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

Haozhe Zeng | Cornell University | hz657@cornell.edu

ABSTRACT

To explore the impact of retail traders' sentiments, as articulated on forums like WallStreetBets, on equity market movements. This investigation aims to discern the duration of this correlation (short-term or mid-long term), ascertain if the correlation is more pronounced in specific stock categories like penny stocks or tech giants, or determine if such a correlation might be absent altogether.

INTRODUCTION:

The digital transformation of trading platforms has emboldened retail traders, making them a formidable force in the equity market. Forums such as WallStreetBets have not only served as discussion hubs but have also catalyzed significant stock market events: GameStop short squeeze. This project endeavors to dissect the sentiments echoed in these forums and their potential correlation with stock market trends, while also probing the depth and breadth of this correlation across different stock categories.

APPROACHES

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:

A comprehensive dataset combining retail traders' sentiments and stock market data. Insights into the influence of online forums on stock market trends. Identification of specific events or sentiments that have a clear impact on stock prices.

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.

Future Scope:

Institutional Trader Influence: Investigate the impact of posts, analysis, and reports made by institutional traders such as banks, hedge funds, and other financial institutions. This can help differentiate between the influence of retail versus institutional sentiments.

Deep Dive into Specific Sectors: Focus on specific sectors (e.g., tech, pharmaceuticals, energy) to understand how sentiments on different platforms influence sector-specific stocks.

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.