Evaluating Stock Return Predictability Using Historical Prices and Twitter Sentiment

Raayan Hemrajani

Received July 14, 2025
Accepted February 15, 2026
Electronic access March 15, 2026

Predicting short-term stock returns remains a longstanding challenge in financial research due to high noise levels and the rapid incorporation of publicly available information into asset prices. While prior studies suggest that sentiment extracted from social media may improve market forecasting, many report implausibly strong results under evaluation frameworks that risk data leakage. This study examines whether sentiment derived from Twitter provides incremental predictive value for next-day stock returns when combined with historical price information under a leakage-free, time-series-appropriate evaluation protocol. Daily adjusted closing prices were collected for 25 publicly traded U.S. companies over a one-year period from September 2021 to September 2022, covering 252 trading days. Over 80,000 tweets referencing these firms were analyzed. Tweet sentiment was quantified using the VADER sentiment analyzer and aggregated at the daily, per-stock level using mean sentiment, tweet volume, and a missingness indicator. Linear regression, ridge regression, and gradient boosting regression models were evaluated using a strict chronological train–test split. Model performance was assessed using mean absolute error, root mean squared error, and R^2, and compared against a no-skill baseline predicting zero return. Results indicate that none of the evaluated models materially outperform the baseline, and that aggregated Twitter sentiment provides minimal incremental predictive value for next-day stock returns after controlling for lagged price information. These findings suggest that daily aggregated Twitter sentiment does not materially improve next-day stock return prediction under leakage-free evaluation, highlighting the challenges of short-horizon financial forecasting.

Keywords: stock return prediction, investor sentiment, Twitter data, time-series forecasting, machine learning

Introduction

Investor sentiment has long been studied as a potential influence on financial market behavior[1], particularly during periods of heightened uncertainty or market stress[2], and broader research has examined the extent to which observable news and other public information explain short-term stock price movements[3]. With the growth of digital media, large-scale textual data from sources such as news articles, firm disclosures, and social media platforms have become increasingly accessible, motivating interest in sentiment-based approaches to financial prediction[4–9]. However, predicting short-horizon stock returns remains difficult, as financial markets rapidly incorporate publicly available information into prices, a core implication of the efficient market hypothesis[10].

Despite extensive prior work, a gap remains in the literature regarding rigorous evaluation of sentiment-based prediction models under time-series–appropriate protocols. Many studies report unusually high predictive performance while relying on random data splits or same-day price targets[11,12], raising concerns about data leakage and overstated conclusions[13–15]. This study addresses this gap by evaluating whether Twitter sentiment provides incremental predictive value for next-day stock returns when assessed under a leakage-free, chronological framework.

Twitter is selected as the sentiment source due to its widespread use in prior financial sentiment studies[11,16] and the reproducibility of large-scale data collection. The scope of the study is limited to daily aggregation and next-day forecasting, with an emphasis on methodological rigor rather than model complexity[17,18].

Accordingly, this study tests whether daily, Twitter-derived sentiment adds incremental predictive value for next-day stock returns after controlling for lagged price information, using a leakage-free chronological evaluation. The analysis is limited to U.S. equities over a one-year period with daily aggregation and a lexicon-based sentiment model, which may understate intraday effects and finance-specific language[8,19–22].

Methods

This study employs an observational, retrospective time-series research design. The sample consists of 25 publicly traded U.S. companies selected based on data availability across both

price and Twitter datasets. Daily adjusted closing price data were collected over 252 trading days from September 30, 2021 to September 29, 2022. Twitter data were collected over the same calendar period using keyword-based queries corresponding to company names and ticker symbols.

To ensure chronological consistency and prevent look-ahead bias, tweet timestamps were mapped to the next available trading day. Sentiment was computed using the VADER sentiment analyzer[23], a rule-based model designed for short, informal text. For each stock and trading day, the mean sentiment score and tweet count were calculated, along with a binary indicator denoting days with no associated tweets.

The dependent variable is the next-day log return[24–26] of the adjusted closing price. Independent variables include lagged stock returns, rolling volatility measures, daily mean sentiment, tweet volume, and the sentiment missingness indicator. The analytical procedure consisted of data collection, preprocessing, feature construction, model training, and out-of-sample evaluation.

Multicollinearity diagnostics. To assess potential multicollinearity among predictors, Variance Inflation Factors (VIF) were computed for the full feature set used in model estimation. Table 2 reports VIF values for all predictors.

Table 1 Variance inflation factors (VIF) for model predictors.

Feature	VIF
ret_1_lag	1.001
vol_5	1.070
ma_5	1.146
sent_mean_lag1	1.208
tweet_count_lag1	1.225
Volume	1.307
sent_missing_flag_lag1	1.309

All VIF values are close to 1 and well below commonly cited thresholds (e.g., 5 or 10), indicating minimal multicollinearity and suggesting that coefficient instability due to correlated predictors is unlikely. Nevertheless, ridge regression is included as a robustness check, as regularization mitigates coefficient sensitivity in the presence of correlated inputs.

Three models were evaluated: linear regression, ridge regression, and gradient boosting regression. All models were trained and tested using a strict chronological split, with earlier observations used for training and later observations reserved for testing. Performance was evaluated using mean absolute error, root mean squared error, and R^2. A no-skill baseline predicting zero return was used for comparison. All data used are publicly available and contain no personal identifying information. No human subjects were involved, and the study complies with NHSJS ethical guidelines for publicly available data.

Results

Table 1 summarizes out-of-sample predictive performance across all evaluated models on the held-out test set.

Table 2 Out-of-sample predictive performance across models (MAE, RMSE, R^2).

Model	MAE	RMSE	R^2
Baseline (0 return)	0.01973	0.02806	—
Linear Regression	0.01973	0.02809	−0.0117
Ridge Regression (tuned)	0.01973	0.02809	−0.0117
Gradient Boosting Regressor	0.02040	0.02878	−0.0620
GBR (no sentiment)	0.02034	0.02878	−0.0622

Note: R^2 is undefined for the constant baseline predictor.

As shown in Table 1, predictive performance is similar across all evaluated models and closely matches that of the no-skill baseline[14,18]. R^2 values are negative for all models, indicating that none explain variance in next-day stock returns beyond the baseline. Models incorporating Twitter sentiment do not exhibit lower error metrics than price-only variants, suggesting minimal incremental contribution from aggregated sentiment features.

Discussion

The results indicate that aggregated Twitter sentiment provides little to no incremental predictive value for next-day stock returns when evaluated under a leakage-free, chronological framework. This finding is consistent with the efficient market hypothesis[10], which posits that publicly available information is rapidly incorporated into asset prices[24]. The research objective of assessing the incremental value of sentiment under conservative evaluation assumptions was met, with results suggesting that daily aggregated sentiment is insufficient for short-horizon forecasting.

Several limitations should be acknowledged. Sentiment was aggregated at the daily level, potentially obscuring intraday dynamics. The analysis focuses on a one-day forecast horizon and does not evaluate economic profitability or trading strategies. Additionally, the use of a lexicon-based sentiment model may limit sensitivity to nuanced financial language[19–22]. Future research may explore higher-frequency data, alternative sentiment representations[20,21], or longer-term prediction horizons to further investigate conditions under which sentiment may play a larger role. Taken together, these findings emphasize that methodological rigor is essential when evaluating sentiment-based financial models and that conservative benchmarks are critical for avoiding misleading conclusions[13,15].

References

1 M. Baker and J. Wurgler, *Journal of Economic Perspectives*, **21**, 129–152,.

2 P. Tetlock, *Journal of Finance*, **62**, 1139–1168,.

3 D. Cutler, J. Poterba and L. Summers, *Journal of Portfolio Management*, **15**, 4–12,.

4 J. Bollen, H. Mao and X. Zeng, *Journal of Computational Science*, **2**, 1–8,.

5 W. Antweiler and M. Frank, *Journal of Finance*, **59**, 1259–1294,.

6 P. Tetlock, M. Saar-Tsechansky and S. Macskassy, *Journal of Finance*, **63**, 1437–1467,.

7 A. Smailović, J. Grčar, N. Lavrač and M. Žnidaršič, *Information Sciences*, **285**, 181–203,.

8 H. Nassirtoussi, S. Aghabozorgi, T. Wah and A. Ngo, *Expert Systems with Applications*, **41**, 7653–7670,.

9 J. Engelberg and C. Parsons, *Journal of Finance*, **66**, 67–97,.

10 E. Fama, *Journal of Finance*, **25**, 383–417,.

11 S. Gu, B. Kelly and D. Xiu, *Review of Financial Studies*, **33**, 2223–2273,.

12 G. Ranco, M. Aleksovski, G. Caldarelli, M. Grčar and I. Mozetič, *PLOS ONE*, **10**, 0138441,.

13 M. Prado, *Journal of Portfolio Management*, **44**, 120–133,.

14 A. Goyal and I. Welch, *Review of Financial Studies*, **21**, 1455–1508,.

15 C. Harvey, Y. Liu and H. Zhu, *Review of Financial Studies*, **29**, 5–68,.

16 T. Sprenger, A. Tumasjan, P. Sandner and I. Welpe, *European Financial Management*, **20**, 926–957,.

17 M. Prado, *Advances in financial machine learning*, Wiley, Hoboken, NJ.

18 J. Campbell and S. Thompson, *Review of Financial Studies*, **21**, 1509–1531,.

19 T. Loughran and B. McDonald, *Journal of Finance*, **66**, 35–65,.

20 D. Araci, Proceedings of the ACL Workshop on Financial Technology and Natural Language Processing, pp. 1–7,.

21 C. Kearney and S. Liu, *International Review of Financial Analysis*, **33**, 171–185,.

22 T. Loughran and B. McDonald, *Journal of Accounting Research*, **54**, 1187–1230,.

23 C. Hutto and E. Gilbert, Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, pp. 216–225,.

24 B. Malkiel, *Journal of Economic Perspectives*, **17**, 59–82,.

25 J. Campbell and R. Shiller, *Review of Financial Studies*, **1**, 195–228,.

26 N. Jegadeesh and D. Wu, *Journal of Financial Economics*, **110**, 712–729,.