**MSDS 451 - Check In 3**

Ethan Vanlerberghe

School of Professional Studies, Northwestern University

MSDS 451: Financial Engineering

Dr. Miller

17 August 2025


NORTHWESTERN
UNIVERSITY

**Introduction (10 points). Why are you conducting this research? Identify potential users of the knowledge base and application(s) that you intend to develop.**

The purpose of conducting this research is to create a buy-and-hold model focused on sentiment analysis gathered via social media to algorithmically trade assets. To do so, thorough research must be performed to ensure proper decision-making can be made every step of the way, leading to the best possible outcome for any future use. This research, however, can extend beyond just the creation of this ETF, and instead help others who are looking to set up their "Own ETF's", as it provides a strong basis of understanding for those beginning to understand the purpose of ETFs and how to create one successfully. This research should be used by individuals looking to create their "Own ETF's", to establish and understand how to select items to build their ETF successfully. Specifically, it should be used to start learning how to build a basic model on 25 years of large-cap fund data. The end application is designed to be used by others as a teaching tool to better see the possibilities of creating their "Own ETFs" and allow users to track the performance and decision-making of the model.

**Literature review (10 points).. Who else has conducted research like this?**

A great deal of research has already been done in this area, and serves as a massive knowledge base on how to begin this process. Key examples lie in GitHub projects (Hansen-han, 2025), or (Shaidev, 2025). In these examples, assets are pulled from Yahoo Finance using YFinance and manipulated using Monte Carlo simulations to predict uncertainty in the markets. Then, based on the assets chosen, a Twitter query is performed to gather sentiment regarding said assets. Based on the sentiment, a weight is calculated to aid in asset valuation prediction. Without this uncertainty modeling, backdating data to describe performance would not allow for an accurate forecast into the future.

**Methods (10 points). How are you conducting the research? Make sure you address the issues that are the focus of this checkpoint assignment.**

The four large-cap 25-plus-year companies that were selected were Apple ($AAPL), Microsoft ($MSFT), IBM ($IBM), and Caterpillar ($CAT). These four were chosen due to their long-term financial stability since at least 1999. The data for these four companies was pulled from Yahoo Finance via the YFinance package. To benchmark the performance of this portfolio, a benchmark of GSPC was used, modeling the portfolio against the S&P 500 from 1999 to the year-end 2024. In the first code chunk, the management fees are also established. A management fee of 2% annually was inserted, as it is in the average range of 1% - 2% mutual charge (ICI, 2024). The 20% performance fee is based on the fund's ability to outperform the benchmark, in this case, the S&P 500, well within the average range for a management fee (Investopedia, 2025).

After establishing variables, the Twitter data scraping began. First, the number of lookback days, limit per ticker, rate limit, and max retries were created (Hansen-han, 2025). The number of lookback days tells the model how far back in time to look for tweets regarding the topic. The tweet limit per ticker sets a limit on the number of tweets to observe (Hansen-han, 2025). The rate limit is due to the newer limitations on X's API platform, no longer allowing for academic licenses (Hansen-han, 2025). The free license only allows for 100 observations per 15 minutes. This limit helps to keep this number under the limit. Finally, the number of retries restricts the number of times the model tries to run before it moves on. After these values are established, the bearer token from the API account is inserted (Shaidev, 2025). Then, the Vader lexicon package is loaded to help observe the sentiment of the tweets (Hansen-han, 2025). As the Vader program processes the tweets, and outputs a modifier weight for use later in calculating the forecasted value of the asset (Shaidev, 2025).

After pulling the tweet sentiment, it was time to begin creating functions. The first function, compute_log_returns, returns the log price return for the daily returns of the portfolio. Log units are used to more easily compare daily returns and movements across markets and securities within those markets

(Miller, 2025). After these values are gathered and calculated, it is time to move on to the benchmark returns. The benchmark returns are then gathered and logged.

Next, a Monte Carlo simulation is created to model uncertainty in the financial markets. The model estimates the mean vector and covariance matrices of historical returns for the assets and simulates thousands of return paths over one year (Raghav, 2025). By doing this, the uncertainty of future returns can be estimated while also using backdated data. Random portfolio combinations are created using random weights applied to the different assets over the lifetime of the portfolio. These weights can simulate market shifts, economic swings, or even major scandals that companies may endure (Quimbayo, 2025). For this rendition, some very basic mean reversion was added to the modeling strategy to try and better overall model performance. Each portfolio run's expected return, volatility, and Sharpe ratio are computed based on these simulated outcomes (Shapo, 2025). The portfolio run with the highest Sharpe ratio is selected and used for backdating. This portfolio run, nicknamed 'Run 42' (Adams, 1979), backtests the portfolio over the 25 years to provide a basic estimate of returns over time. This estimate will allow any potential investors to gain more confidence in the decisions being made by the management.

Lastly, KPI's are created to measure the portfolio's performance. These include alpha, beta, and Sharpe ratios. These all serve the same purpose: To build confidence in any potential investment (Wilmott, 2008).

**Results (10 points).. What did you learn from your research so far?**

This week's research taught me quite a bit about how to go about building a portfolio, combined with the simulation of the unknown. While week 2's assignment taught me how to go about creating my trading portfolio and measuring it against statistics, it had very little uncertainty built in. Week 5's assignment was all about uncertainty modeling with Monte Carlo simulations, but very little about active portfolio management. This week was a large-scale combination of both. Bringing the two ideas together

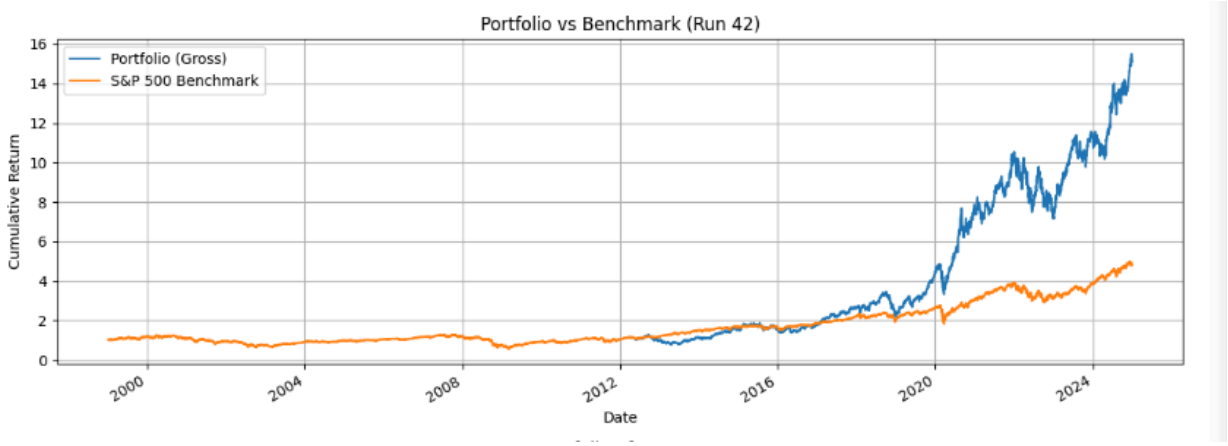
required much more time and research than I had anticipated; however, the result is that much more rewarding because of it.

Overall, while the model appears to be MUCH outperforming the benchmark, in terms of performance statistics, not so much. With an Alpha of 0.0003, a beta of 1.1587, and a Sharpe ratio of -0.548, I can see that we barely beat the market, the portfolio is 15.87% more risky than the market, and when adjusted for risk, the investment is not worth it according to the Sharpe ratio. This means that when backtested, this portfolio does not show investors that it is worth investing in relative to the S&P 500, which is a problem. I look forward to spending more time trying to understand why that is. I would assume it is because of the risk taken when picking only a small handful of stocks, but I will need more research to know.

**Conclusions (10 points).. So, what does it all mean? Do you have any concerns about the term project at this point?**

After this check-in, I want to spend more time looking at how I chose to do my sentiment analysis. I thought that the X.com API's would be my best bet, as Twitter-based trading bots have always been something that I found extremely useful and interesting. However, with the data caps without the academic licensing, I feel that it may no longer be the best tool. Instead, I would like to explore using Reddit data, hopefully they will have a more friendly API fan, or at least an academic package for students. As for the performance of the model, it is slightly better, but I believe the main issue is due to the performance of the model based on the data it was provided. As they say, "Garbage in, garbage out!".

Figure 1: Portfolio Performance



**Works Cited**


Abraham, Stephan A. “How to Create Your Own ETF.” *Investopedia*, 1 Apr. 2015, investopedia.com/articles/investing/040115/how-create-your-very-own-etf.asp. Accessed 20 July 2025.

Adams, Douglas. *The Hitchhiker's Guide to the Galaxy*. Pan Books, 1979.

Dierking, Dave. “ETFs 101: My Basic Criteria for Choosing an ETF.” *Seeking Alpha*, 17 Dec. 2018, seekingalpha.com/instablog/5366071-dave-dierking-cfa/5210721-etfs-101-basic-criteria-for-choosing-etf.

Grinold, Richard C., and Ronald N. Kahn. *Active Portfolio Management: Quantitative Theory and Applications*. Probus, 1995.

hansen-han. “alpaca_sentiment_trader.” *GitHub*, 2025, github.com/hansen-han/alpaca_sentiment_trader. Accessed 17 Aug. 2025.

“Holdings & Sector Allocations of Invesco QQQ.” *Invesco*, invesco.com/qqq-etf/en/about.html. Accessed 20 July 2025.

Miller, Tom. “451 Feature Engineering: Programming Assignment 1.” 2025.

—. “Week 6 Term Project Checkpoint B.” Lecture, 2025.

“Modern Portfolio Theory.” *Wikipedia*, Wikimedia Foundation, 29 July 2025, en.wikipedia.org/wiki/Modern_portfolio_theory. Accessed 30 July 2025.

Quimbayo, C. A. Z. “Robust Bayesian Portfolio Optimization.” *Journal of Operations Management*, 2025, doi:10.1016/j.jom.2025.05.004.

Raghav. “MonteCarloSimStockPrices.” *GitHub*, 2025, github.com/RaghavsScarletSplendour/MonteCarloSimStockPrices. Accessed 30 July 2025.

Ross, Sheldon M. *An Elementary Introduction to Mathematical Finance*. Cram101, 2016.

Shapo, Misha. "Simulating Stock Prices with Monte Carlo Methods." *GitHub*, 2025, github.com/MishaShapo/Monte_Carlo_Stocks/blob/master/Simulating_Stock_Prices_with_Monte_Carlo_Methods.ipynb. Accessed 30 July 2025.

"Simulating Stock Prices with Monte Carlo Methods." *YouTube*, uploaded by Misha Shapo, 2025, youtube.com/watch?v=6-dhdMDiYWQ. Accessed 30 July 2025.

"Stock Trading Principles." *YouTube*, uploaded by Misha Shapo, 2025, youtube.com/playlist?list=PLvcbYUQ5t0UHdm6bNx3Rnj1fdpt9sGNsm. Accessed 30 July 2025.

shirosaidev. "stocksight." *GitHub*, 2025, github.com/shirosaidev/stocksight . Accessed 17 Aug. 2025.

Thompson, Cedric. "Fundamental vs. Technical Analysis: What's the Difference?" *Investopedia*, investopedia.com/ask/answers/difference-between-fundamental-and-technical-analysis/. Accessed 20 July 2025.

VanEck. "$BUZZ Prospectus." *SEC.gov*, sec.gov/Archives/edgar/data/1137360/000113736025000087/vanecksocialsentimentetfbu.htm. Accessed 20 July 2025.

—. "Buzz– VanEck Social Sentiment ETF: Holdings & Performance." *VanEck*, 2 June 2023, vaneck.com/us/en/investments/social-sentiment-etf-buzz/.

Wilmott, Paul. *Frequently Asked Questions in Quantitative Finance*. Playaway Digital Audio, 2008.