A Preliminary Investigation into the Relationship Between Social Media Sentiment and Future Equity Returns

1 Introduction

Social media has transformed how information is disseminated and consumed by investors. Social media platforms allow for real-time discussions on financial markets, where users often express opinions and sentiments regarding equities. Prior studies have indicated that such sentiment, especially on platforms with significant user bases, may influence trading behavior and price formation. This paper investigates the relationship between sentiment expressed on a social network and future stock returns. Specifically, we examine whether aggregate sentiment on equities predicts their future returns for 2,000 liquid U.S. stocks between 2015 and 2020.

2 Methodology

We model the stock return over the next 5 days ($d_{5,t+1}$) as a function of social media sentiment and trading volume. We construct two key variables: rolling sentiment and abnormal volume.

- Rolling sentiment is the sum of daily stock-level sentiment over a five-day period.
- Abnormal volume is meant to capture how information-rich the data is for a given stock and date. It is calculated as:
    1. Define "normalized volume" as the stock-level message volume divided by the log of that stock's Average Daily Trading Volume (ADTV) in dollars.
    2. Define $\text{NormalizedVolume}_{\text{mean},5}$ as the five-day rolling mean of normalized volume. Similarly, define $\text{NormalizedVolume}_{\text{mean},63}$ and $\text{NormalizedVolume}_{\text{sd},63}$ as the 63-day rolling mean and standard deviation of normalized volume respectively.

3. Then, abnormal volume is the "z-score"

$$\frac{\text{NormalizedVolume}_{\text{mean},5} - \text{NormalizedVolume}_{\text{mean},63}}{\text{NormalizedVolume}_{\text{sd},63}}$$

The final right-hand side variables are z-scored before regression to be approximately normal.

2.1 Time-Weighted Rolling Sums

When computing rolling sums, means, and standard deviations on sentiment and volume, we apply a time-weighted linear-decay to give more importance to recent data over earlier data. Specifically, the weight over a w-day window is:

$$\text{Weight}(t, w) = \max(0, t + w)$$

This gives a linear decay between $(-\text{window}, 0)$ and $(0, \text{window})$. We then normalize so that weights sum to one.

2.2 Returns

The y-variable is five-day close-to-close returns. The returns are lagged by one trading day to account for time trading in and out of positions. Specifically, for date d:

- All right-hand side variables are available before market-open on d.
- Y is the return of each ticker from the close of d to $d + 5$.

2.3 Regression Models

To estimate the relationship between stock returns and sentiment, consider the following two regressions:

1. In the first regression, we use **rolling sentiment** as the sole independent variable:
$$d_{5,t+1} \sim \beta_1 \text{Rolling Sentiment}_t + \epsilon_t$$
2. The second regression uses an interaction between **rolling sentiment** and **abnormal volume**:
$$d_{5,t+1} \sim \beta_1 \text{Rolling Sentiment}_t \times \text{Abnormal Volume}_t + \epsilon_t$$

We fit both models cross-sectionally, pooling data across tickers.

Model	R^2	Estimate	Std. Error	t value
Model 1	4.068e-06	-7.558e-05	6.442e-05	-1.173
Model 2	6.034e-06	-9.196e-05	1.018e-04	-0.904

Table 1: Coefficients and R^2 values for both regression models

3 Results

3.1 Coefficients and R^2 Values

Our analysis reveals that neither models explains a significant portion of equity returns, as indicated by the very low R^2 values. However, we strongly believe in the underlying concept—that social media sentiment could play a meaningful role in predicting future stock returns. There are several avenues for future work that could improve upon this initial study:

- Optimizing feature construction: Specific choices when constructing even simple features can make a large difference. Perhaps modifying the existing features will have a large effect on predictive power.
- Creating additional features: Developing more sophisticated features could help capture more information and relationships that might be missed with the current set.
- Exploring other social networks: Different platforms may have varying levels of information richness. Incorporating data from other social networks could improve the robustness of the models.
- And so much more...

While this study shows that the current model is not sufficient, we believe that further research in these directions could unlock significant predictive power from social media sentiment.