Analyzing the Correlation Between Retail Traders' Sentiments and Equity Market Movements

Haozhe Zeng | Cornell University | hz657@cornell.edu

Zixiao Wang | Cornell University | zw699@cornell.edu

ABSTRACT

This research seeks to explore the impact of retail traders' sentiments, primarily from forums like WallStreetBets, on equity market movements. The investigation will discern the duration of this correlation, whether it's short-term or extends to mid-long term. It will also ascertain if the correlation is more pronounced in specific stock categories like penny stocks or tech giants or if such a correlation might be absent altogether.

INTRODUCTION:

In recent years, the financial landscape has undergone a significant shift, largely driven by the digital transformation of trading platforms. This transformation has democratized access to financial markets, enabling a surge of retail traders to participate actively in stock trading. These traders, often characterized by their nimbleness and ability to mobilize quickly, have emerged as a formidable force in the equity market, challenging traditional institutional players.

One of the most prominent platforms that has come to symbolize this new wave of retail trading is the WallStreetBets forum on Reddit. Serving as a discussion hub, WallStreetBets has become a focal point for retail traders to share insights, strategies, and sentiments about various stocks. The power of such collective sentiment became glaringly evident during events like the GameStop short squeeze, where concerted buying actions driven by discussions and emotion on the forum led to unprecedented stock price surges, catching many institutional investors off guard.

However, while isolated events like the GameStep short squeeze have been widely reported, a systematic analysis of the broader influence of such forums on the equity market remains to be explored. This project endeavors to dissect the sentiments echoed in these forums and their potential correlation with stock market trends. Beyond just identifying correlations, the study aims to probe the depth and breadth of this influence. Questions arise such as: Are these correlations consistent across different stock categories? Do sentiments from these forums have a more pronounced effect on certain types of stocks, like penny stocks or tech giants? And importantly, can these sentiments be harnessed predictively to anticipate stock market movements?

By delving into these questions, this project seeks to provide a comprehensive understanding of the interplay between retail trader sentiments and equity market dynamics in the digital age.

METHODOLOGY

Data Collection:

Use web scraping tools to extract comments and posts from sites like WallStreetBets (Reddit) and Twitter

Data Organization using Hash Tables

Organize and index the scraped data efficiently using ticker symbols as keys and associated posts as values. Ensure quick retrieval of relevant posts for specific stock analysis.

Data Deduplication using Bloom Filters:

Implement Bloom filters to identify and eliminate duplicate entries, ensuring data integrity. Use Bloom filters for membership queries to determine mentions of specific stocks or keywords.

Data Labeling:

Manually label a subset of data to gauge general sentiment. Use this labeled dataset as a benchmark to evaluate automated sentiment analysis tools or train custom models if necessary.

Noise Reduction:

Filter out irrelevant posts using keyword filtering. Identify and remove spam or bot-generated content to maintain data purity.

Data Analysis:

Align sentiment data with stock market data for concurrent time frames. Conduct correlation analysis to determine the relationship between sentiment scores and stock market movements. Utilize visualization techniques to provide a clear representation of findings.

Expected Outcomes:

Upon completion of this research, we anticipate deriving a comprehensive dataset that seamlessly integrates retail traders' sentiments with stock market data, offering a panoramic view of the intricate interplay between online sentiments and market dynamics. We aim to gain profound insights into the tangible influence that online forums, notably platforms like WallStreetBets, exert on stock market trajectories. This exploration will not only spotlight specific events or dominant sentiments that tangibly impact stock valuations but also delve into the nuanced dynamics between different stock categories. For instance, while we hypothesize that retail traders might wield significant influence over penny stocks, given their lower market capitalization and trading volumes, their sway over major sectors like QQQ or SPY is expected to be minimal, if not negative. This potential inverse correlation could stem from the overwhelming presence of institutional investors in these sectors or the sheer trading volume that buffers these stocks against sentiment-driven volatilities. Through this multifaceted analysis, our research seeks to paint a detailed picture of the modern stock market, where age-old trading conventions intersect with the democratized trading ethos of the digital age.

Future Scope:

While our current research primarily focuses on the influence of retail traders in the stock market, the role of institutional traders remains an intriguing avenue for future exploration. Institutional traders, with their vast resources and market knowledge, have traditionally been the dominant players in the equity market. As we move forward, we are keen on delving deeper into understanding the magnitude and nuances of their market impact. Key questions we aim to address include: How do institutional traders' strategies and decisions shape market trends? Is there a discernible correlation between stock movements and the reports or analyses released by these institutions? By juxtaposing the influences of retail and institutional traders, we hope to provide a more holistic understanding of the modern stock market's dynamics and the interplay of its various actors.

Plan (ONE MONTH):

First Week

1. Determine the specific data categories we aim to gather. Evaluate potential data sources, including Youtube, brokerage platforms like Robinhood, Twitter, and Reddit.

2. Develop strategies and tools for efficient data extraction from the identified sources. This could involve exploring APIs, web scraping tools, or third-party data providers.

Second Week

1. Cleaning: Identify and handle missing values, outliers, and any inconsistencies in the data.

2. Merging: Combine datasets from different sources in a coherent manner, ensuring alignment in terms of time frames, data types, and other relevant parameters.

3. Labeling: Manually or using automated tools, label the data to identify the sentiment (e.g., positive, negative, neutral). This step is crucial for subsequent sentiment analysis tasks.

Third Week

1. Model Selection: Determine the types of NLP models suitable for the task. Consider traditional models like Naive Bayes, SVM, or more advanced ones like LSTM, BERT, or Transformer-based models.

2. Training: Use the labeled data to train the selected NLP model(s). This involves splitting the data into training and validation sets, setting up the training loop, and monitoring the model's performance.

3. Quantification: Explore quantitative methods to represent the data. This could involve techniques like TF-IDF, word embeddings (Word2Vec, GloVe), or even transformer embeddings. The goal is to convert textual data into numerical form for model training and analysis.

Fourth Week

1. Model Selection for Correlation Analysis: Based on the results from Week 3, decide on the type of model to use for correlation analysis. This could be:

    a) Linear Models: Such as Linear Regression if the relationship appears to be linear.

    b) Non-linear Models: Such as Decision Trees, Random Forests, or Support Vector Machines if the relationship seems to have non-linear patterns.

    c) Deep Learning Models: Such as Neural Networks or LSTM if the data has complex patterns or sequential dependencies.

2. Training: Train the chosen model using the quantified data from Week 3 and stock market movement data. Ensure to split the data into training, validation, and test sets to evaluate the model's performance accurately.

3. Evaluation: Assess the model's ability to discover correlations between retail trader sentiments and market movements. Use metrics like R-squared, Mean Absolute Error, or others relevant to the model type to quantify the model's accuracy.

4. Temporal Analysis: Given that stock market data is time-series data, consider analyzing the lag between sentiment changes in online forums and stock market reactions. This can help in understanding how quickly the market reacts to shifts in retail trader sentiments.

5. Comparative Analysis: Rank the platforms based on the strength and significance of their correlation with stock market movements. Use visualization tools to represent these findings.

Potential Citations:

Online Social Media and Stock Market:

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8

Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. Journal of Economic Behavior & Organization, 107, 730-743.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies, 27(5), 1367-1403.

Sentiment Analysis and Opinion Mining:

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

Kumar, A., & Lee, C. M. (2016). Retail investor sentiment and return comovements. The Journal of Finance, 61(5), 2451-2486.

NLP Techniques for Financial Markets:

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.