Bryan Zhao, Darren Nguyen, Zayaan Atif
ECON-UB 232: Data Bootcamp
Midterm
10 October 2025

Purpose

Stock analysis has become an increasingly popular means of generating wealth and creating economic growth. Still, many investors struggle to make money as investment strategies are difficult to master and understand. Through our project, we developed a model to analyze a company's profitability, valuation, and reputation through web scraping to recommend potential investment options.

Essential Questions:

- How can we combine current sentiment, price momentum, and underlying valuation metrics to create a relative composite score for each stock?
- On average, what sector performs best under our scoring method?
- Based on these metrics, which stocks from the S&P 500 should be purchased?
- Similarly, what stocks should be shorted?
- How do price moving averages with different windows compare relative to each other?
- What similarities exist between the graphs of two different stocks in the same sector? How about the graphs of two different stocks' moving averages?

Data Sources

YFinance

An important aspect when considering an investment is a company's strength. In order to assess strength, we look at a company's profitability through its return on equity and return on assets ratios.

$$ROE = \frac{\text{Net Income}}{\text{Shareholders' Equity}}$$

$$ROA = \frac{\text{Net Income}}{\text{Total Assets}}$$

Each ratio tells us the ability of a company to use its equity and assets to generate profits, a strong indicator of the strength, stability, and potential growth of the company.

Another important aspect to consider is the momentum of stocks. We utilized a company's one-month and three-month moving averages to determine if a stock is trading above its long-term moving average and identify the current momentum.

Reddit

Finally, evaluating a company's reputation is important, as favorably viewed companies are more likely to find new investors to drive growth within the company. We scraped Reddit, more specifically r/wallstreetbets, r/stockmarket, and r/stocks, as they are some of the largest investment discussion platforms with the most active communities. We scraped the top 500 posts of the last month of each subreddit for mentions of tickers in the S&P 500, as these are likely to be the most discussed stocks. We then collected the titles and bodies of the posts and ran sentiment analysis on them using VADER.

Methodology

Pulling YFinance Data

The YFinance API is a powerful webscraper, allowing us to efficiently retrieve public information on a stock's financials, statistics, and price history. Using the.Ticker() method, we created an object that allows us to scrape all publicly available information on the Yahoo Finance database. With the object, we could use the .info method to create a dictionary, which could extract a company's name, real-time price, sector, market capitalization, ROE, and ROA.

Additionally, we could use the .history() method with parameter period = 1y to create a DataFrame with information on historical prices. To more accurately calculate moving averages, we extrapolated the ["Close"] price column to create a series, allowing us to use a rolling method to separate data by one-month and three-month time blocks. Using the time blocks, we used the .mean() method to create a one-month and three-month moving average over a period of time, allowing us to graph past moving average trends. Finally, we used the iloc[] method to extract the most recent moving averages to calculate our momentum.

Reddit Sentiment Analysis

The sentiment analysis section used VADER (Valence Aware Dictionary and Sentiment Reasoner) as it was designed to score the sentiment of social media posts, being aware of things like emojis and slang. SentimentIntensityAnalyzer() was initialized to compute the sentiment score, which can take a range of -1 (extremely negative) to +1 (extremely positive)

Functions were created to streamline the process of finding the sentiment of text, such as the titles or bodies of Reddit posts. The get_sentiment() function returns the VADER sentiment score of a given text, and 0 for a null value or non-string. The overall_sentiment function averages the sentiment of the title and body if both are non-zero values, allowing every post to have a single overall sentiment that represents its general tone.

Using Praw (Python Reddit API Wrapper), we scraped the top 500 posts of the last month on the aforementioned subreddits. We extracted our attributes of interest, such as the text of the title and the body, and stored them in a Pandas DataFrame. We then applied the VADER functions that were created earlier to find the sentiment of each title and body, and then created a column for the overall sentiment. Next, we iterated through the combined title and body to cross-reference them against a predefined list of S&P 500 tickers and summed the sentiment for that ticker in the subreddit. For every post, if the ticker

was present anywhere in the title or body, we added 1 to the mention count. We also added the sentiment of the text in that post to the sum of sentiments for each ticker in that subreddit. To maintain the accuracy of our data, we made a list of common words and phrases that could be misconstrued as tickers (and were likely not being mentioned), and cleaned the dataframe by removing these "false tickers."

After each subreddit's posts were processed, they were appended to a larger list and concatenated into a single combined DataFrame when the loop ended. Total mentions were calculated, and the average sentiment was found by dividing the sum of the sentiments by total mentions.

A DataFrame was created with columns of ticker, mentions, and average sentiment, with tickers being the index. It was ordered first by descending mentions and then by descending average sentiment.

Calculating Composite Scores

Our next goal was to calculate relative composite scores to compare each stock on the S&P 500. Since we had the data frames, stock_df and sentiment_df, we joined them on the ticker index, where sentiment for stocks in stock_df that did not have an entry in sentiment_df was populated with NaN values.

We wanted to calculate a composite score based on five metrics: return on assets, return on equity, mo, amount of mentions, and average sentiment. To combine them, we calculated the z-scores for each metric and weighted them according to their importance in our model. We focused our stock insights on short-term movements due to the near-sighted nature of sentiment and our moving average crossover strategy. Our final equation, with each metric in terms of z-scores, ended up as

Composite Score = (Sentiment Weight x Average Sentiment x Mentions) + (Momentum Weight * MA Momentum) + (ROE Weight * ROE) + (ROA Weight * ROA).

We then estimated weights based on short-term impact.

ROE (12.5%) and ROA (12.5%): As return on equity and return on assets are closely related and measure long-term performance, we aimed to capture the underlying value of stocks. This led us to allocate a total of 25% of our scoring on these valuation metrics, or 12.5% each.

Sentiment (12.5%): For sentiment, we noted large possibilities for bias in a subreddit, but we also saw the importance of current market chatter. We decided to maintain a minimal but impactful allocation of 12.5%.

Note: Many stocks may not have a sentiment score or mentions. This does not negatively impact the integrity of our scoring, as the absence of attention is also an indicator of a stock's movement.

Momentum (50%): Because they can indicate a stock's short-term trends, we heavily favored momentum in our scoring. So, we weighted it at 50%.

Findings

What stocks should be purchased? Through our composite score, we identified 3 strong stocks that reflect a combination of good sentiment, high momentum, and strong fundamentals: NVIDIA (NVDA), UnitedHealth Group (UNH), and Advanced Micro Devices (AMD).

Ticker	Name	Composite Score	Sector	Market Cap	ROE	ROA	Current Short MA	Current Long MA	MA Momentum	Mentions	Average Sentiment	Prices	Short Term MA	Long Term MA	ROE (Z-Score)	ROA (Z-Score)	MA Momentum (Z-Score)	Mentions (Z-Score)	Average Sentiment (Z-Score)
NVDA	NVIDIA Corporation	5.996176	Technology	4.424709e+12	109.417	53.094	181.588001	172.984214	4.973741	10.0	0.293195	Date 2024-10-22 00:00:00-04:00 143.548843 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	0.079557	5.632000	0.450399	5.466480	3.700394
UNH	UnitedHealth Group Incorporated	3.616476	Healthcare	3.309059e+11	21.653	6.464	352.254301	311.202205	13.191454	6.0	0.276375	Date 2024-10-22 00:00:00-04:00 557.763550 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	-0.052165	0.060006	1.606471	3.232006	3.480514
AMD	Advanced Micro Devices, Inc.	3.599385	Technology	3.862855e+11	4.699	2.190	187.054666	166.777222	12.158401	34.0	0.057710	Date 2024-10-22 00:00:00-04:00 154.089996 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	-0.077611	-0.450711	1.461141	18.873320	0.622011

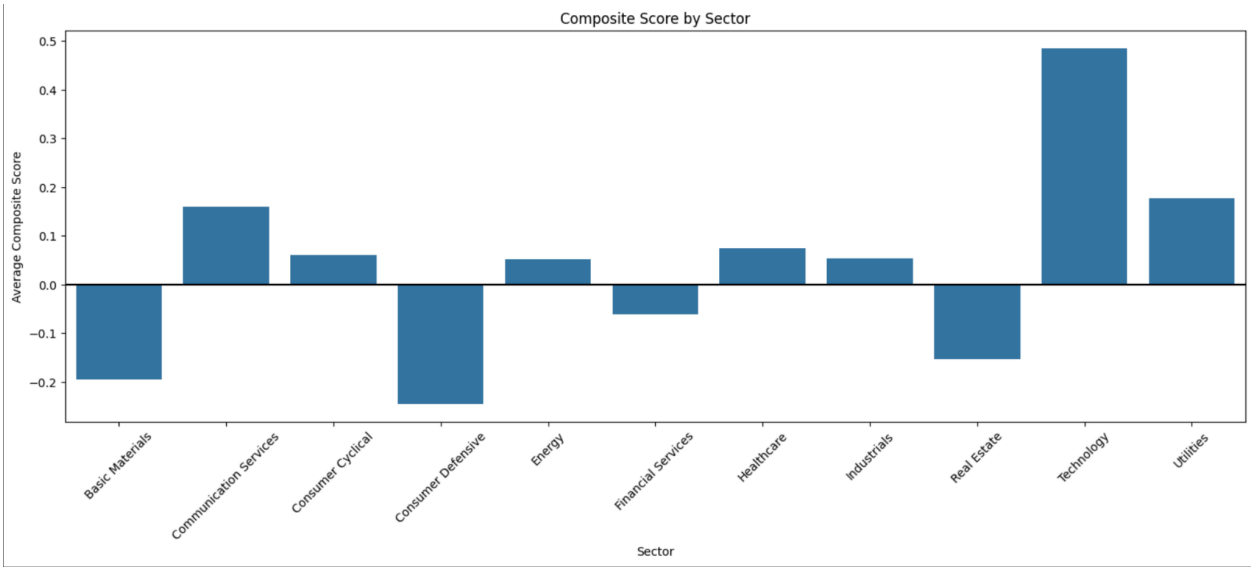
It is interesting to note that the top 3 stocks had strong sentiment scores or mentions, each at around a z-score of 3, showing a bullish market attitude. NVDA also had an ROA z-score of 5, signaling its high efficiency and profitability. Moreover, AMD had an astronomically high z-score of 18, showing how much talk is surrounding the company compared to the rest of the market.

What stocks should be shorted? We identified the three weakest stocks: Align Technologies (ALGN), Goodyear Tire & Rubber (GT), and Xerox Holdings (XRX).

Ticker	Name	Composite Score	Sector	Market Cap	ROE	ROA	Current Short MA	Current Long MA	MA Momentum	Mentions	Average Sentiment	Prices	Short Term MA	Long Term MA	ROE (Z-Score)	ROA (Z-Score)	MA Momentum (Z-Score)	Mentions (Z-Score)	Average Sentiment (Z-Score)
ALGN	Align Technology, Inc.	-1.266675	Healthcare	9.874093e+09	11.410	6.500	130.267332	155.610445	-16.286254	0.0	0.0	Date 2024-10-22 00:00:00-04:00 210.809998 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	Date 2024-10-22 00:00:00-04:00 NaN 2...	-0.067538	0.064307	-2.540467	-0.119704	-0.132409
GT	The Goodyear Tire & Rubber Comp	-1.316733	Consumer Cyclical	2.045232e+09	8.986	1.510	7.694667	9.152889	-15.931824	0.0	0.0	Date 2024-10-22 00:00:00-04:00 8.19 2024-10...	Date 2024-10-22 00:00:00-04:00 NaN 202...	Date 2024-10-22 00:00:00-04:00 NaN 202...	-0.071176	-0.531967	-2.490606	-0.119704	-0.132409
XRX	Xerox Holdings Corporation	-1.437577	Technology	4.039124e+08	-72.183	0.279	3.595560	4.340992	-17.171914	0.0	0.0	Date 2024-10-22 00:00:00-04:00 9.759312 202...	Date 2024-10-22 00:00:00-04:00 NaN 202...	Date 2024-10-22 00:00:00-04:00 NaN 202...	-0.193000	-0.679063	-2.665062	-0.119704	-0.132409

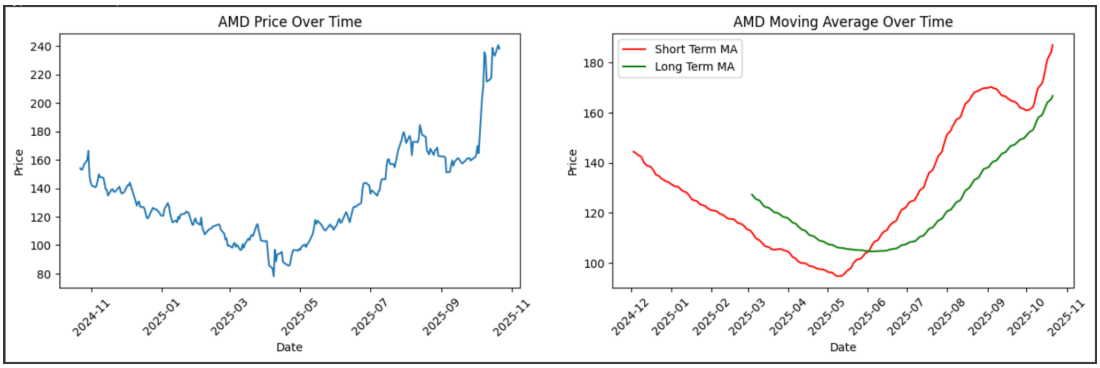
The bottom 3 stocks had a downward momentum, around a z-score of -2.5, that pulled down their overall rankings.

On average, what sector performs best under our scoring method? Using the average composite score of each sector, we developed the following chart.



There is a clear winner in the technology sector stocks, and the lowest ranking sector was consumer defensive, with basic materials and real estate trailing close behind. However, it is important to note that these sectors may get less coverage than technology stocks, and they could be overshadowed due to a lack of sentiment.

How do price moving averages with different windows compare relative to each other? By graphing both the price and moving averages side by side for a given stock, we can visualize how moving averages of different window sizes smooth out the price in different ways. More importantly, there is a clear difference between the long-term moving average and the short-term moving average, as shown below.



This difference captures a critical part of our composite scoring, where short-term averages rising above the long-term averages indicate upward momentum and vice versa.

What similarities exist between the graphs of two different stocks in the same sector? How about the graphs of two different stocks' moving averages? When comparing two stocks in the same sector, we can notice how prices change in tandem. For example, AMD and NVDA prices rise and dip at similar times, so their moving averages follow the same trend lines. This is an important insight when selecting the highest-ranking stocks according to our criteria, as the momentum and sentiment are likely to be similar across an industry.