Is social sentiment statistically associated with stock price volatility?

Abstract

This project investigates the relationship between online investor sentiment and short-term stock price volatility. Using financial news headlines, Reddit discussions, and market price data, a unified dataset is constructed that captures both news-driven and user-generated sentiment for nine actively traded stocks from March 1^{s}, 2025 to April 15th, 2025. Sentiment scores are derived using VADER analysis, and volatility is computed using 5-day rolling standard deviation of log returns.

I conducted a bidirectional analysis to determine whether sentiment leads, lags, or moves concurrently with volatility. Our findings show that **lagged sentiment metrics are most correlated with future volatility**, suggesting predictive potential. I also uncovered distinct differences in how news vs. Reddit sentiment influences volatility across different stocks and variation in the strength and direction of sentiment-volatility alignment, indicating that certain stocks (like TSLA and PLTR) are more sentiment-sensitive. A Ridge regression model trained on lagged sentiment and volume features achieves an R^2 score of **0.91 on the test set**. This analysis highlights the nuanced yet measurable influence of social sentiment and attention on market behavior.

Motivation

In this project, I explored how investor sentiment affects short-term stock price volatility, particularly from Reddit discussions and financial news headlines. I hypothesize that *daily changes in social and news sentiment impacts next-day stock price volatility*.

Reddit communities like **r/stocks** and **r/wallstreetbets**, along with news aggregators such as **Google News**, provide a window into public opinion and speculation around trending stocks.

I focus our analysis on a selected set of **9 actively discussed stocks** including **AAPL, TSLA, NVDA, META, PLTR, AMZN, BAC, INTC, and UNH**. These were chosen for their strong retail investor following and frequent appearance across Reddit and financial news platforms during the study period.

The core questions I aim to answer are:

- Does sentiment lead, lag, or move together with volatility?
- Is there a difference in predictive strength between news-driven and user-generated sentiment?
- Can lagged sentiment features improve volatility forecasting in a regression model?

By combining structured market data with textual sentiment scores and attention metrics, I explore whether investor mood can explain or anticipate fluctuations in short-term volatility.

Data Collection and Cleaning

Data Sources:

- **Google News Headlines**
  https://news.google.com/

  Dataset - News headlines and timestamps
  Purpose - To generate news sentiment and volume

- **Python Reddit API Wrapper - PRAW**
  https://praw.readthedocs.io/en/stable/

  Dataset - Reddit posts from r/stocks and r/wallstreetbets
  Purpose - To derive user sentiment and activity

- **Yahoo! Finance API – yfinance (API dataset)**
  https://pypi.org/project/yfinance/

  Dataset - Daily OHLCV stock prices

Purpose - To compute daily volatility for each stock

Data Cleaning Steps:

Market Data

- Pulled daily prices for 9 stocks from **March 1 to April 15, 2025** using yfinance.
- Computed **log returns** and a **5-day rolling standard deviation** as our volatility metric.
- Missing dates or weekends were not interpolated, aligning with trading days.

News Data

- Scraped headlines from Yahoo Finance using requests and BeautifulSoup, extracting publication date, ticker, and title.
- Cleaned text using nltk stopwords and applied **VADER Sentiment Analysis** to assign a compound sentiment score per headline.
- Aggregated to a **daily average sentiment per ticker**, along with **daily headline count** as news volume.

Reddit Data

- Collected submissions using **PRAW API** from r/stocks and r/wallstreetbets for the same time period.
- Extracted post title, creation date, and ticker mentions.
- Applied **VADER** on post titles and computed **daily sentiment and volume per ticker**.

The cleaned versions of the three datasets are available as:

- combined_market_data_clean.csv
- combined_news_data_clean.csv
- combined_reddit_data_clean.csv

Merging Strategy:

- Merged all datasets on **['ticker', 'date']**, using an **inner join** to ensure aligned availability.
- Filled missing sentiment scores with per-ticker means, and sentiment changes/volume with zeros.
- Created **lag features** (e.g. previous day sentiment and volume) using groupby().shift(1).

The final combined dataset is available as:
merged_sentiment_volatility.csv

Technical Solution

Feature Engineering:

To capture both investor sentiment and its potential impact on stock volatility, I engineered the following **lagged features** per ticker:

Feature	Description
news_sentiment_lag1	Previous day's average sentiment from financial news
reddit_sentiment_lag1	Previous day's average sentiment from Reddit discussions
news_sentiment_change	Day-over-day change in news sentiment
reddit_sentiment_change	Day-over-day change in Reddit sentiment
news_volume_lag1_z	Z-score normalized count of news headlines (lagged)
reddit_volume_lag1_z	Z-score normalized Reddit post count (lagged)
volatility_lag1 ***(optional)***	Previous day's volatility used in an alternate model

All volume features were **z-score normalized per ticker** to ensure comparability across stocks with different levels of activity.

Exploratory and Bidirectional Analysis:

To investigate temporal relationships between sentiment and stock price volatility, I conducted a structured lead-lag correlation analysis. The procedure included the following steps:

- **Sentiment Alignment**: I computed daily average sentiment scores for both news and Reddit data, aggregated by (date, ticker). Lag features were generated using .shift(1) grouped by ticker, including news_sentiment_lag1, reddit_sentiment_lag1, and their respective day-over-day changes.
- **Correlation Matrices**: I created separate correlation matrices to quantify relationships between volatility and sentiment features under three temporal configurations:
    - **Same-day analysis**: correlation between daily volatility and sentiment metrics of the same day.
    - **Lagged sentiment analysis**: correlation between current volatility and previous day's sentiment.
    - **Lagged volatility analysis**: correlation between sentiment and previous day's volatility to test for reverse influence.
- **Visualizations**:
    - Heatmaps were used to display pairwise correlations.
    - Scatter plots were generated using seaborn.regplot() to visualize linear patterns for each lag configuration.
    - Feature-wise correlation bar plots were created to compare the magnitude and direction of influence across different sentiment-based predictors.
- **Per-Ticker Analysis**: The entire correlation analysis pipeline was repeated within each ticker group using groupby('ticker'), allowing us to compute and compare per-stock sentiment-volatility correlation structures.

These exploratory steps helped define the final feature set and informed the modeling strategy, particularly in determining the predictive utility of lagged sentiment and attention metrics.

Modeling Approach:

I used **Ridge Regression** to model and predict **daily volatility** as a function of sentiment-driven and attention-driven features.

Steps:

1. Filtered data for **March 2025**, aligning with 20 trading days.
2. Dropped rows with missing values due to lagging or normalization.
3. Standardized input features using StandardScaler.
4. Used RidgeCV for automatic alpha selection via cross-validation.
5. Built **two models**:
    - Without volatility_lag1 (purely sentiment-based)
    - With volatility_lag1 (controls for autoregressive effect)
6. Evaluated both models using **RMSE** and **R^2 score** on a **chronological train-test split** (shuffle=False).
7. Visualized model coefficients and performance through bar plots and volatility time series graphs.

Implementation Challenges and Solutions:

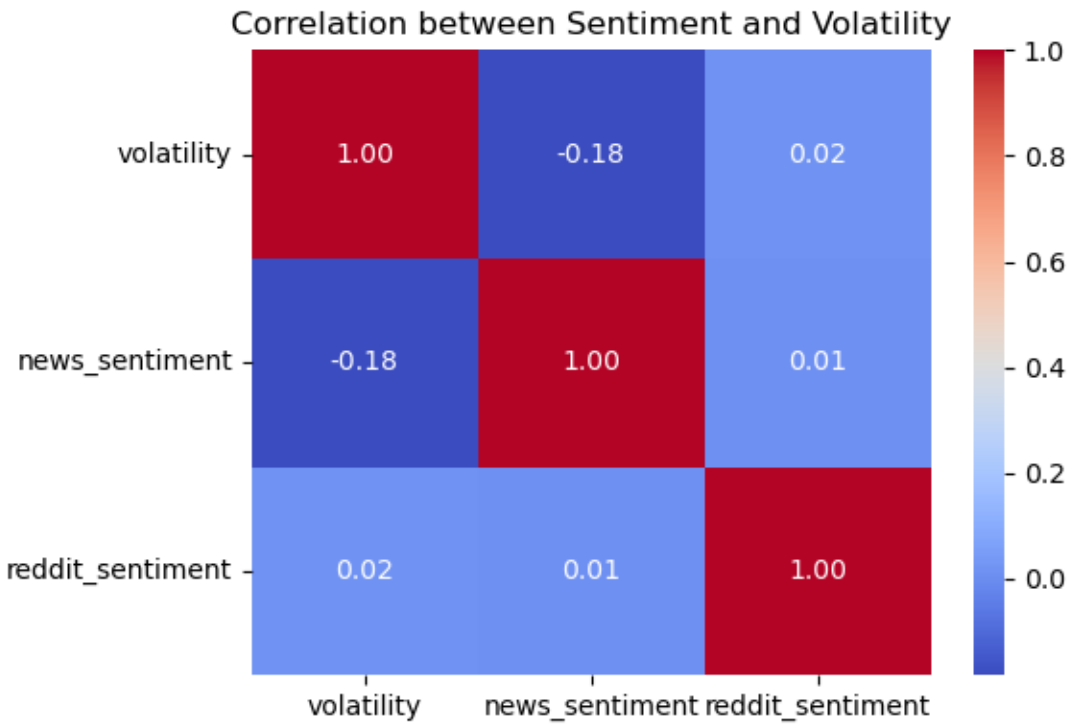
Challenge	Solution
Mixed timezones in raw datetime columns	Parsed with utc=True and removed timezone info before merging
Sparse sentiment data on some dates	Imputed missing sentiment with ticker-wise mean; volume with 0
Volume skew due to outliers	Applied **rolling z-score normalization** ticker-wise
Data leakage risk from rolling volatility	Ensured volatility was computed **only using past 5 days**
Maintaining time order in model	Used train_test_split(..., shuffle=False) to preserve date integrity

Analysis, Insights, and Conclusions

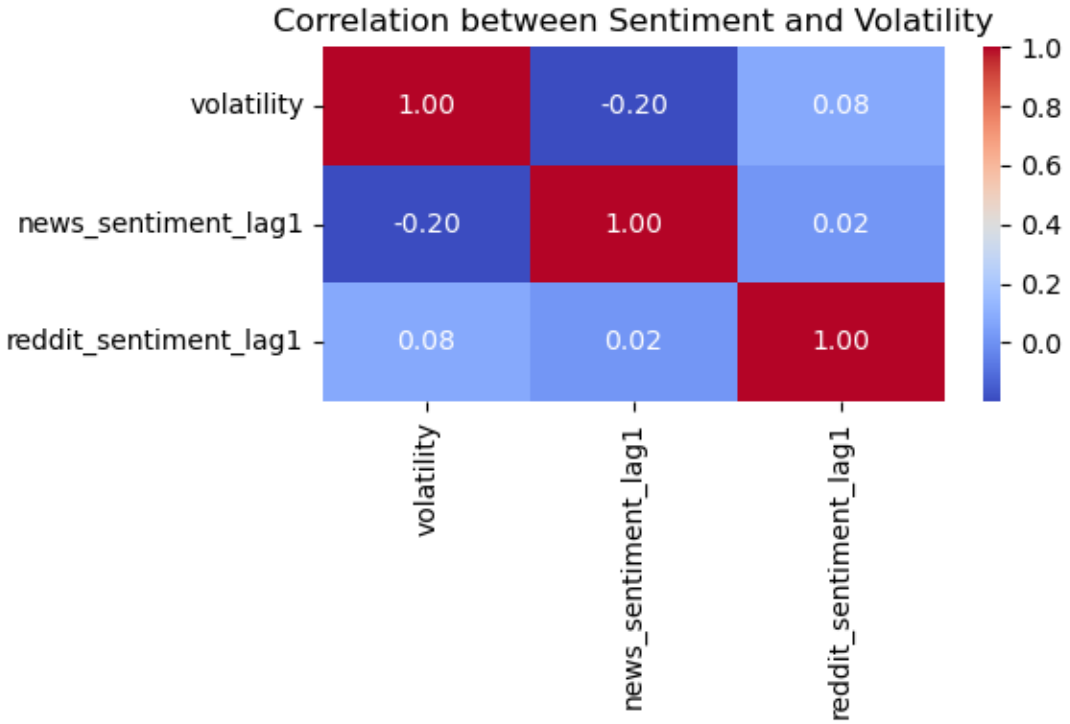
Same-Day, Lead, and Lagged Correlation:

I first explored the **temporal relationship** between sentiment and volatility by calculating Pearson correlations:

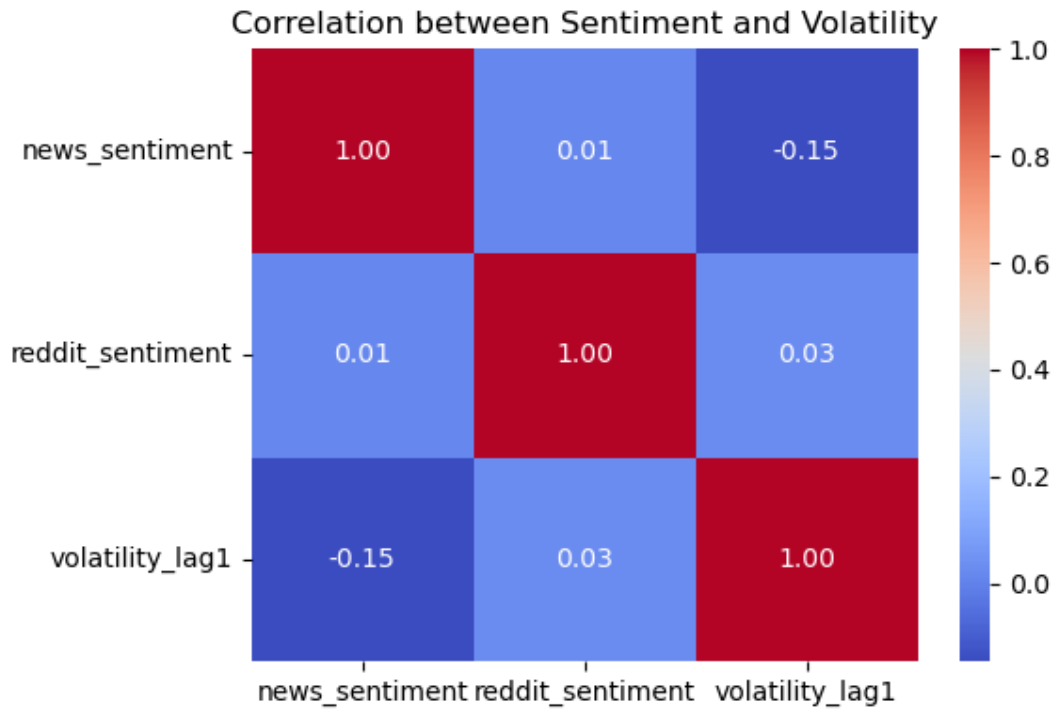
Same-day correlation:



Lagged sentiment (previous day sentiment vs current day volatility):



Lead sentiment (previous day volatility vs current day sentiment):

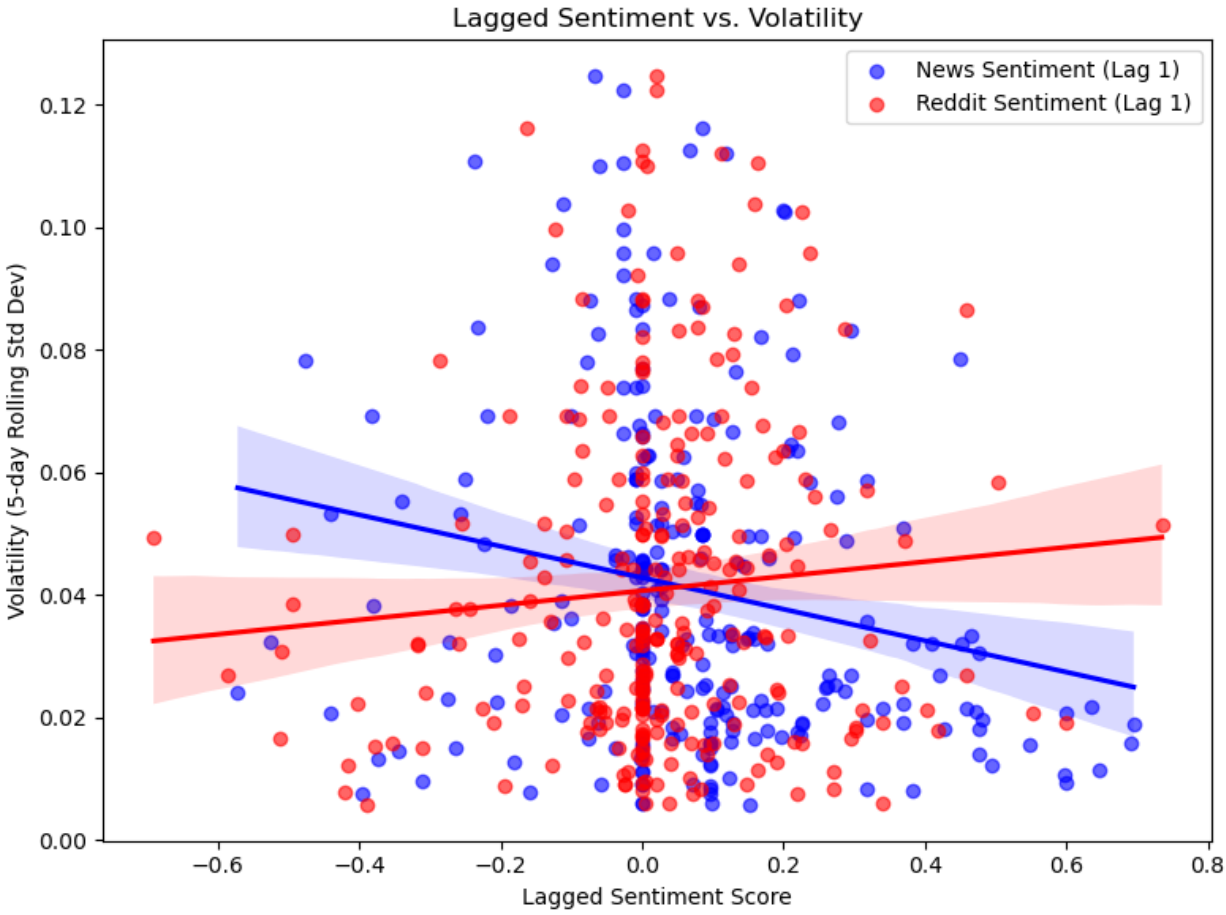


Correlation Type	News Sentiment	Reddit Sentiment
Same-Day	-0.18	+0.02
Lagged Sentiment → Volatility	-0.20	+0.08
Volatility → Lead Sentiment	-0.15	+0.03

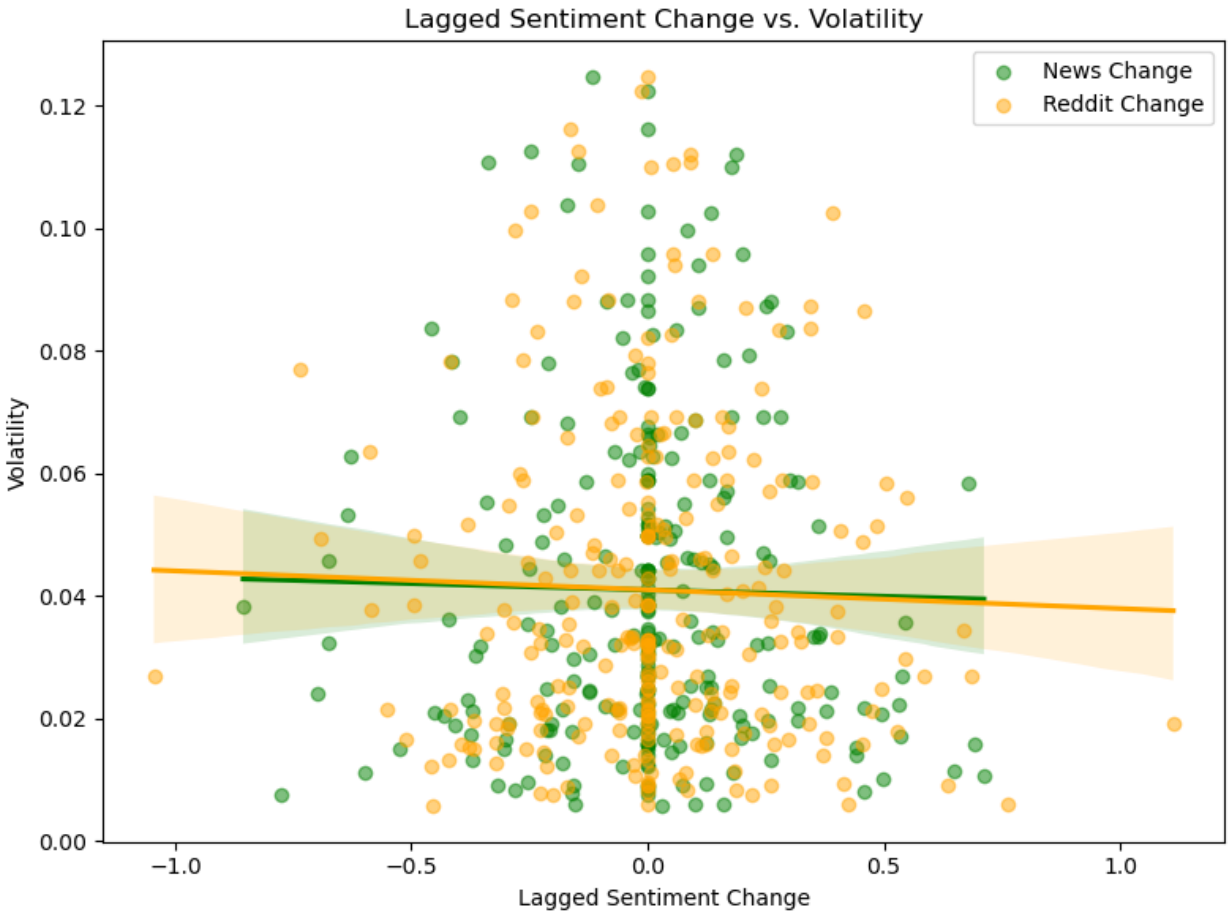
The strongest correlations appear when sentiment is lagged, suggesting **predictive potential** — especially for news sentiment.

This is in line with our hypothesis that daily changes in social and news sentiment correlates with next-day stock price volatility.

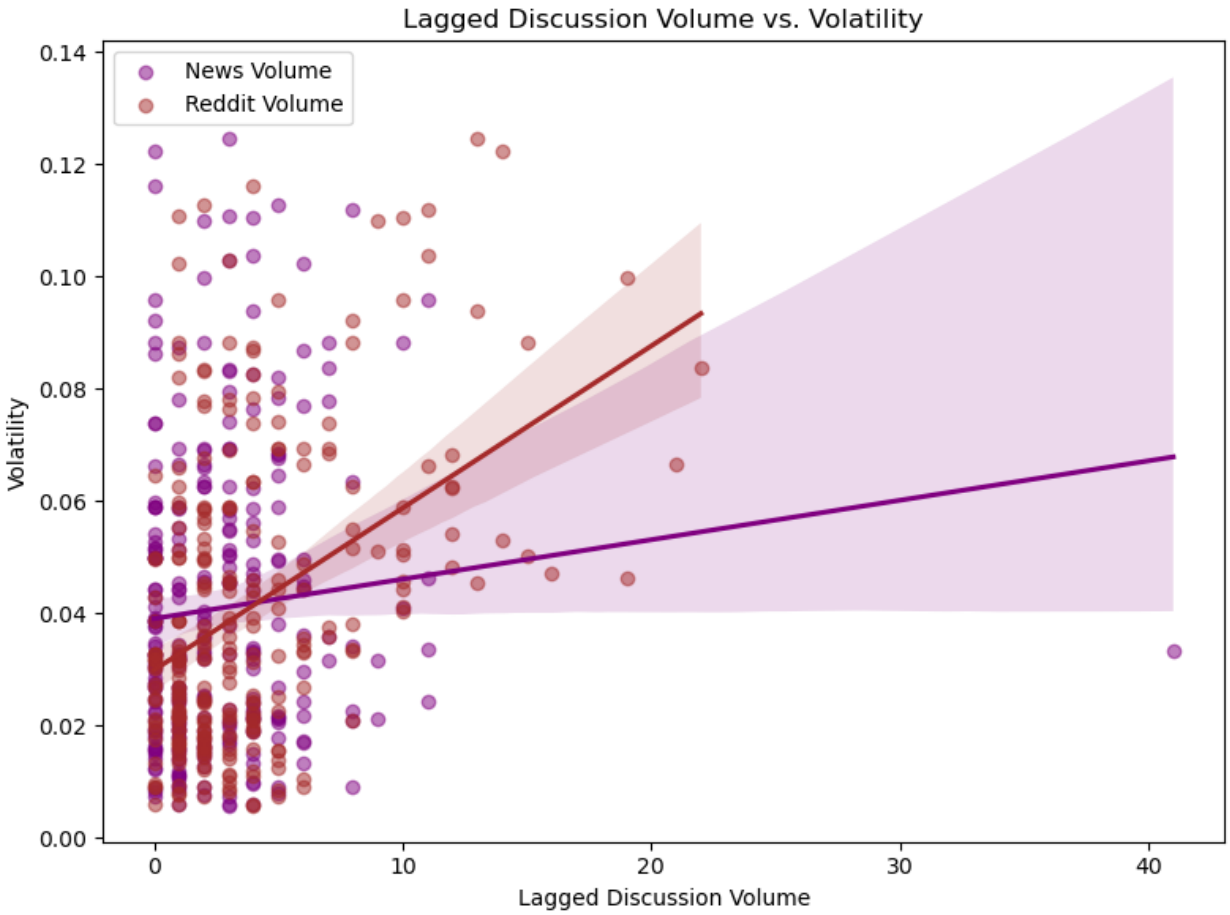
Sentiment and Volume vs. Volatility:



In the first plot, lagged sentiment scores from news and Reddit were plotted against 5-day rolling standard deviation of stock returns. The distribution shows a weak, but directionally distinct relationship: negative news sentiment was slightly associated with higher future volatility, while positive Reddit sentiment appeared to correspond with modest increases in volatility. This aligns with the hypothesis that user sentiment may reflect speculative attention, while negative news might signal risk.

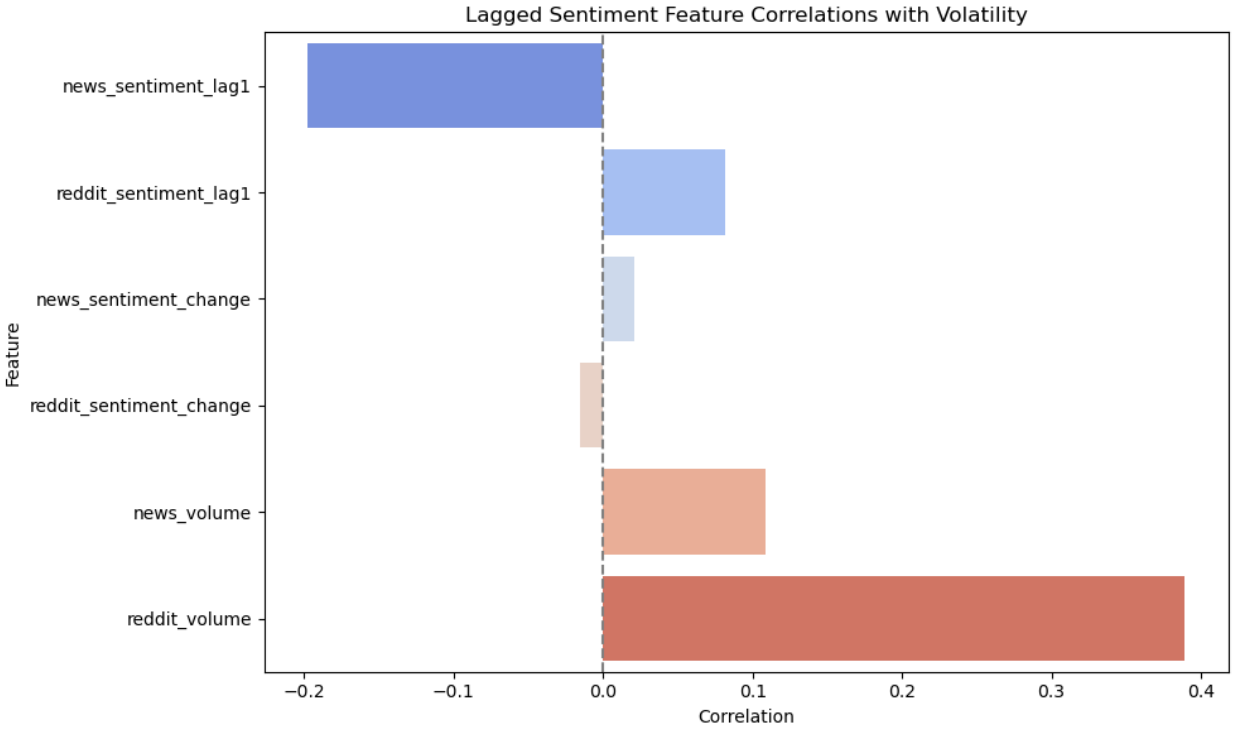


In the second plot, I analyzed the day-over-day change in sentiment scores. While the correlation was minimal overall, the visual dispersion suggests that sudden sentiment swings—especially on Reddit—can coincide with volatility spikes, though the relationship is not linear and exhibits noise.



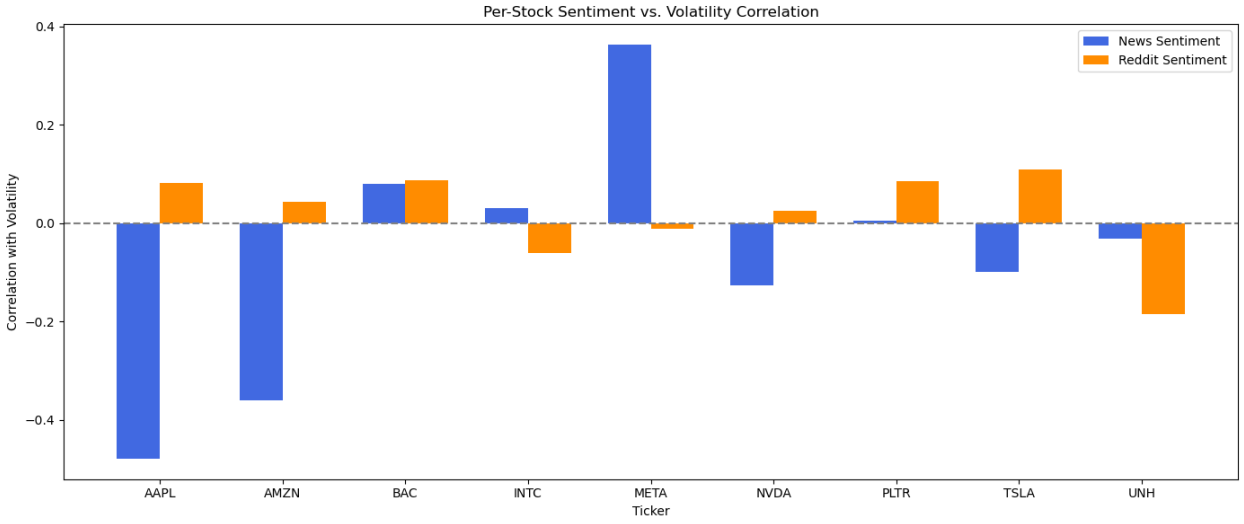
The third plot shows a clearer trend: increases in Reddit discussion volume were associated with higher volatility, suggesting that user engagement may act as a proxy for market uncertainty or trading interest. News volume showed a weaker, more stable positive relationship, reflecting how traditional media may capture broader market narratives but with less short-term variability than social media.

Correlation Plot:



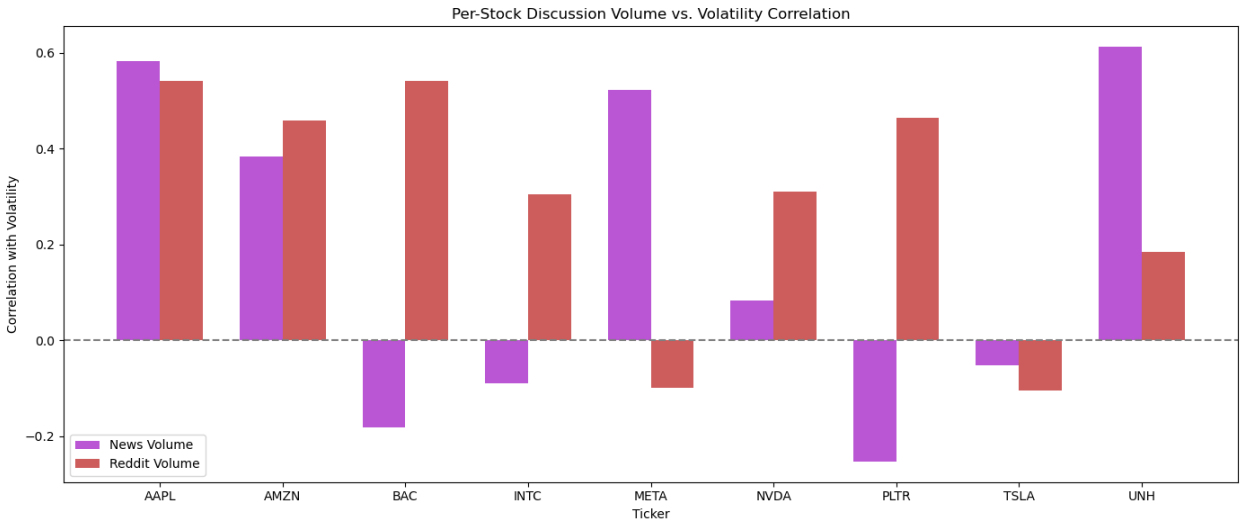
- **Reddit volume** had the **strongest correlation** with volatility (+0.39 overall), more than sentiment itself.
- **News sentiment** had a consistent **negative correlation** with next-day volatility.

Stock-Wise Effects:



I computed the Pearson correlation between daily volatility and previous day sentiment scores (both news and Reddit) on a per-ticker basis. The results show substantial heterogeneity. For instance, **META** exhibited a strong positive correlation with news sentiment, while **AAPL** and **AMZN** showed negative correlations. Reddit sentiment, in contrast, showed moderate positive alignment for some tickers such as **TSLA**, **PLTR**, and **BAC**, hinting at a possible community-driven reaction mechanism.

These patterns suggest that different companies respond to different information channels. Heavily institutional stocks may be more reactive to traditional news sentiment, while retail-favorite stocks may align more with Reddit discussions.



A similar per-stock breakdown of lagged discussion volume (i.e., number of articles or Reddit posts from the previous day) revealed a generally positive correlation with volatility. **UNH**, **AAPL**, and **META** in particular showed strong positive associations with news volume, suggesting heightened informational activity prior to volatility spikes. Interestingly, **BAC** and **PLTR** showed divergent behavior, with negative or flat correlations in news-driven volume, hinting at possible noise or desensitization in those tickers.

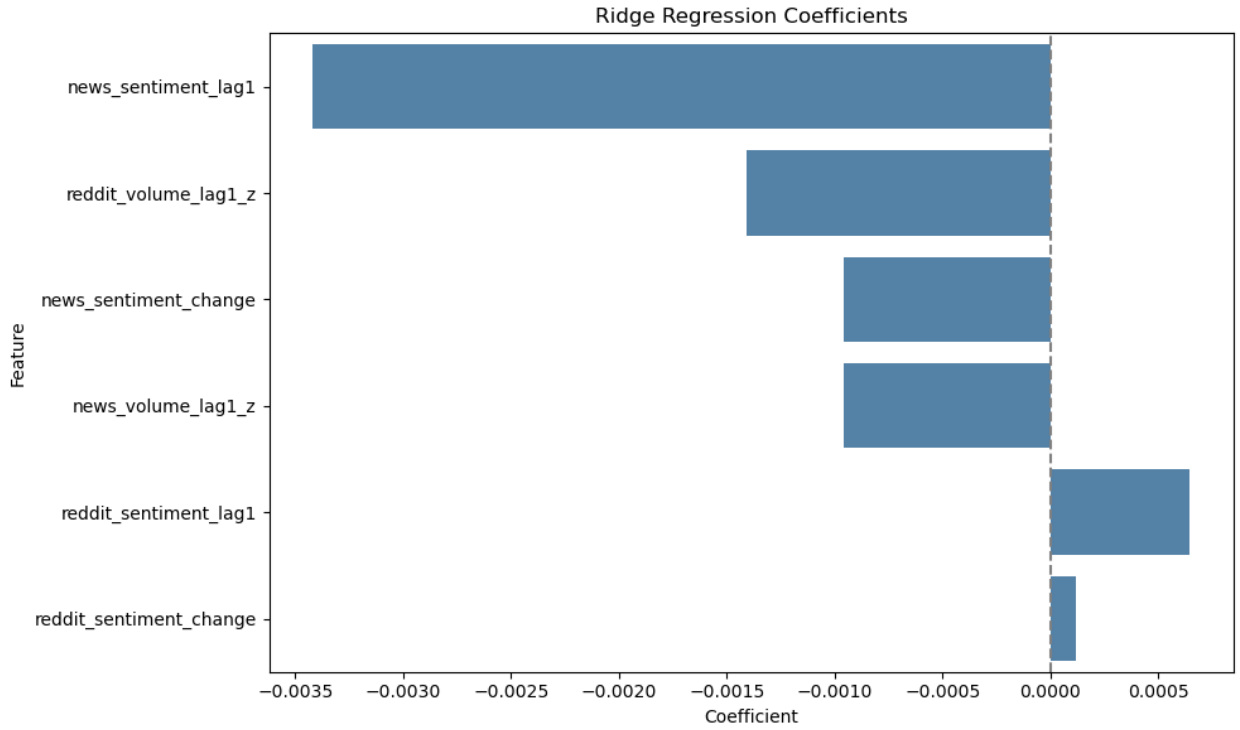
Reddit volume, however, showed more consistent positive correlations across all stocks—suggesting that increased retail engagement is broadly predictive of greater price movement, regardless of company fundamentals or sentiment direction.

Regression Modeling:

To evaluate the predictive power of sentiment features on market volatility, I built Ridge regression models using lagged sentiment, sentiment change, and discussion volume as predictors for the month of March 2025. Two variants of the model were constructed: one excluding the previous day's volatility (volatility_lag1), and one including it as an additional explanatory feature.

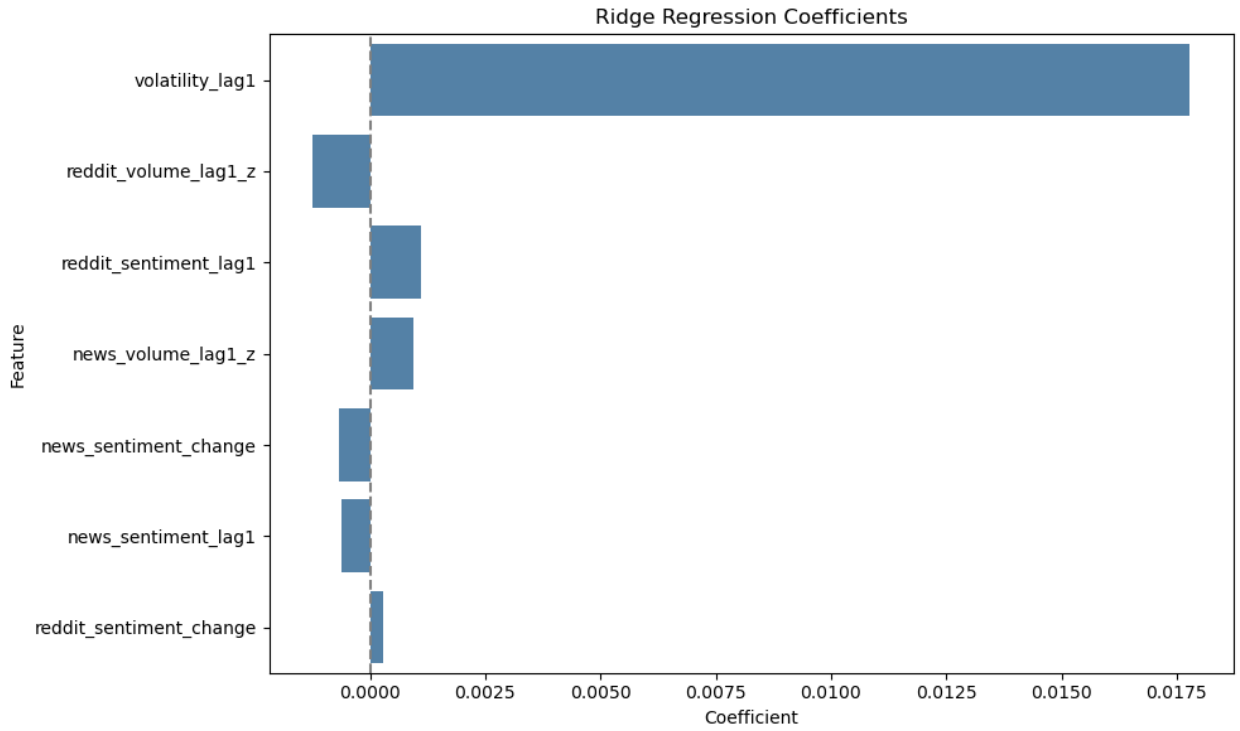
First model:

The **first model**, which relied solely on sentiment-driven features, the model showed limited explanatory power. It achieved a **Train RMSE of 0.0180**, **Test RMSE of 0.0259**, and **Test R^2 of 0.0256**, suggesting weak generalization performance. The strongest negative influence came from news_sentiment_lag1 (-0.0034), while reddit_volume_lag1_z also had a small negative contribution. This result aligns with earlier correlation observations indicating a weak linear relationship between isolated sentiment metrics and future volatility.

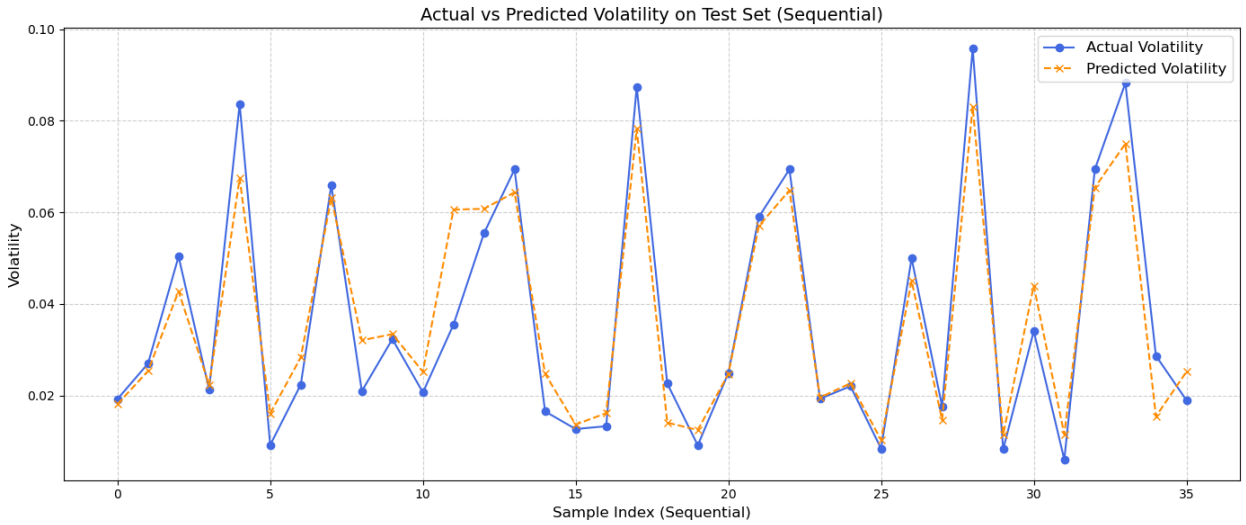


Second Model:

Introducing volatility_lag1 as a feature led to a **dramatic improvement in model performance**, with a **Train RMSE of 0.0090**, **Test RMSE of 0.0079**, and a **Test R^2 of 0.9088**, indicating that the model could explain over 90% of the variance in volatility during testing. The coefficient of volatility_lag1 (0.0178) dwarfed those of the sentiment features, underscoring its dominant predictive role.



Predicted vs. actual volatility:



Final takeaways:

- **Lagged sentiment** — especially from news — provides usable signal for **predicting volatility**.
- **Volume/attention metrics** (Reddit posts, headline counts) are **good predictors.** This suggests that attention, not just opinion, matters.
- Including **past volatility** enhances predictive power, but at the cost of **pure sentiment interpretability**.
- Model generalizes well across tickers, but **some stocks are more sentiment-sensitive** than others.

Limitations and Future Work

While our analysis provides compelling evidence of a relationship between online sentiment and stock price volatility, some limitations should be acknowledged.

1. Limited Time Frame:

Our dataset spans a relatively short period, encompassing only about 30 trading days. This narrow window restricts the generalizability of findings and may overfit to short-term market conditions or event-driven volatility.

2. Exclusion of Broader Market Indicators:

Our model focuses on sentiment and volume metrics but omits macroeconomic variables (e.g., interest rates, inflation, sector indices) and company fundamentals (e.g., earnings reports, financial ratios), which are known to influence volatility. Their inclusion could improve predictive accuracy and interpretability.

3. Simple Sentiment Scoring:

I used VADER, a lexicon-based sentiment analyzer, which while effective for short texts like headlines and Reddit posts, may oversimplify nuanced or sarcastic language. Fine-tuned transformer-based models (e.g., FinBERT) could better capture financial sentiment.

4. Uniform Modeling Across Tickers:

A single global model was used for all stocks, assuming uniform behavior across different tickers. However, our stock-wise analysis showed that sensitivity to sentiment varies. A per-ticker or sector-wise model could better capture these idiosyncrasies.

5. No Modeling of Interaction Effects:

Our ridge regression model does not account for interactions between sentiment and volume or nonlinear relationships. Future work could explore more expressive models such as random forests, XGBoost, or neural networks — potentially with attention to interpretability.

Future Work Directions:

- **Longer Time Horizon:** Extend the dataset to 6–12 months to analyze persistent patterns and reduce temporal noise.
- **Sentiment Momentum Modeling:** Include rolling average sentiment trends and variance to detect rising hype or fear.
- **Multi-source Attention Signals:** Incorporate Google Trends, Twitter, or YouTube metrics to enhance the "attention" signal.
- **Event Tagging:** Tag and isolate sentiment patterns around known events (e.g., earnings, policy announcements) to isolate causal relationships.
- **Alternative Target Variables:** Explore direction prediction (up/down) or classification of high vs. low volatility regimes.

Summary

This project explored the relationship between online sentiment and short-term stock price volatility using financial news and Reddit data. By engineering lagged sentiment and volume features, I examined whether sentiment precedes or reacts to market volatility. Our findings suggest that **lagged sentiment and discussion volume show modest but meaningful correlations with future volatility**.

A ridge regression model trained on March 2025 data achieved strong predictive performance, with prior volatility and Reddit volume emerging as the most influential predictors. Scatter plots, correlation matrices, and per-stock analyses further supported these insights. While the study demonstrates the potential of sentiment-driven forecasting, it also highlights the need for more complex models and longer time frames to build robust and generalizable systems.