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. Platforms like X, formerly known as Twitter, 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, referred to as “Gamma”, and its impact on stock returns. Specifically, we examine whether aggregate sentiment on equities predicts their future residual returns for 2,000 liquid U.S. stocks between 2015 and 2020.

2 Data

The data is sourced from the Gamma platform using Natural Language Processing (NLP) algorithms. First, raw messages are tagged with relevant companies using Named Entity Recognition (NER). Then, for each company-relevant message, the message is classified as negative, neutral, or positive. Messages with multiple companies or low confidence in sentiment estimation are ignored. For each stock, the data is aggregated daily at the stock level. Specifically, it has the sum of sentiment scores from messages, the total number of messages, and the number of unique users who posted those messages.

3 Methodology

We model the residual stock return over the next 5 days ($sd5_{t+1}$) 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 $\text{NormalizedVolume}_{\text{mean},10}$ as the ten-day rolling mean of normalized volume. Similarly, define $\text{NormalizedVolume}_{\text{mean},126}$ and $\text{NormalizedVolume}_{\text{sd},126}$ as the 126-day rolling mean and standard deviation of normalized volume respectively.
  3. Then, abnormal volume is the "z-score"

$$\frac{\text{NormalizedVolume}_{\text{mean},10} - \text{NormalizedVolume}_{\text{mean},126}}{\text{NormalizedVolume}_{\text{sd},126}}$$

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

3.1 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 weights 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.

3.2 Returns

The y-variable is five-day close to close residual 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 constructed by residualizing the market return of each security from the close of d to *d + 5*.

Residual returns are the part of a stock's return that cannot be explained by overall market movements or other known factors, like industry trends. It's what's left after accounting for the influence of those factors. In other words, the stock's performance based on its own unique characteristics or events, independent of broader market influences.

Here, residual returns are calculated using the Fama-French 3-factor model, which accounts for three key factors: market risk, company size, and value versus growth. The model captures how much of a stock's return can be attributed to overall market movements (market risk premium), whether the stock is from a small or large company (size premium), and whether the company is categorized as value or growth based on its book-to-market ratio (value premium). The residual return is the difference between the actual return of the stock and the return predicted by these three factors.

Mathematically, that is:

$$R_i = \beta_1 R_m + \beta_2 \text{SMB} + \beta_3 \text{HML} + \epsilon_i$$

Where:

- R_i is the stock's market return,
- R_m is the market return,
- SMB (Small Minus Big) is the size factor,
- HML (High Minus Low) is the value factor,
- ϵ is the residual return of stock i - the return unexplained by the model.

We construct the y variable of residual returns in the above fashion.

3.3 Regression Models

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

1. In the first regression, we use **rolling sentiment** as the sole independent variable:
$$sd5_{t+1} \sim \beta_1 \text{Rolling Sentiment}_t + \epsilon_t$$
2. The second regression uses an interaction between **rolling sentiment** and **abnormal volume**:
$$sd5_{t+1} \sim \beta_1 \text{Rolling Sentiment}_t \times \text{Abnormal Volume}_t + \epsilon_t$$

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

4 Results

4.1 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:

Model	R^2	Estimate	Std. Error	t value
Model 1	9.063e-07	3.424e-05	2.841e-05	1.205
Model 2	8.643e-07	3.305e-05	2.962e-05	1.116

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

- 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.