulted in higher MAE and lower R^2 values. Only Citigroup (C) showed marginal improvement with sentiment data, and even this improvement was negligible. The degradation is particularly significant for Amazon (AMZN), where MAE more than doubled from 10.21 to 23.24 when sentiment was incorporated, while R^2 dropped substantially from 0.9514 to 0.7908. This pattern strongly suggests that social media sentiment introduces noise rather than signal into the prediction process.

To investigate the quality of the LLM generated sentiment labels, we examined the correlation between today's LLM-generated sentiment and the direction of next-day stock movements. Table 4 shows the correlation strength measured using Cramer's V coefficient. The analysis reveals consistently weak and statistically insignificant correlations between sentiment and price movements (all $V < 0.1$, $p > 0.05$). Prediction accuracies (29.72%-34.99%) were only slightly better than random chance, indicating social media sentiment offers minimal predictive value for stock movements.

We further examine Amazon as a case study. Table 5 shows the distribution of sentiment labels and price movements. Chi-square analysis revealed no statistically significant relationship between sentiment and subsequent price movements (p-value = 0.8572).The nearly uniform distribution across all cells in Table 6 visually confirms this lack of predictive relationship - each sentiment category shows almost identical distributions across all three price movement outcomes.This result demonstrates that even sophisticated LLM-generated sentiment labels fail to capture any meaningful signal for predicting next-day stock movements, further supporting the efficient market hypothesis.

Table 1: Classification Performance of LSTM Models With and Without Social Media Sentiment Data

Stock	Up (%)	Down (%)	Accuracy (No Sent)	Accuracy (With Sent)
AAPL	37.50%	**62.50%**	0.5618	0.5618
AMZN	**55.32%**	44.68%	0.5057	0.5057
BAC	42.22%	**57.78%**	0.4969	**0.5247**
C	40.91%	**59.09%**	0.5212	**0.5636**
D	46.67%	**53.33%**	0.5410	**0.5532**
GOOG	50.00%	50.00%	**0.5328**	0.5214

The "Up (%)" and "Down (%)" columns show the distribution of positive and negative movements in the test set. Accuracy values in bold indicate improved performance with sentiment data. Notably, most models achieve equal or lower accuracy than always predicting the dominant class (highlighted in **bold**), suggesting strong overfitting.

Table 2: Regression Performance of LSTM Models With and Without Social Media Sentiment Data

Stock	MAE (No Sent)	MAE (With Sent)	R^2 (No Sent)	R^2 (With Sent)
AAPL	**2.079**	2.264	**0.6230**	0.5726
AMZN	**10.21**	23.24	**0.9514**	0.7908
BAC	**0.215**	0.218	**0.8744**	0.8707
C	0.787	**0.759**	0.6815	**0.7044**
D	**0.660**	0.732	**0.8718**	0.8559
GOOG	**9.56**	11.13	**0.9477**	0.9327

The social media sentiment data generated by the LLM generally degrades model performance across both metrics. For five of six stocks, incorporating sentiment leads to higher Mean Absolute Error (MAE) and lower R^2 values compared to models using only historical price data. Only Citigroup (C) shows marginal improvement, suggesting sentiment features introduce more noise than signal in most cases.

Table 4: Correlation Between LLM-Generated Sentiment and Next-Day Price Movement

Stock	Correlation	p-value	Accuracy
AAPL	0.0371	0.8456	32.36%
AMZN	0.0768	0.2342	29.72%
GOOG	0.0300	0.9312	34.04%
D	0.0683	0.3127	34.99%
C	0.0572	0.4891	33.56%
BAC	0.0550	0.5243	32.04%

We categorized next-day price movements as positive ($\geq 0.5\%$), neutral (-0.5% to 0.5%), and negative ($\leq -0.5\%$). All correlations are weak ($V < 0.1$) and statistically insignificant ($p > 0.05$), with prediction accuracies only slightly better than random chance (29.72%-34.99%).

Table 5: AMZN: Distribution of Sentiment Label and Price Movements in Test Set

Sentiment	%	Price Movement	%
Positive	40.00	Up ($\geq$0.5%)	35.63
Negative	33.96	Neutral (-0.5% to 0.5%)	32.50
Neutral	26.04	Down ($\leq -0.5\%$)	31.88

Table 6: AMZN: Conditional Distribution of Next-Day Price Movements Given Sentiment Label

Sentiment	Down (%)	Neutral (%)	Up (%)
Negative	31.29	34.97	33.74
Neutral	32.80	28.80	38.40
Positive	31.77	32.81	35.42

Table 3: Comparison of R^2 Values Between LSTM Predictions and Naive Persistence Model

Stock	Avg Daily Variation	R^2 (LSTM)	R^2 (Persistence)
AAPL	-0.10%	0.6230	**0.8695**
AMZN	+0.39%	0.9514	**0.9329**
BAC	+0.16%	0.8744	**0.8686**
C	+0.03%	0.6815	**0.8745**
D	-0.11%	0.8718	**0.8964**
GOOG	+0.29%	0.9477	**0.9842**

The daily variation shows that stock prices typically change by less than 1% on average. The R^2 values from the LSTM models are consistently lower than those from a naive persistence model that simply predicts today's price for tomorrow. This comparison demonstrates that the LSTM models fail to outperform even this simple baseline, suggesting they lack genuine predictive power. The persistence model's higher R^2 values indicate that the LSTM's apparent performance is not due to learned patterns but rather reflects the strong autocorrelation inherent in stock price time series.

5 Discussion: What ML Can Do — and What It Can't

Our research demonstrates both the capabilities and limitations of machine learning in financial markets. While ML models can effectively classify sentiment in financial news (achieving high accuracy in our sentiment analysis task), they fail to translate this capability into meaningful stock price predictions. This highlights an important distinction: ML excels at pattern recognition in structured data but struggles with the inherently unpredictable nature of financial markets. The efficient market hypothesis provides a theoretical framework for understanding this limitation—new information is rapidly incorporated into prices, eliminating exploitable patterns. Any potential ML-based predictive advantage might only exist in extremely short timeframes (milliseconds to seconds) before market efficiency eliminates the opportunity. Beyond these scenes, ML applications in finance may be more valuable for tasks like risk assessment, portfolio optimization, and anomaly detection rather than direct price prediction.

6 Conclusion

This study demonstrates that many current ML-based stock prediction approaches are unreliable and misleading. Our experiments show that neither historical price data nor social media sentiment analysis provides meaningful predictive value for next-day stock movements. LSTM models achieve seemingly impressive R^2 values but consistently underperform simple persistence models, revealing their inability to learn patterns beyond autocorrelation. LLM-generated sentiment labels show no statistically significant correlation with price movements, confirming that publicly available information is quickly incorporated into stock prices, as suggested by the Efficient Market Hypothesis. While ML excels at pattern recognition in structured domains, it fundamentally struggles with the inherently stochastic, noisy, and adversarial nature of financial markets. Future research might explore alternative timeframes or data sources, but our results caution against overreliance on ML for short-term stock prediction using readily available information.

References

[1] A. W. Lo, "The adaptive markets hypothesis: Market efficiency from an evolutionary perspective," *Journal of Portfolio Management*, vol. 30, no. 5, pp. 15–29, 2004.

[2] G. M. Caporale and A. Plastun, "Forecasting financial market efficiency using machine learning methods," *Journal of Empirical Finance*, vol. 74, pp. 123–139, 2024.

[3] T. Sundqvist, *Machine Learning for Stock Price Prediction: A Comparison with Naive Models*. PhD thesis, Uppsala University, 2021.

[4] V. M. Nguyen, T. H. Tran, and N. H. Nguyen, "Predicting emerging market stocks using lstm: Evidence from vietnam," *Humanities and Social Sciences Communications*, vol. 11, no. 1, p. 110, 2024.

[5] N. Elangovan and K. Prasad, "A systematic review of machine learning techniques for stock market prediction," *Journal of Information Technology in Management*, vol. 31, no. 2, pp. 12–24, 2020.

[6] R. Ghosh, A. Dutta, and S. Majumder, "Recent trends in stock market prediction using machine learning:

A review," *PeerJ Computer Science*, vol. 9, p. e1435, 2023.

[7] D. H. Bailey, J. Borwein, M. L. d. Prado, and Q. J. Zhu, "The probability of backtest overfitting," *Journal of Computational Finance*, vol. 20, no. 4, pp. 39–69, 2014.

[8] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*. Random House, 2007.

[9] J. Street, "Jane street market prediction," 2021.

[10] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1970–1979, 2018.

[11] N. Hu, Z. Tian, X. Lu, G. Zhang, Y. Su, and Y. Shi, "Attention-based hierarchical neural network for interpretable visio-linguistic sentiment analysis," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4831–4840, Association for Computational Linguistics, 2018.

[12] A. Leccese, "Machine learning in finance: Why you should not use lstms to predict the stock market," *Bluesky Capital Management*, 2019.