Simulating the portfolios of investing subreddits using sentiment analysis

Eric Kearney

December 9, 2022

1 Introduction

reddit.com is a popular online discussion board, divided into "subreddits", which are user-created and dedicated to a particular topic of discussion. This means that there are often multiple subreddits dedicated to the same topic broadly, however because they grew separately they will have often fostered their own individual communities, each with their own beliefs, rules, goals, etc.

I've identified three subreddits focussed on investing into the U.S. stock market, r/investing, r/stocks, and r/wallstreetbets (content warning: These online communities, in particular r/wallstreetbets, are known for using language that some may find offensive).

The goal of this project was to use sentiment analysis to determine how each of these subreddits viewed individual stocks, assemble a portfolio for each subreddit based on this analysis, and compare the performance of these portfolios against each other.

2 Collecting data

I started this project with this Stock Market Subreddits dataset I found on kaggle. I then wrote a script (collect_stock_data) using this API which uses Yahoo Finance to allow me to download historic Stock Market Data.

3 Data Pre-Processing

The next task for this project was the pre-process of 'clean' the reddit dataset. I began with boilerplate language processing techniques, such as using nltk to remove words that are common, yet carry little meaning (ex: "The", "I", "What", etc.). Finally, I removed all puncuation from the text.

I then scanned each post and comment in the reddit dataset for text that could be found in the list of NASDAQ symbols obtained from nasdaqtrader.com (e.g., 'AAPL' for 'Apple', 'TSLA' for 'Tesla'), once a symbol was found, I attached it as a new column in the dataframe to that post. If a single post contained multiple symbols, I copied the post and added a new entry into the dataset, with the second symbol attached in the new column, and repeated as necessary.

For example, a post titled: "Video explaining the huge growth of large cap growth stocks (AAPL, AMZN, MSFT, TSLA, FB, GOOGL)","I've seen this question everywhere on this sub: ""why did apple/tesla/[insert large growth company name here] growth so much so quick! They still haven't made a lot of profit! We are in a bubble!"* from r/investing became:

'AAPL', 'Video explaining the huge growth of large cap . . .'
'AMZN', 'Video explaining the huge growth of large cap . . .'
'MFTS', 'Video explaining the huge growth of large cap . . .'
etc.

*Unfortunately the Kaggle dataset did not include the usernames of those who posted, so I cannot properly give credit to this quote.

4 Using VADER

For the sentiment analysis portion of this project, I elected to use VADER (Valence Aware Dictionary and sEntiment Reasoner)[1]. VADER is a pre-trained sentiment analysis model trained using social media data, making it a powerful tool for this task. One potential cavaet is that the subreddit r/wallstreetbets is well-known for having a more unique lexicon and culture, which could throw a wrench in VADER's ability to determine the sentiment of statements originating from it.

I had originally hoped to be able to train my own model, which would've

naturally led to more accurate results, however due to a shortage of time and resources I decided it would be more prudent to simply use a pre-trained model instead.

[1]VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (by C.J. Hutto and Eric Gilbert) Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

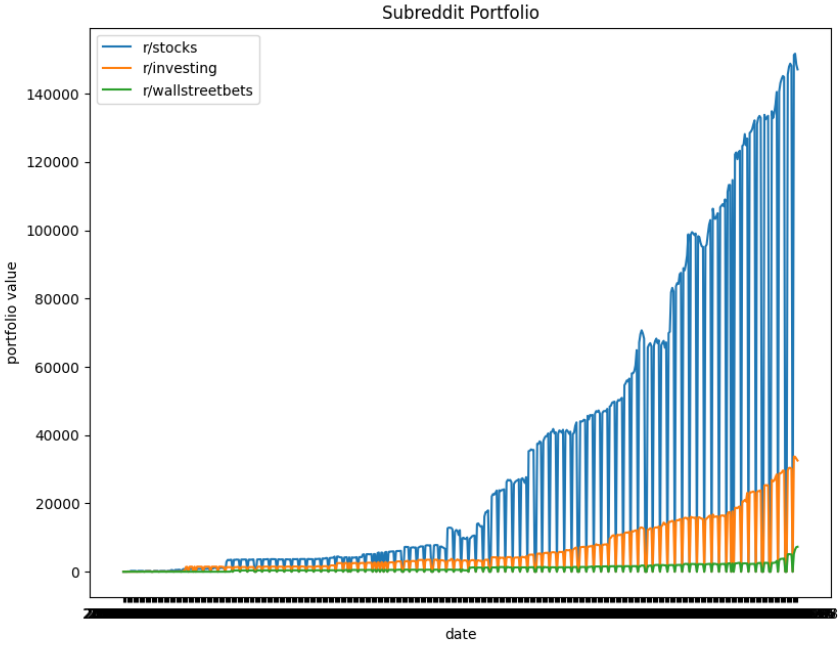
5 Simulating the Stock Market

Using the analyzed posts and comments from the three subreddits along with the historic stock market data, I created a stock market simulation, in which three traders, one for each subreddit, exists. My simulation began on April 27th, 2019 and ended January 27th, 2021. On each day between those dates, the simulation found all posts/comments from the reddit dataset that were posted on that day, then if the sentiment expressed in that post was positive, the trader representing the subreddit from which the post originated from would 'buy' one share of the stock associated with that post. If, on the other hand, the post had a negative sentiment, the trader would 'sell' one share of that stock (assuming it actually owned a share of it to begin with). There were times where there'd be multiple posts concerning the same stock from the same subreddit, and the sentiment in each of those posts diverged. To handle this situation, I created a 'score' which was caluclated by using the number of 'upvotes' each post had received from other members in the community. I multiplied the upvotes of posts with a negative sentiment by -1, and then summed the score for all the posts concerning the same stock together. If that score ended up being ¿ 0, the trader 'bought' the stock, otherwise, it 'sold' the stock.

I had originaly hoped to simulate around five years of the stock market, however I realized that r/wallstreetbets hasn't been active that long, so instead I had to scale back to simulating 21 months.

6 Results

The following is the value of all the stocks in the portfolio of each subreddit, minus the amount that it would've taken to buy those stocks, on the last day of the simulation.


Subreddit Portfolio
r/stocks
r/investing
r/wallstreetbets
140000
120000
100000
80000
60000
40000
20000
0
portfolio value
date

- r/stocks : $147,138.64
- r/investing : $32,556.34
- r/wallstreetbets : $7,302.13

While it appears that investing was the clear winner, there's much more to this story (See Errors & Future Work)

7 Errors & Future Work

While developing this project, it became clear that there are many errors currently embedded within it. Resolving these errors will require a great deal of time and effort, and I do have to turn in **something**. I hope to continue work on this as a personal project after class ends. Here are the errors I am currently aware of:

- The pre-processing currently only looks for NASDAQ symbols within the text, not for company names. For example, a post that contains 'AAPL' will be correctly flagged as discussing Apple, Inc., however a post containing 'Apple' will not. This means that not all the companies being discussed are being correctly 'bought' in the simulation, and that companies with names that are commonly abbreviated to their NASDAQ symbols, such as 'GME', will be overrepresented in the simulation compared to companies that are often fully spelled out when discussed online.
- Some companies have NASDAQ symbols that happen to be equivalent to common words, most apparently is Agilent Technologies Inc., which has the NASDAQ symbol 'A'. This means that stocks were being 'bought' when they shouldn't have been. The subreddit hurt most by this error was clearly r/wallstreet bets. Whose portfolio was comprised almost entirely of 'YOLO', which is the AdvisorShares Pure Cannabis ETF. It is clear from the context of the comments that the intention behind saying 'YOLO' was almost always as the acronym "You Only Live Once". An example comment containing "YOLO" from r/wallstreetbets:

  "Keynes said we're all dead in the long run, I think he meant always YOLO"*

- Another issue that particularly hurt r/wallstreetbets were sentiment mis-identifications. For example the post:

  "CLASS ACTION AGAINST ROBINHOOD. Allowing people to only sell is the definition of market manipulation. A class action must be started, Robinhood has made plenty of money off selling info about our trades to the hedge funds to be able to pay out a little for causing people to loose money now","LEAVE ROBINHOOD. They dont deserve to make money off us after the millions they caused in losses. It might take a couple of days, but send Robinhood to the ground and GME to the moon."*

  Scored a negative sentiment for GME. While this appears to be correct as in the sentiment of the statement itself is negative, that negativity is not intended to be directed towards GME. As a matter of fact, a great deal of the discussion on r/wallstreetbets was targeted towards this Robinhood class action lawsuit, and similar lawsuits. The end result of this was that r/wallstreetbets bought almost no shares of GME, which is the community's stock of choice.

- Subreddits were allowed to buy or sell no more than one share of each company a day. A better simulation probably would allow for the purchase of numerous shares of a stock if the subreddit's sentiment towards that company was extremely positive, and conversely, allow for the sale of numerous shares if the sentiment were extremely negative.

- Clearly, more data from each subreddit, as well as model trained on the text of the subreddits themselves would've improved accuracy.

*I unfortunately cannot quote the users as their usernames were not included in the dataset.

8 Limitations

It should be made explicity clear to anyone who wishes to work on a project like this, make a similar project, or base their financial decisions based on a project like this that it is incredbily easy to manipulate the results. Almost any portfolio can be made to look incredible if one is willing to cherry pick the dates. I personally would under no circumstances recommend that anyone invest money based on the advice of strangers on the Internet, without

thoroughly doing their own research. It may be tempting to start following r/stocks now and just invest in whatever the users there seem to be buying, but it should always be stated again and again that past performance does not indicative of future results.