Correlating Reddit Sentiment and Market Returns

Abe Moore Odell

Mentor: Andrew Forney

Discipline: Computer Science

Introduction

In the past few years there has been an emergence of "meme stocks," defined by Investopedia as "shares of a company that have gained a cult-like following online and through social media platforms" [5]. One of the main social media drivers for these meme stocks has been the internet platform Reddit—a website that allows users to share blog-like posts and interact with other users through comments and voting mechanics. Reddit is the most popular online forum on the internet [6], but its effects on the stock market remain relatively unstudied. While Reddit has proven to be an effective platform for influencing the price of individual stocks such as in the short-squeeze of GameStop, which saw the stock price jump from $13 to $200 a share in just a year [3], the influence of Reddit on broader macroeconomic trends is less clear. In my research I will ask, "is there a broader correlation between the sentiments of Reddit posters and wider market performance?" To answer this question, I will compare the "mood" of the top 1,000 daily Reddit posts to market index performance from the same day in order to find any potential correlation.

Background and Related Work

Another social media platform Twitter, with similar blogging and commenting mechanics as Reddit, has already been widely studied in search of market correlations. One study has shown a strong correlation between Tweet volume and stock trading volume [13]. A study similar to the one in this proposal has found a strong correlation between the language used in Tweets and the market performance of related stocks [11].

Additional research has been done to find methods for finding the sentiment of Twitter posts using both NLP (natural language processing) [11] as well as more simplistic methods involving lists of negative and positively associated words [9]. The researchers in "Correlating

Financial Time Series with Micro-Blogging Activity," found that this correlation can be used as an effective tool for stock traders. Other research, such as "Mining of Concurrent Text and Time Series," has found similar methods of data mining and analytics effective as a trading tool when analyzing news sources.

In order to simplify my research and decrease programming and data collection time, I will utilize the same list of negatively and positively associated words used in "Detecting Subjectivity and Tone with Automated Text Analysis Tools." Which, given a large enough sample size of Reddit posts, should prove to be effective in determining tone and sentiment.

Motivation

Knowing to expect a bear (downward) or bull (upward) market is extraordinarily useful for both individual investors and large investment firms. Imagine that you had the information to sell your stocks before a market crash, or buy at the market's nadir. At its core, this study searches for a market indicator; does a more positive Reddit correlate with a bull market? Does a more negative Reddit correlate with a bear market?

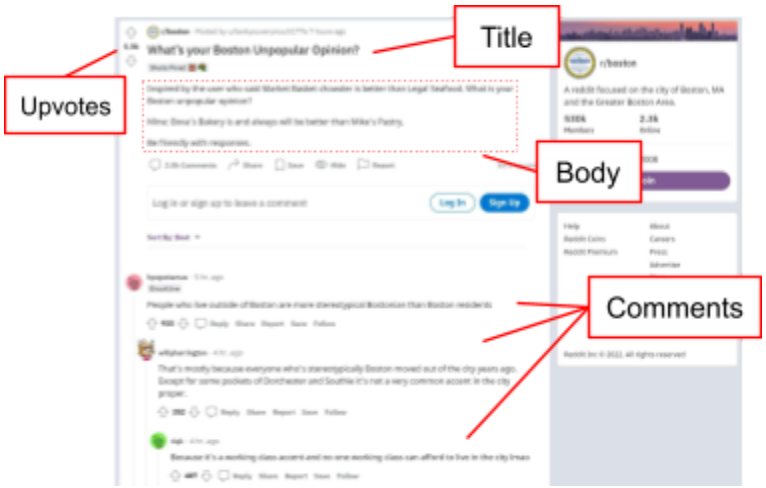
This study, if successful, would also lay the groundwork and provide justification for further in-depth studies. If the experiment does yield the expected results, some interesting questions are raised for further studies, such as utilizing NLP (natural language processing done through machine learning) or artificial intelligence applied to both examining the posts, and analyzing the data once collected.

Additionally, it would be beneficial for further, higher budget studies to conduct longer term data collection for more conclusive results. A more selective method of gathering Reddit posts, based on language patterns or keywords that map more closely to specific stock tickers selected based on the language of the post may also lead to a stronger correlation. For example,

using NLP to determine that a post relates to agriculture, and relating that to agricultural commodity index funds could lead to more directly usable findings, but is outside the limited scope of this study.

A more simple variation on the previous experiment would be to collect the top posts from different popular Subreddits (forums on Reddit sorted by category and topic). This is a variation that I would consider doing if I am able to get the program working soon enough. However, this more in depth method of data collection would require a larger amount of data collection, storage space, and processing power.

Methods



[Figure 1.] An example Reddit post with key elements highlighted. This research would gather the text from the Title and the Body sections only.

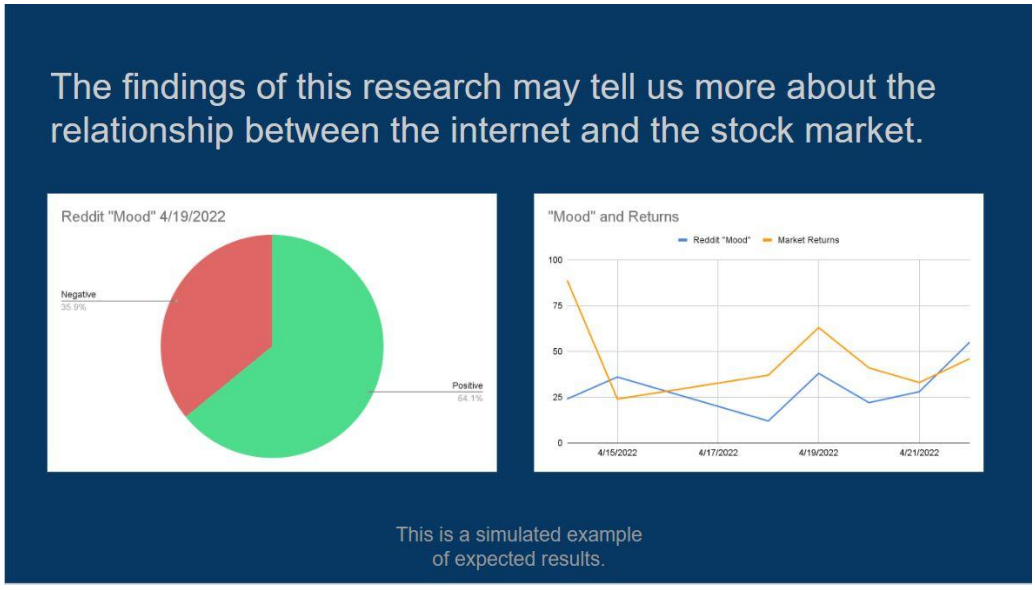
In order to conduct this research, I will collect the top 1,000 reddit posts from each market trading day. To do this, I will use a custom program that will utilize the Pushshift API and PRAW (the Python Reddit API Wrapper) to gather the title and body text of these posts. The program will then analyze the title and body text of each post for key terms using a list of positive and negatively associated words, specifically the same Harvard IV-4 list used in "More

Than Words : Quantifying Language to Measure Firms' Fundamentals," which uses language analysis of investing firms reports, correlating the language used, and whether it was mostly negative or mostly positive to the returns experienced by the firm over the period the report covers [10].

I will also collect market data—specifically, the change in price in three major market indices: the S&P 500, the Dow-Jones Industrial Average, and the Nasdaq-100—from each corresponding market day. As this is just the opening and closing prices of these three indexes, this information will be gathered manually from one of any number of easily accessible online sources such as the Wall Street Journal's historical prices page [14].

I will then graph these two metrics (aggregate Reddit "mood" and the overall stock market indicators) and analyze the data for long and short term trends and correlations. If any trends or correlations are found, this could prove very useful for both trading large firms and average investors.

Expected Results



[Figure 2.] Sample results for both an individual day and wider time span.

I expect to see some correlation between average Reddit “mood” and stock market returns. It is likely that the market will be down on days with more negative post language, and up on days with more positive language. This relationship will also likely show that days that are with negative language will likely have larger drops in the market and very positive language days will have larger gains.

However, it may be unclear in which direction this correlation goes, if it does prove to exist. Do reddit posts affect stock prices (as we saw with GameStop), or do down market days lead to worse moods across the platform? Or, do Reddit sentiments and markets move together for the same external reasons?

Conclusion

While Reddit is a relatively new platform, its influence on the stock market through meme stocks is undeniable. By comparing Reddit sentiment and market returns, we can examine the platform's influence on the broader market and gather more insight into how the internet affects stock prices and vice versa. We will ask, “does a pessimistic Reddit correlate with lower market prices? Does an optimistic market correlate with a more positive Reddit?” Using computer aided data collection techniques such as the Pushshift API and PRAW I will explore the relationship between Reddit and the market, and answer the question: is there a correlation between Reddit sentiment and market returns?

References

[1] “2010 flash crash,” *Wikipedia*, 19-Feb-2022. [Online]. Available: https://en.wikipedia.org/wiki/2010_flash_crash#Explanation. [Accessed: 24-Feb-2022].

[2] C. D. Aenlle, “A.I. has arrived in investing. humans are still dominating.,” *The New York Times*, 12-Jan-2018. [Online]. Available:

https://www.nytimes.com/2018/01/12/business/ai-investing-humans-dominating.html. [Accessed: 24-Feb-2022].

[3] Dailey, Natasha. “Reddit Traders Are Waxing Nostalgic over the Pre-Short Squeeze Days of Gamestop - from $13 A Year Ago to $200 Today, Many Are Holding on for More Gains Ahead.” Business Insider. Business Insider, November 24, 2021. https://markets.businessinsider.com/news/stocks/reddit-traders-nostalgic-gamestop-stock-short-squeeze-gme-2021-11.

[4] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, “Correlating financial time series with micro-blogging activity,” *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, 2012.

[5] Hayes, Adam. “Meme Stock.” Investopedia. Investopedia, March 24, 2022. https://www.investopedia.com/meme-stock-5206762.

[6] J. Grimes, “Reddit scraper 2022: How to scrape data from Reddit,” *Best Proxy Reviews*, 31-May-2020. [Online]. Available: https://www.bestproxyreviews.com/reddit-scraper/. [Accessed: 24-Apr-2022].

[7] L. Kramer, “An overview of Bull and Bear Markets,” *Investopedia*, 08-Feb-2022. [Online]. Available: https://www.investopedia.com/insights/digging-deeper-bull-and-bear-markets/#:~:text=Bull%20Market%20vs.,-Bear%20Market&text=Bear%20Market-,A%20bull%20market%20is%20a%20market%20that%20is%20on%20the,stocks%20are%20declining%20in%20value. [Accessed: 23-Feb-2022].

[8] M. Ettredge, J. Gerdes, and G. Karuga, “Using web-based search data to predict macroeconomic statistics,” *Communications of the ACM*, vol. 48, no. 11, pp. 87–92, 2005.

[9] O. Lam, “Detecting subjectivity and tone with Automated Text Analysis Tools,” *Medium*, 30-Jul-2018. [Online]. Available: https://medium.com/pew-research-center-decoded/detecting-subjectivity-and-tone-with-automated-text-analysis-tools-5f0e662224b8. [Accessed: 23-Feb-2022].

[10] Paul C. Tetlock, Saar-Tsechansky, and Macskassy, “More than words: Quantifying language to measure firms' fundamentals,” *The Journal of Finance*, vol. 63, no. 3, pp. 1437–1467, 2008.

[11] R. Silipo and K. Melcher, “Sentiment analysis: What's with the tone?,” *InfoQ*, 27-Nov-2018. [Online]. Available: https://www.infoq.com/articles/sentiment-analysis-whats-with-the-tone/. [Accessed: 23-Feb-2022].

[12] S. Seth, “Basics of Algorithmic Trading: Concepts and examples,” *Investopedia*, 08-Sep-2021. [Online]. Available: https://www.investopedia.com/articles/active-trading/101014/basics-algorithmic-trading-concepts-and-examples.asp. [Accessed: 24-Feb-2022].

[13] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, “Mining of Concurrent Text and Time Series.”

[14] W. S. Journal, “DJIA | dow jones industrial average historical prices - WSJ,” *The Wall Street Journal*. [Online]. Available: https://www.wsj.com/market-data/quotes/index/DJIA/historical-prices. [Accessed: 24-Apr-2022].

[15] Y. Mao, W. Wei, B. Wang, and B. Liu, "Correlating S&P 500 stocks with Twitter data," *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research - HotSocial '12*, 2012.

Itemized Budget and Materials

Item	**Explanation**	**Cost (USD)**
Salary	Based on my detailed timeline below, I believe I will need three weeks of dedicated time in order to conduct this research. At eight hours a day and $20/hr, this comes to a total of $2400.	2400
Housing	I also will require housing for the duration of the experiment. If conducted at LMU, this housing could be provided at very little or no cost to the university.	*free*
Food Stipend	Additionally a living expenses and food stipend of $20 per day adds an additional $420	420
APIs	Free APIs are available for both Reddit (Pushshift api for gathering historic Reddit data) and the stock market (such as Polygon.io, though many such APIs exist)	*free*
		Total Cost: 2820

Detailed Timeline

TASK TITLE	DURATION (Days)	WEEK 1					WEEK 2					WEEK 3				
		M	T	W	R	F	M	T	W	R	F	M	T	W	R	F
Programming																
Plan and Blueprint Program	2															
Select APIs	1															
Initial Programming	3															
Debugging	2															
Data Collection																
Run Data Collection	1															
Visualize and Analyze Data	3															
Write Report																
Write Report	4															