ML Final project

Yun Lin, Mishael Khan

April 2025

1 Introduction

This project aims to develop a predictive model to support novice investors in making more informed portfolio decisions, specifically targeting small-cap stocks. By combining sentiment analysis of social media discussions with financial market data, the model is designed to forecast the daily direction of stock returns (up or down). Our analysis focuses on the period from 2022 to 2024 and incorporates a diverse set of data sources, including historical stock prices, Reddit posts from investment-related subreddits, and key fundamental indicators. The small-cap stocks analyzed in this study are selected based on the *U.S. News & World Report* article, "Best Small-Cap Stocks to Buy in 2025".

The full code, data, and documentation for this project are available at: `https://github.com/yunl39/STAT3106-ML-Final-Project`.

2 Data Description and Preprocessing

2.1 Stock Data

- Source: Yahoo Finance API (2022–2024)
- Variables collected: Close Price, Volume

2.2 Macroeconomic Indicators

Recognizing the influence of macroeconomic conditions on stock prices and investor expectations, we incorporated several macroeconomic indicators from the Federal Reserve Bank's database:

- Expected Inflation: `https://fred.stlouisfed.org/series/EXPINF5YR`
- Federal Funds Rate: `https://fred.stlouisfed.org/series/DFF`
- Real GDP: `https://fred.stlouisfed.org/series/GDPC1`
- Consumer Price Index (CPI): `https://doi.org/10.26509/frbc-ec-201002`
- Industrial Production Index: `https://fred.stlouisfed.org/series/INDPRO`

2.3 Reddit Data

We used Reddit post data available through Academic Torrents, which provides compressed `.zst` files containing full Reddit submissions. To parse these files, we followed the Python script provided in the Pushshift-Dumps GitHub repository.

From the dataset, we extracted posts from six finance-focused subreddits: `r/investing`, `r/stocks`, `r/wallstreetbets`, `r/StocksAndTrading`, `r/stockstobuytoday`, and `r/investingforbeginners`. We limited our extraction to posts created between January 1, 2022 and December 31, 2024.

For each post, we collected the full text content, number of comments, score (an indicator of engagement and influence), upvote ratio, and the creation timestamp. The initial extraction yielded approximately 266,000 posts. However, many entries were incomplete or missing the full post content. To ensure data quality, we filtered out all posts lacking complete content, resulting in a final dataset of 92,237 high-quality posts.

All subsequent analyses presented in this study are based on this cleaned and filtered dataset.

3 Feature Engineering

3.1 Stock Features

- Daily log return
- Daily change in Volume
- 7-day rolling variance
- Lagged daily log return and variance

These measures capture both the performance and volatility of individual stocks, as well as the time-dependent characteristics of returns.

3.2 Sentiment Features

We engineered several sentiment-related features from Reddit posts to capture investor mood and engagement, both at the market-wide and asset-specific level. Our sentiment analysis pipeline consists of the following steps:

- **Entity Tagging:** For each post, we created eight binary variables indicating whether the post mentioned one of the firms or sectors of interest. These dummy variables allow us to distinguish between general market discussions and posts specifically related to our target assets.
- **Mention Ratios:** Using the entity tagging indicators, we calculated the daily ratio of posts mentioning each firm or sector to the total number of posts. This feature quantifies the relative attention or hype surrounding specific entities, which can serve as a proxy for investor interest.
- **Sentiment Scoring with VADER:** We applied the VADER (Valence Aware Dictionary and sEntiment Reasoner) model to calculate a compound sentiment score for each post, ranging from -1 (most negative) to +1 (most positive). VADER is well-suited to social media content.
- **Daily Aggregation:** Post-level data were aggregated into daily metrics to align with financial data. These included:

- Total number of posts per day
- Daily average number of comments
- Daily average post score (influence)
- Daily average upvote ratio (community consensus)
- Daily average sentiment score, weighted by post score

These features reflect both the volume and quality of Reddit activity, which are known to affect short-term market dynamics.

- **Overall vs. Targeted Sentiment Indices:** We distinguish between two types of sentiment signals:
  - The *overall market sentiment score*, computed from all posts
  - *Firm- and sector-specific sentiment scores*, computed only from posts mentioning relevant entities

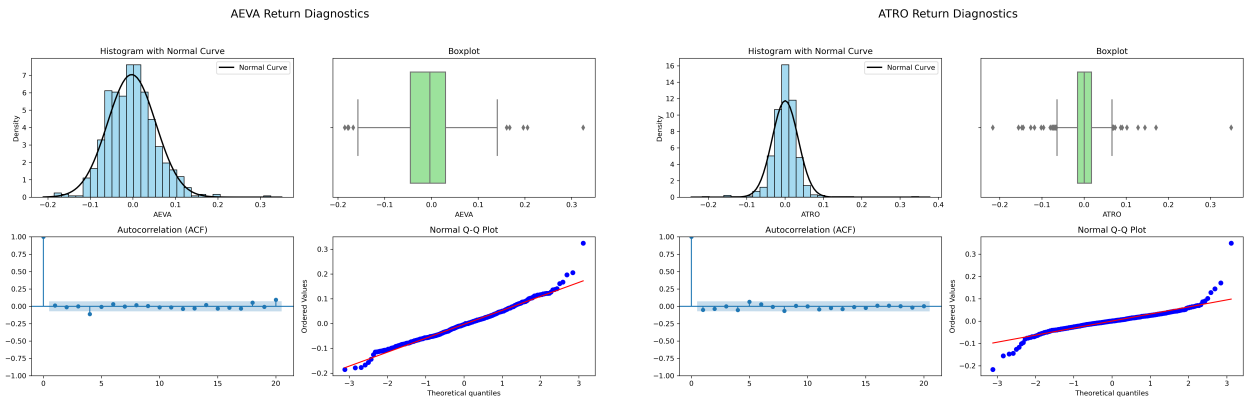
  This separation enables the model to learn both macro-level market mood and asset-specific sentiment effects.

- **Sentiment Categorization:** To improve interpretability and support classification-based modeling, we converted continuous sentiment scores into categories:
  - Positive (score > 0.05)
  - Neutral (between -0.05 and 0.05)
  - Negative (score $<$ -0.05)

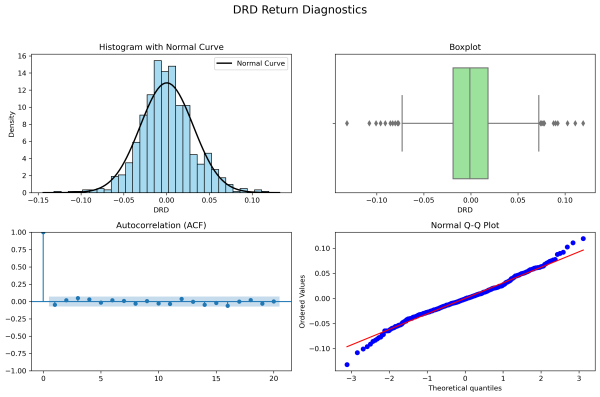
4 Exploratory Data Analysis

4.1 Return Diagnostics

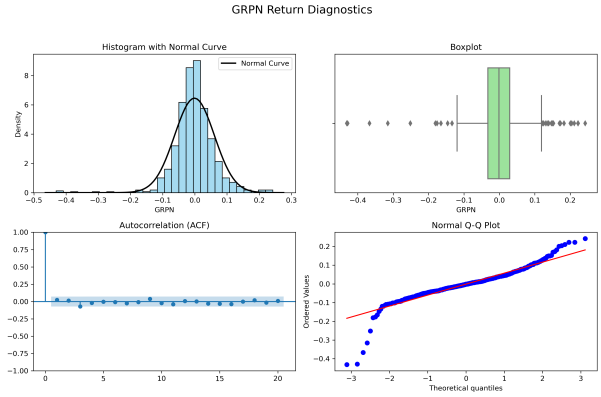
To evaluate the statistical properties of daily log returns, we generated diagnostic plots for each stock. Each plot includes a histogram with a fitted normal curve, a boxplot to identify outliers, an autocorrelation function (ACF) plot, and a Q-Q plot to assess normality.



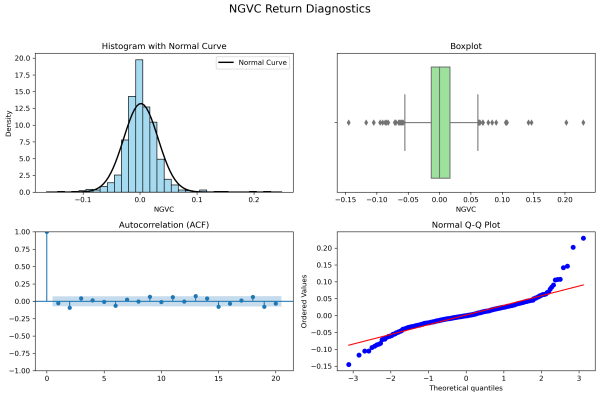
(a) AEVA (b) ATRO



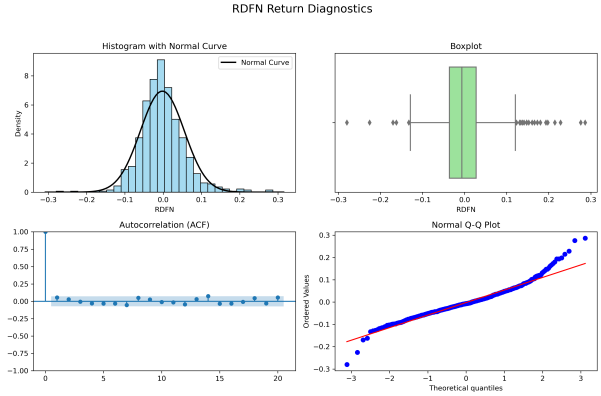
(a) DRD



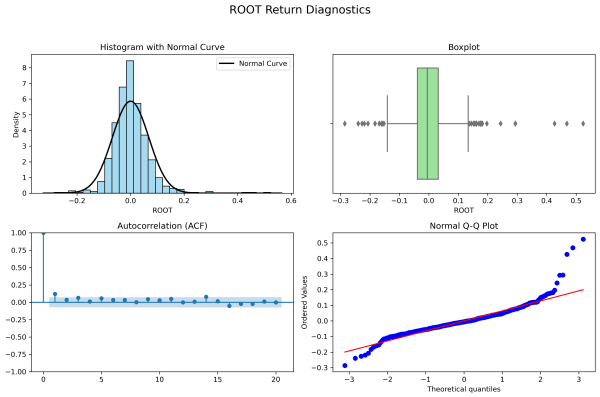
(b) GRPN



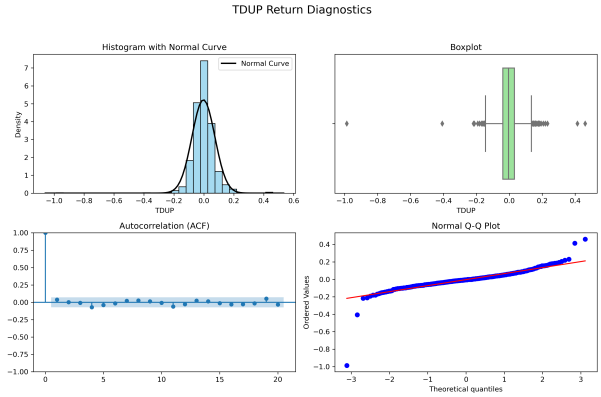
(a) NGVC



(b) RDFN



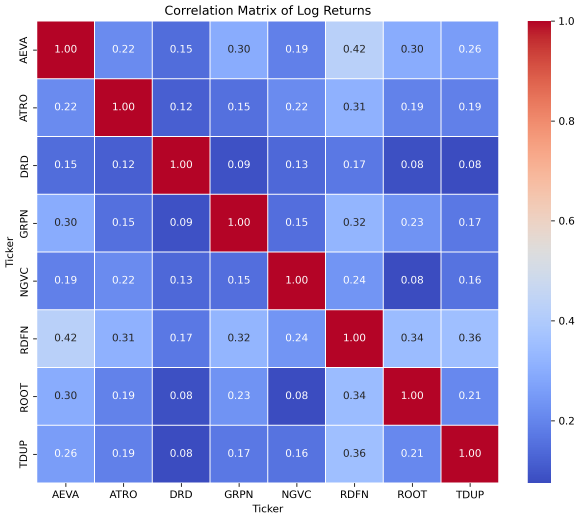
(a) ROOT



(b) TDUP

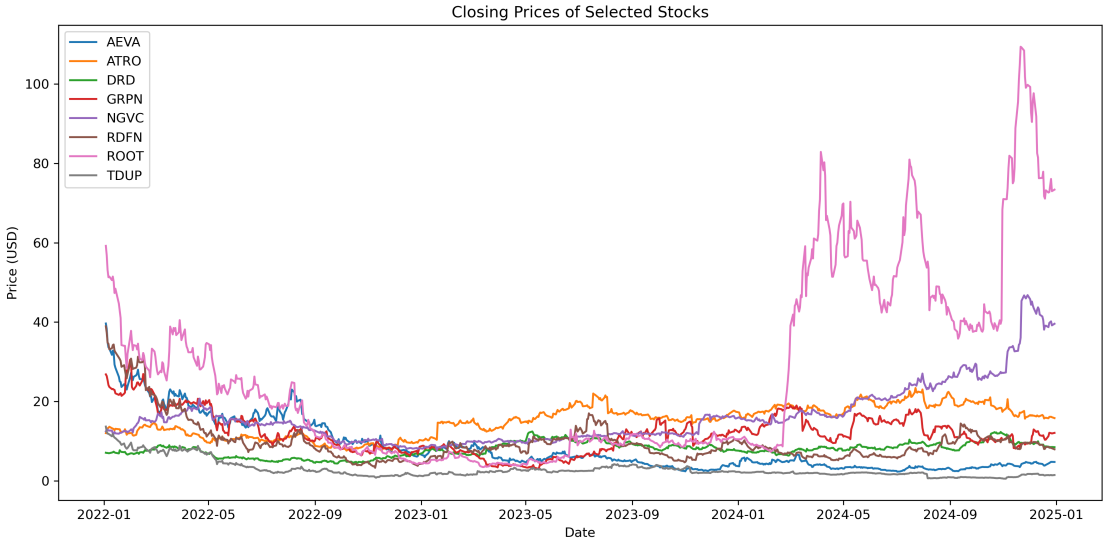
The return diagnostics suggest that most stocks exhibit approximately normal return distributions with some deviations in the tails and the presence of outliers. Autocorrelation is minimal, indicating little time dependence in daily returns. Overall, the return behavior is consistent with typical financial time series.

4.2 Correlation Analysis and Time Series Properties of Price and Return



(a) Heatmap of log return correlations

The correlation analysis reveals that the daily log returns of the selected small-cap stocks exhibit generally low to moderate correlations. This suggests that the returns are not strongly synchronized, indicating potential diversification benefits. As a result, constructing a portfolio using these stocks may help reduce overall portfolio risk through imperfect correlation among assets.



(a) Closing Prices of Selected Stocks

Closing prices are non-stationary and follow a stochastic trend, log returns are approximately stationary as the mean doesn't change over time. Therefore, we focus on modeling and predicting returns.

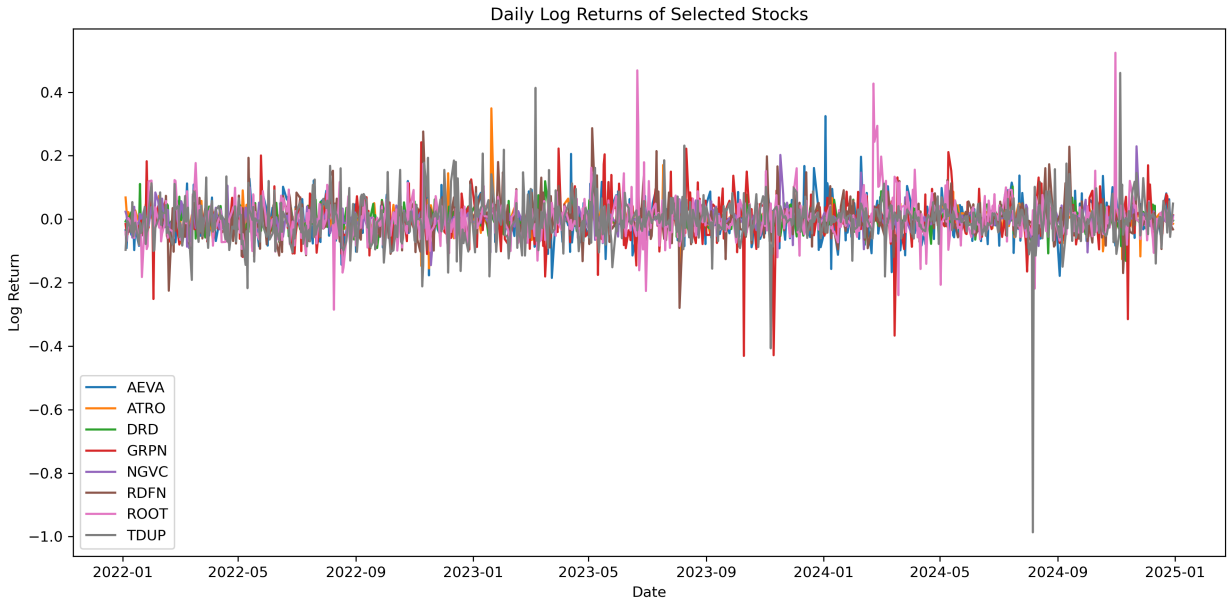


Figure 7: Daily Log Returns of Selected Stocks

4.3 Sentiment Distribution

4.3.1 Mention Frequency

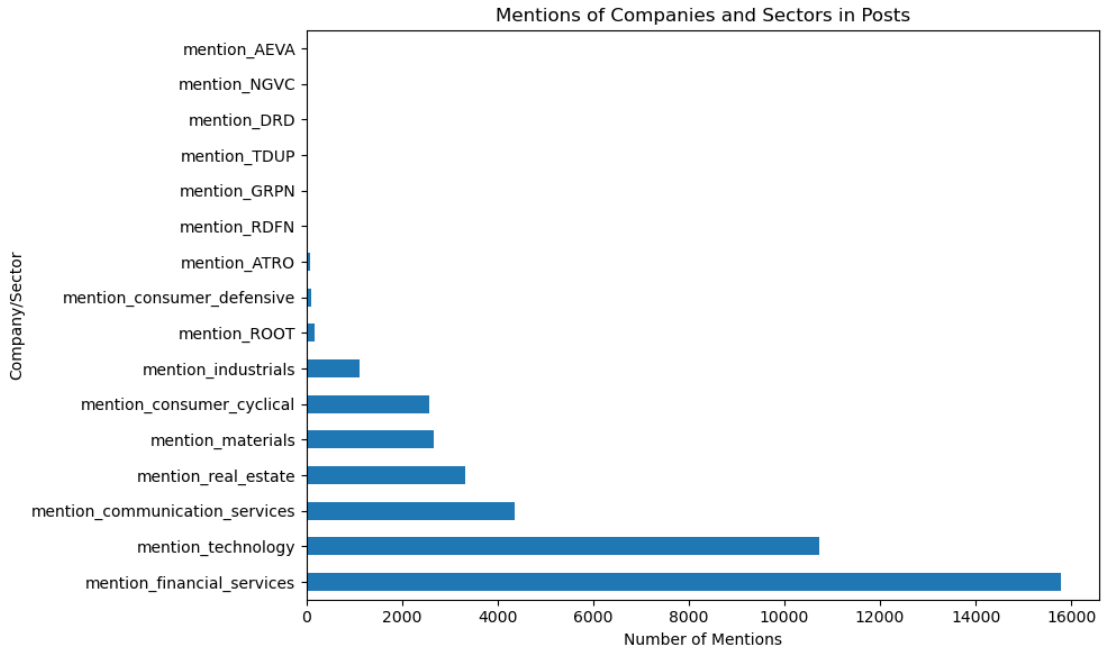
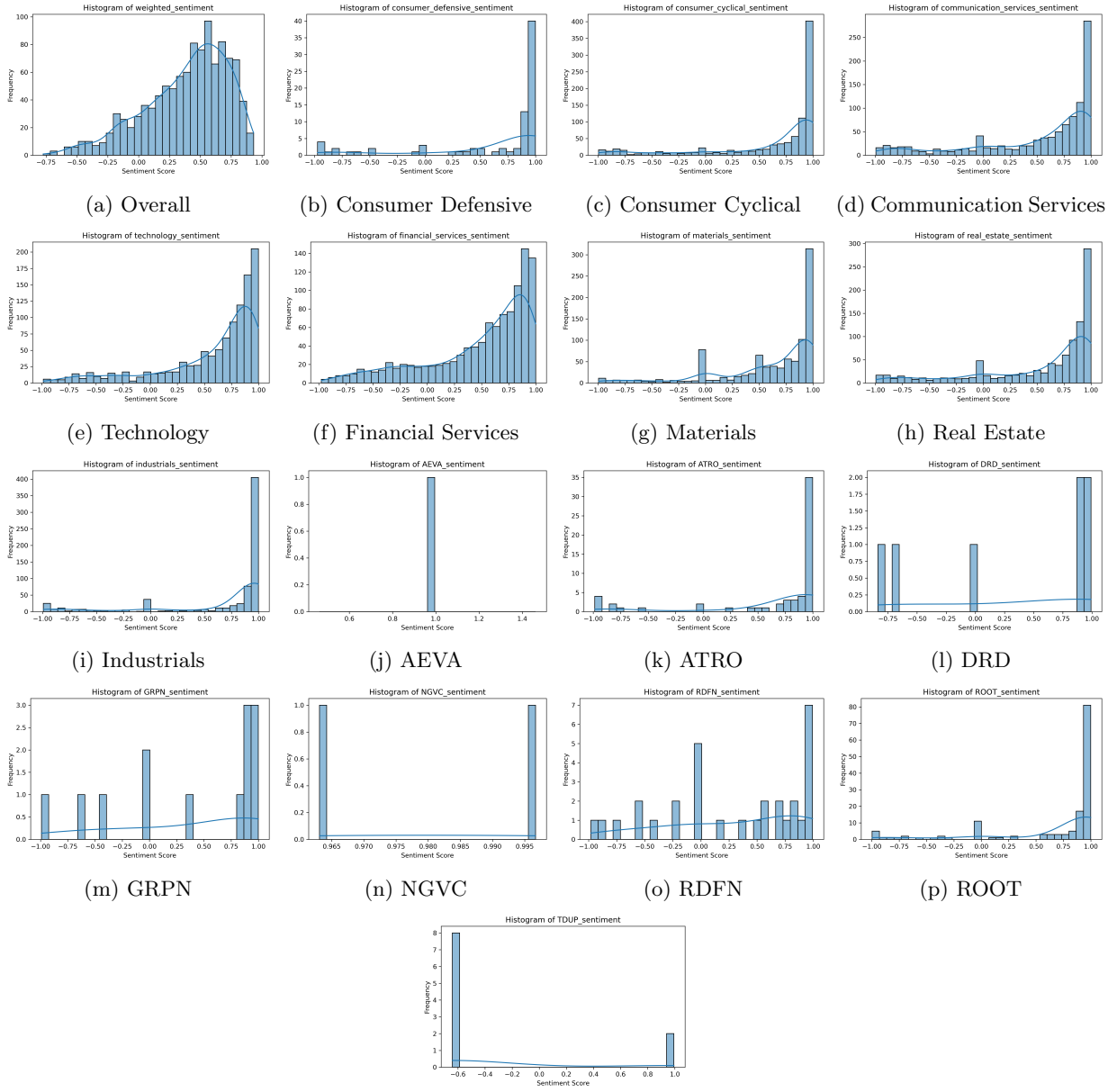


Figure 8: Number of mentions for each firm and sector in Reddit posts. Financial services and technology are mentioned far more frequently than other sectors or individual stocks.

4.3.2 Sentiment Distribution



(a) Overall (b) Consumer Defensive (c) Consumer Cyclical (d) Communication Services

(e) Technology (f) Financial Services (g) Materials (h) Real Estate

(i) Industrials (j) AEVA (k) ATRO (l) DRD

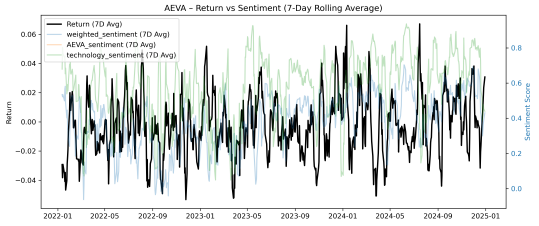
(m) GRPN (n) NGVC (o) RDFN (p) ROOT

(q) TDUP

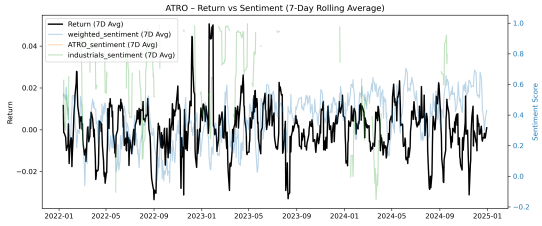
Figure 9: Sentiment score distributions for all posts (overall), individual sectors, and companies.

The sentiment distributions for sectors are strongly skewed, with most values clustered near the upper end, reflecting generally positive Reddit discussions. Firm-level distributions are often sparse due to limited mentions. Given the skew and non-normality, a log or similar transformation is recommended to improve symmetry and stabilize variance for modeling.

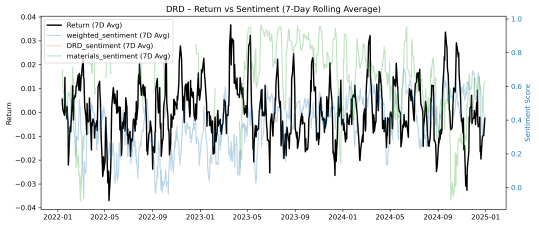
4.4 Return and Sentiment Dynamics



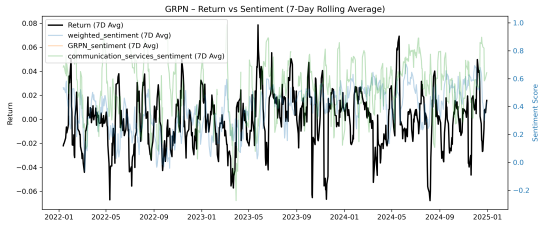
(a) AEVA



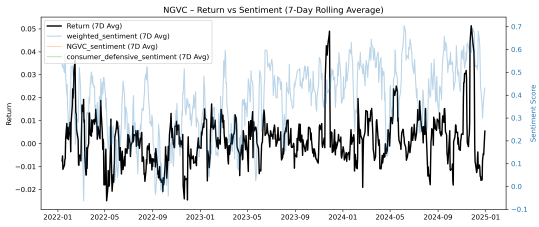
(b) ATRO



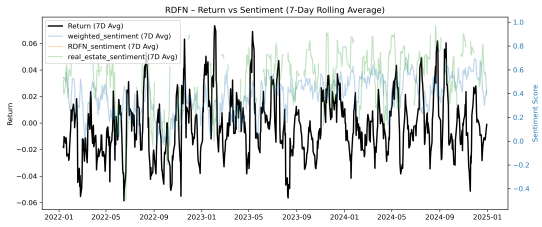
(c) DRD



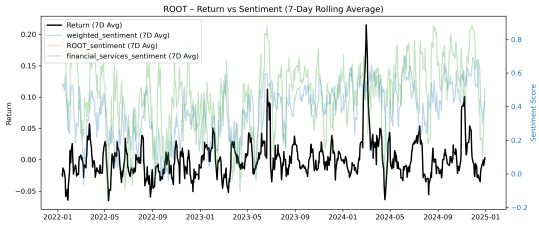
(d) GRPN



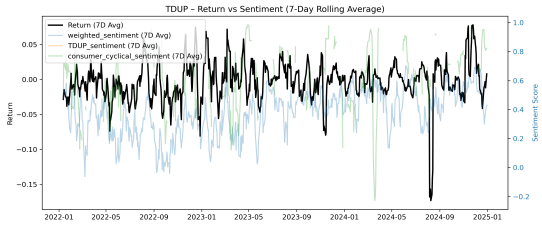
(e) NGVC



(f) RDFN



(g) ROOT



(h) TDUP

Figure 10: Return vs. Sentiment Time Series with 7-Day Rolling Average. Return is plotted on the left y-axis and sentiment scores on the right y-axis for each stock.

The time series analysis reveals an inconsistent but informative relationship between sentiment and stock returns. Sector-level sentiment occasionally moves in the same direction as returns, while overall market sentiment shows weaker alignment. Firm-level sentiment remains sparse due to limited mentions. Notably, both sentiment and return series often display spikes in volatility around the same periods, suggesting they respond to similar market events or shifts in investor attention. Further exploration revealed that sector sentiment tends to lead return movements by 2–3 days. This observed lag, along with the co-occurrence of volatility spikes, indicates that lagged sentiment features may capture useful predictive signals and should be considered in modeling.

5 Modeling and Evaluation

5.1 Problem Framing

Predicting the exact value of stock returns is inherently challenging due to market volatility and noise. To simplify the modeling task and focus on directional movement, we transform the continuous log return variable into a binary target defined as:

$$\text{Log\_Return\_Binary} = (\text{Log\_Return} > 0)$$

This transformation results in a classification target, where a value of 1 indicates a positive log return, and 0 indicates a negative log return. By framing the problem as a binary classification task, we aim to predict the direction of stock price movements rather than their exact magnitude. This approach is well-suited for decision-making scenarios such as trading signals.

5.2 Baseline Model

We first trained our models using a basic feature set made up of standard financial and macroeconomic variables to serve as a baseline comparison. These include:

- **7D_Rolling_Variance**: a 7-day rolling measure of return volatility.
- **Volume**: daily percentage change in volume.
- **7D_Rolling_Mean**:a 7-day rolling mean of returns.
- **Return_Lag1**: 1-day lagged return.
- **Volume_Lag1**: 1-day lagged percentage change in volume.
- All macro indicators in section 2.2.

We trained Random Forest, XGBoost and Elastic Net classifiers to predict the direction of stock log returns (up or down) for each tickers separately. The following graph display the result.

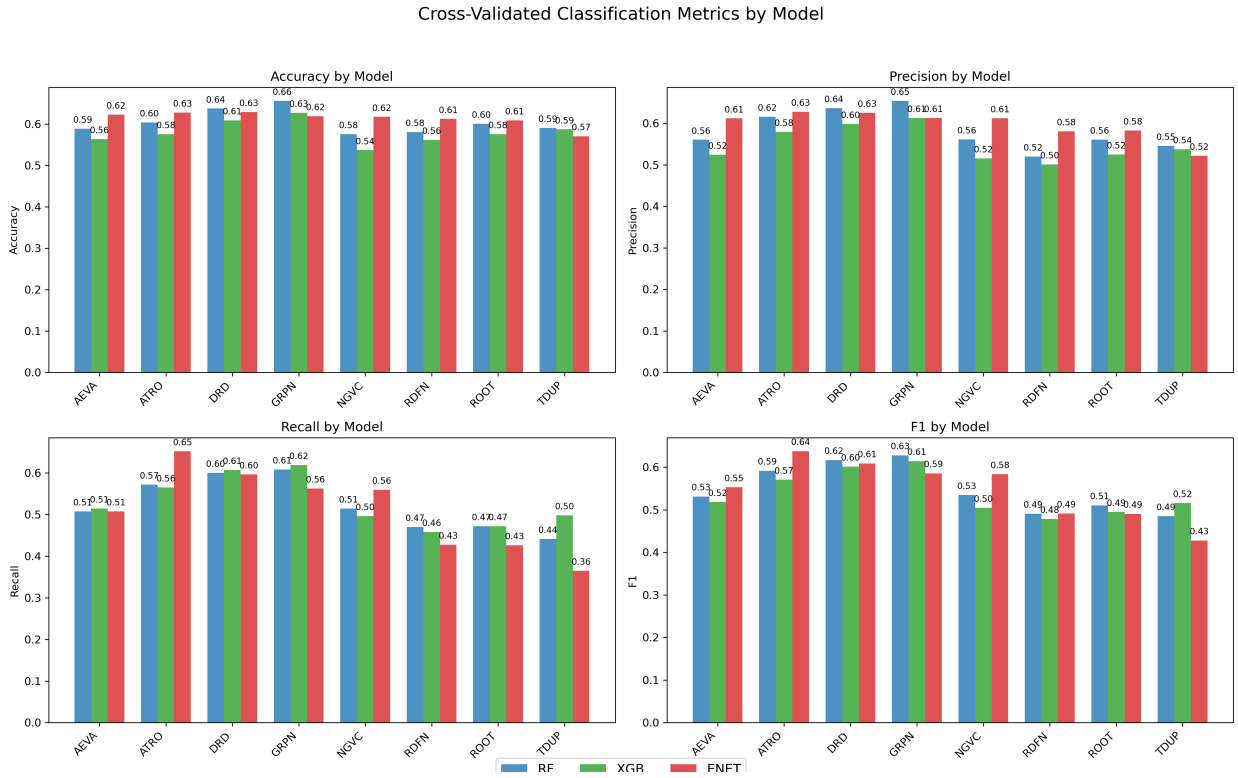


Figure 11: Cross-validated classification metrics (Accuracy, Precision, Recall, F1) for Random Forest (RF), XGBoost (XGB), and Elastic Net (ENET) across different tickers. Values shown are averaged across 5-fold cross-validation on the training set.

The classification performance across all models—Random Forest, XGBoost, and Elastic Net—shows limited predictive power, with most evaluation metrics (accuracy, precision, recall, and F1 score) ranging between 0.50 and 0.65. These results indicate that while the models slightly outperform random guessing (with accuracy 0.5), they struggle to reliably capture the underlying patterns in the data that drive short-term stock return direction. Among the models tested, Elastic Net consistently outperformed Random Forest and XGBoost across multiple metrics and tickers, likely due to its ability to handle multicollinearity and prevent overfitting through regularization.

These results are based on a feature set that excludes Reddit-based sentiment indices. The inclusion of sentiment variables—particularly those related to sector-level and may provide additional explanatory power and help capture investor mood and market signals not reflected in standard technical or macroeconomic indicators. Future modeling will incorporate these sentiment features to assess whether they significantly improve predictive accuracy.

5.3 Models with Reddit Data

5.3.1 PCA on Reddit Features

To address potential multicollinearity among Reddit-related features—such as number of posts, total comments, average comments, and average upvote ratio—we performed Principal Component Analysis (PCA). As shown in Figure 12, the first three principal components capture over 90% of the total variance, allowing us to reduce dimensionality while preserving most of the information. The loading heatmap in Figure 13

highlights that features like average score, total score, and comment-related metrics contribute most strongly to the first principal component, indicating their importance in capturing Reddit activity. These components are used as composite fatcors in subsequent modeling.

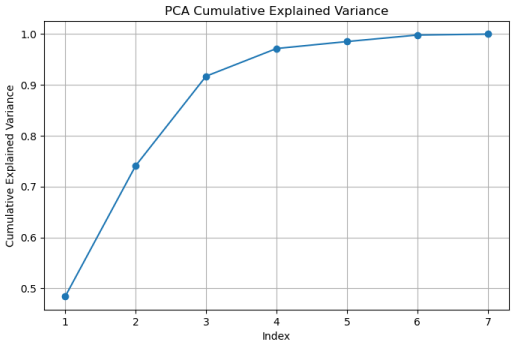


Figure 12: Cumulative explained variance from PCA of Reddit sentiment features.

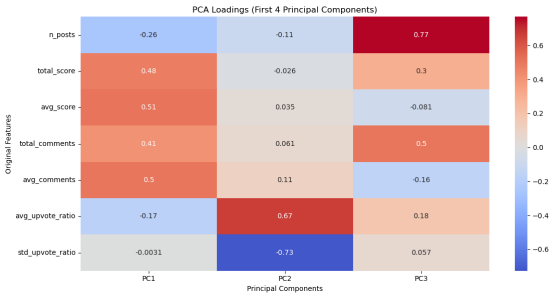


Figure 13: Heatmap of loadings for the first 4 principal components.

To determine the optimal number of lags for capturing autoregressive dynamics in return and volume, we estimated a series of Vector Autoregression (VAR) models using lag orders from 1 to 20. As shown in Figure 14, AIC and BIC increase steadily with more lags, indicating that adding lags beyond one degrades model performance. Based on this result, we retain only the 1-day lag for return and volume in our modeling framework.

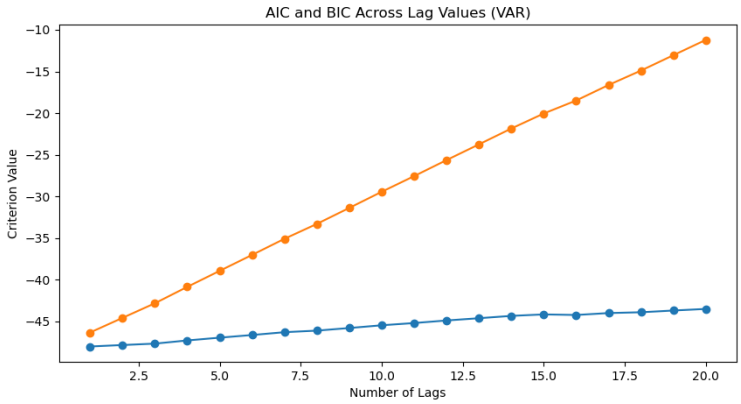


Figure 14: AIC and BIC values across different lag lengths in a VAR model. Both criteria suggest that additional lags increase model complexity without improving fit, supporting the use of a 1-day lag.

5.3.2 Model Training

After inlcuding the Reddit data, we have the following features

- **market_sentiment**: Reddit sentiment index aggregated across all posts.
- **sector_sentiment**: Reddit sentiment index specific to the stock's sector.
- **7D_Rolling_Mean**: 7-day rolling mean of stock returns.

- **7D_Rolling_Variance**: 7-day rolling variance of returns (proxy for volatility).
- **Volume**: Daily percentage change in trading volume.
- **ticker_freq**: Reddit post frequency for the specific ticker.
- **sector_freq**: Reddit post frequency for the stock's sector.
- **Return_Lag1**: 1-day lagged return.
- **Volume_Lag1**: 1-day lagged percentage change in volume.
- **Epi**, **CPI**, **FED_RATE**, **GDP**, **IPI**: Normalized macroeconomic indicators.
- **PC1**, **PC2**, **PC3**: First three principal components from Reddit post metrics (e.g., score, comments, upvotes).

We then refitted the same models including the Reddit-based sentiment features, but observed no improvement in predictive performance. In fact, the results slightly deteriorated. Next, we attempted using the categorized sentiment indices (positive, neutral, negative) as features, but the performance remained suboptimal and continued to underperform compared to models that excluded Reddit data altogether.

To account for potential lagged effects of sentiment, we shifted all Reddit-related features by two days, based on the hypothesis that discussions on Reddit may influence stock prices with a delay. This adjustment led to modest improvements in model performance, with accuracy scores generally ranging between 0.55 and 0.65. The figure below visualizes these results.

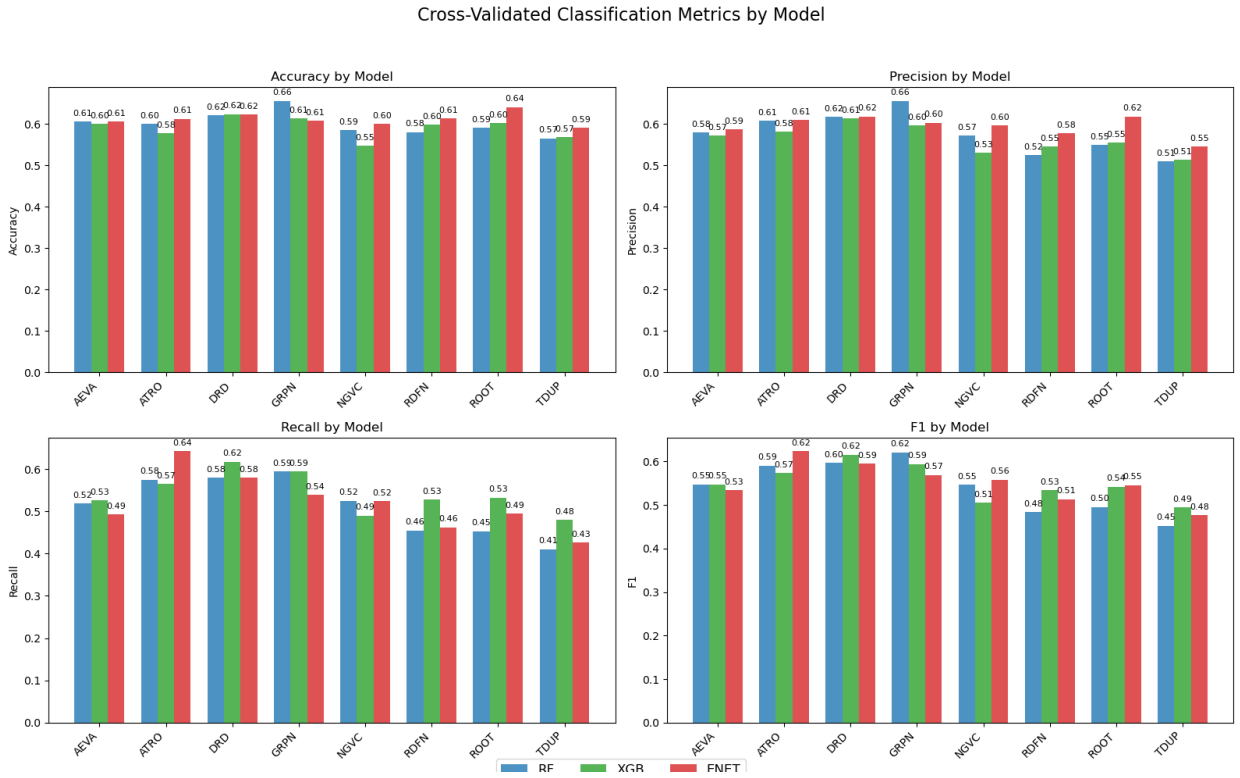


Figure 15: Cross-validated classification performance by model after incorporating categorized and lagged sentiment features. Most models still yield moderate results, with accuracy scores generally ranging from 0.55 to 0.65.

Given the limited predictive performance of our initial models, we applied Lasso regression as a feature selection technique to identify and filter out unimportant predictors. However, Lasso only excluded two features, indicating that most variables had at least some explanatory contribution. We retrained the classification models using the reduced set of features retained by Lasso, but the performance did not improve significantly. The accuracy and other evaluation metrics remained largely unchanged, suggesting that feature selection alone was insufficient to enhance model predictive power in this context.

5.4 Model Selection

Based on the performance of various models, we ultimately decided to include Reddit-based features in our predictive framework. To account for differences in model effectiveness across individual stocks, we selected the best-performing model for each ticker using the F1 score as our primary evaluation metric. The final model assignments are as follows: for AEVA, we use Random Forest; for ATRO, Elastic Net; for DRD, XGBoost; for GRPN, Random Forest; for NGVC, Elastic Net; for RDFN, XGBoost; for ROOT, Elastic Net; and for TDUP, XGBoost.

We trained our models on the training set and evaluated their performance on the held-out testing set. The results are presented below.

Table 1: Model Performance on Test Set by Ticker (Selected Model per F1 Score)

Ticker	**Accuracy**	**Precision**	**Recall**	**F1 Score**
AEVA	0.564	0.521	0.559	0.539
ATRO	0.604	0.595	0.667	0.629
DRD	0.570	0.561	0.514	0.536
GRPN	0.564	0.541	0.564	0.552
NGVC	0.631	0.635	0.556	0.593
RDFN	0.584	0.524	0.508	0.516
ROOT	0.638	0.607	0.515	0.557
TDUP	0.550	0.491	0.424	0.455

Although the model performance remains suboptimal and only slightly better than random guessing, we proceeded to simulate a trading strategy based on the model predictions. We then compared the resulting portfolio value to that generated by a random guessing approach.

We simulate a trading scenario using an initial budget of $10,000. Based on the model's predictions, the final portfolio value reaches $5,098.75, compared to $1,099.43 under a random guessing strategy. While the model does not produce a profit overall, it significantly reduces potential losses relative to random guessing. This demonstrates that our model, despite its limitations, can provide value to investors by improving directional predictions and outperforming naive strategies.

6 Conclusion and Future Work

In this project, we developed predictive models to forecast the daily return direction of selected small-cap stocks using a combination of historical market data, macroeconomic indicators, and Reddit-derived

sentiment features. Despite applying multiple machine learning algorithms—including Random Forest, XGBoost, and Elastic Net—and incorporating social sentiment indicators, the models achieved only modest performance, with accuracy and F1 scores generally ranging between 0.55 and 0.65.

Our findings suggest that Reddit sentiment signals, as currently constructed, do not provide strong predictive power for short-term stock return directions. This may be attributed to noise in social media discussions, low post frequency for small-cap stocks, or a lack of alignment between online discourse and actual market behavior.

A simulated trading exercise showed that, although the model did not generate a profit, it outperformed a random guessing strategy by reducing losses.

For future work, we recommend the following improvements:

- **Alternative Data Sources**: Incorporate additional or higher-quality sentiment data, such as Twitter, news sentiment, or analyst reports, to improve signal strength.
- **Lagged and Aggregated Effects**: Explore alternative time lags and cumulative sentiment windows to better capture delayed market responses.
- **Model Customization**: Fine-tune models for individual tickers, considering sector-specific behaviors and volatility.

In summary, while Reddit sentiment alone does not appear to be a reliable predictor in this setting, the framework established in this project provides a solid foundation for iterative improvement and richer data integration in future research.