Norwegian School of Economics

Bergen, Fall 2022

Eat the Rich

The Aftermath of the Meme Stock Craze: An Empirical Analysis

Martin Nakstad and Jonas F. Hasle

Supervisor: Nataliya Gerasimova

Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

Abstract

Recently, the r/wallstreetbets subreddit has sent shock waves through the financial industry. The subreddit has been the subject of widespread discussion regarding the unintuitive price actions caused by the platform activity. This thesis examines recent changes in the WallStreetBets (WSB) phenomenon, particularly after the meme stock craze in 2021. In addition, it examines relationships between WSB activity and stock data and whether it is possible to utilize findings in a trading algorithm producing abnormal returns. The results identify the meme stock craze as an outlier period, where in the following period the impact of WSB activity returns to normal levels corresponding to the pre-2021 period. In addition, the results demonstrate a significant relationship between WSB activity and stock data, particularly for the mentions activity on the subreddit. Moreover, this thesis finds a profitable strategy adjusted for biases by maximizing the Sharpe ratio, which also provides a strong abnormal return. In a broader sense, its results are promising in relation to the utilization of big data and quantitative methods in analyzing social media and applying the findings to a financial strategy.

Keywords: WSB, r/wallstreetbets, quantitative finance, factor strategy, Fama-French

Acknowledgements

First, we would like to thank our supervisor, associate professor Nataliya Gerasimova, for her commitment to this thesis. She has shown great motivation for our topic and provided us with feedback and valuable insight, which have been of great value this year. Unfortunately, Nataliya is leaving the Norwegian School of Economics, and we wish her all the best in the future. This thesis marks the end of our master's degree in Finance, and we would like to extend our sincere gratitude to friends and family for their support during the five years in Bergen.

Martin Nakstad *Jonas F. Hasle*

Contents

List of Figures

List of Tables

1 Introduction

Since the GameStop rally in 2021, the online society r/wallstreetbets, also known as WallStreetBets (WSB), has grown to become the 25th largest subreddit (r/wallstreetbets, 2022). The activity and attention directed towards the platform have become a popular subject among investors and reporters as the community has received recognition for drastically increasing stock prices, volume and initiating short squeezes. This section presents our motivation for selecting WSB as a topic and introduces the research questions that guide this thesis. Finally, an overview of the WSB community follows.

1.1 Our Motivation

In recent months, there has been a drastic change in the worldwide economy. Inflation rates in most countries are far above target, yields are rising, and the Western world is facing an energy crisis. The current economic conditions are creating considerable fluctuations in the stock markets, and most major stock exchanges experienced bear markets in 2022. As most countries are in economic decline, WSB still seems to discuss stocks experiencing remarkable gains with the goal of "sticking it to the man"[1].

In January 2022, the *Wall Street Journal* (WSJ) published a podcast about the changes on WSB. In this podcast, Luke Vargas and Caitlin McCabe discuss how the tremendous changes have led users to leave the online community (Vargas & McCabe, 2022). Today there is much less activity on the platform than in the golden era of 2021. Even though there has been a considerable change, some things remain the same. In the middle of July 2022, the technology firm AMTD-digital (ticker HKD) performed an initial public offering (IPO) on the New York Stock Exchange (NYSE) at $7.8 per share (IPOScoop, 2022). Over the following 14 trading days, the stock price increased to $2,555, a return of approximately 22,000%, before plunging to $43.75 as at 10-05-2022 (Yahoo! Finance, 2022a). In the period after the IPO, the HDK ticker was the most frequently mentioned stock on WSB, as Figure 22 in Appendix A.3 indicates, and the community received some credit for the price increase (Grape, 2022). The HDK stock is not the first that has seen a drastic increase in price after being mentioned at WallStreetBets, and if history repeats

[1] "Sticking it to the man" is a metaphor for inflicting financial pain on professional investors connected to Wall Street.

itself, it will not be the last.

In previous years, there have been several cases where activity on social media has caused fluctuations in the stock market (Thompson, 2013), demonstrating social media's potential influence. Recently, activity by Elon Musk regarding the acquisition of the social platform Twitter (ticker TWTR) has moved the stock price in both directions. For example, when Musk tweeted: "This deal cannot move forward until he does.", referring to the percentage of spam accounts on Twitter, the stock closed down 8% (Espiner, 2022). As social media continues to grow, and since such activity could lead to fluctuations in the stock market, it is naturally fascinating to analyze one of the most popular platforms that discuss stocks (Gendron et al., 2022).

1.2 Research Questions

The new Netflix-produced series *Eat the Rich*, a documentary that investigates the WSB events in 2021, has resulted in the online community receiving more attention worldwide. With today's computing capabilities, advanced software, and big data, it is possible to undertake an in-depth analysis of the WSB phenomenon. Previous research has provided evidence of profitable strategies (Bradley et al., 2021) that got lost during changes on WSB (Bradley et al., 2021; Vargas & McCabe, 2022). As the community still identifies rapidly increasing stocks, the research questions are as follows:

I) How has the effect of WallStreetBets' activity on stock data changed recently?

II) Is it still possible to identify a profitable trading strategy for stocks mentioned on WallStreetBets?

The first question requires an in-depth analysis of how changes in the WSB community have impacted stock data. Then, after identifying significant variables, a factor strategy maximizing the Sharpe ratio is adopted, based on the activity of the WSB community.

1.3 r/wallstreetbets

In mid-June 2005, Steven Huffman, Aron Swartz, and Alexis Ohanian founded the social platform Reddit (Marsh, 2021). Since the introduction of the online community, Reddit has grown drastically. As of November 2021, Reddit was the 11th most visited webpage in the world, with approximately 1.65 billion monthly users (Clement, 2021, 2022). Subreddits are unique communities within Reddit, where specific topics can be discussed and shared between users. Redditors[2] can rank the posts with an upvote (+1) or a downvote (-1), where the post with the highest score would achieve the best position on the subreddit. As of November 2021, there were more than 2.8 million different subreddits (Wise, 2022), and WSB was the 25th biggest subreddit on the online forum (r/wallstreetbets, 2022).

r/wallstreetbets was established 31. January 2013 and is an anonymous forum for retail investors to discuss investment opportunities (r/wallstreetbets, 2022). Jaime Rogozinski, the founder of the subreddit, wanted a place to review investment opportunities with a high risk/reward ratio (Otani, 2021). Since 2013, the community has seen a considerable increase in subscribers and has expanded to other online platforms (r/wallstreetbets, 2022). Figure 1 illustrates the growth of accounts subscribing to WSB from 2013 to August 2022. As of November 2022, there are more than 13 million subscribers and contributors to the subreddit (subredditstats, 2022). The massive surge in subscribers in 2021 relates to the famous meme stock craze (Vargas & McCabe, 2022).

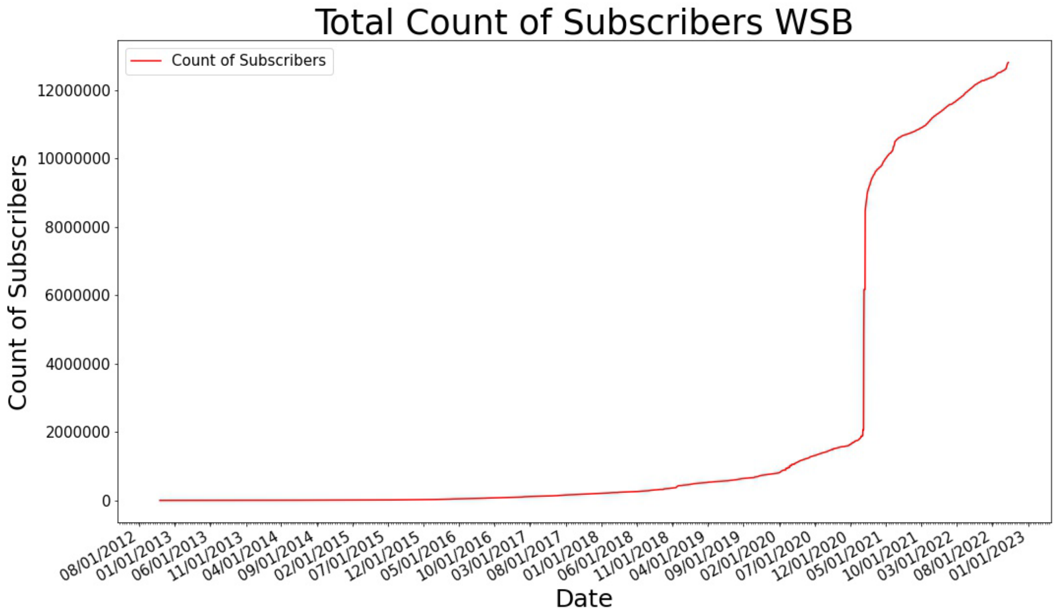

Figure 1: Number of Subscribers on WSB from 2013

[2] "Redditors" refers to registered users on the Reddit platform.

During the meme stock craze, the most famous occurrence relates to GameStop (ticker GME) and is known as the GME rally. The rally started when the WSB society discovered that investment banks and hedge funds had heavily shorted the brick-and-mortar stock. As the short interest of the floating stock to GME reached as high as 150 % (Bloomberg, 2022), retail investors started to invest in the stock. The sudden surge in volume increased the price, which caused investment banks and hedge funds to buy back their short positions. The covering of these positions is known as the GME short squeeze and drove the price up significantly, resulting in an impressive return of approximately 5000% in January 2021 (Yahoo! Finance, 2022d). Figure 2 illustrates the GME short squeeze.

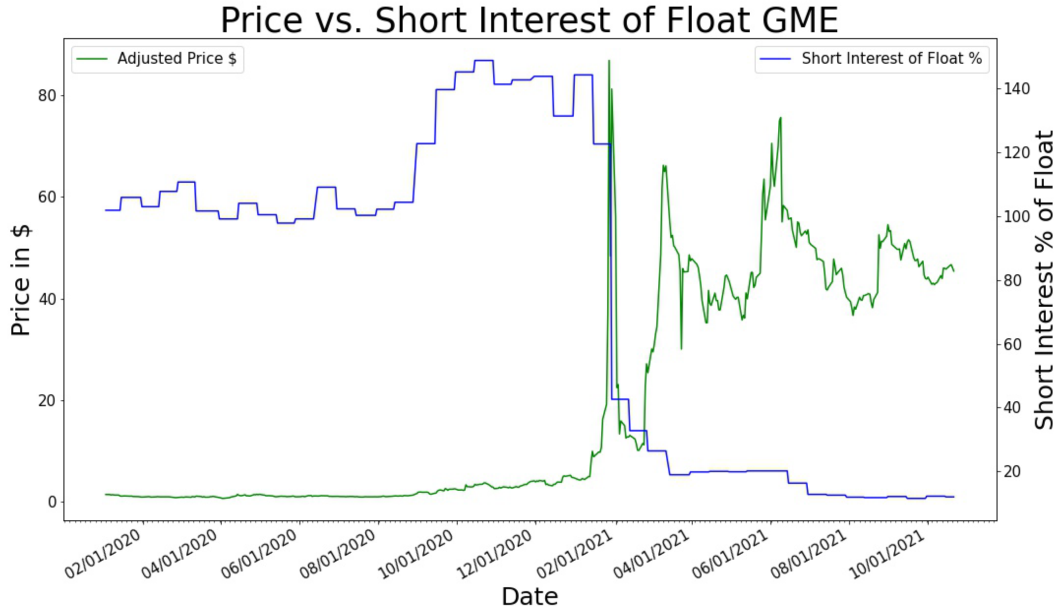

Figure 2: GME Stock Price and Short Interest Percentage of Float

During the meme stock craze, there was a significant increase in subscribers and comments on WSB, as Figure 1 and 4 illustrate. At the start of the rally, there were over 300,000 comments daily. Since then, the activity has slowly decreased and stabilized, averaging 20,000 daily comments (subredditstats, 2022).

As the growth of WSB continued, the platform moderators had to implement different rules regarding the recommendation of stocks. The moderators on the WSB platform introduced a bot that reads the ticker name of the mentioned stock and deletes the post if the company's market cap is below $500 million (r/wallstreetbets, 2022). This rule was implemented to reduce the probability of users manipulating the market. Hence, there are small variations in the stocks mentioned on WSB. Usually, the same stocks are mentioned each year, with some newcomers. Table A.2.1 in the Appendix presents an overview of the top 30 most frequently mentioned stocks on WSB for various sample periods between 2020 and 2022.

Furthermore, the WSB society has become famous for producing meme stocks. The definition of a meme stock is "a stock that has seen an increase in volume not because of how well the company performs, but rather because of hype on social media" (Gobler, 2021). For example, a recent meme stock on WSB has been Bed Bath & Beyond (ticker BBBY). The forum uses a popular song by Britney Spears when discussing this stock and uses the slogan "Hit me $BBBY one more time" (u/bighomie69, 2022). Lastly, the lingo on the platform is unique compared to other social forums that discuss stocks. The subscribers love to use emojis, especially the spaceship emoji. Moreover, they refer to people in a long position as "bagholders" and use the acronym "retard" for traders when recommending buying or selling a stock. Figure 3 below presents the words most frequently used by the WSB community.

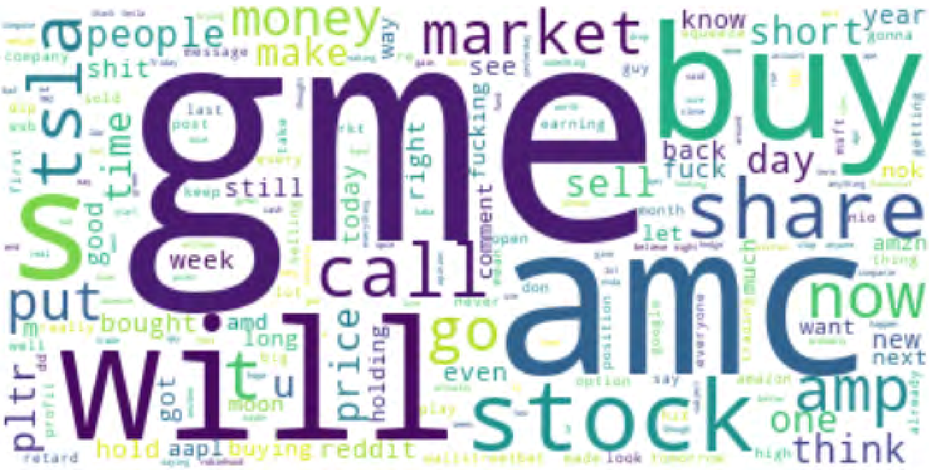

Figure 3: Most Frequently Used Words on WSB

Figures 1 and 4 present relevant data regarding the discussion on the WSJ podcast (Vargas & McCabe, 2022). These figures illustrate that the activity on WSB was much higher during the meme stock craze. Even though the activity is lower than it used to be, the social forum continues to discuss stocks experiencing extraordinary gains. For example, in late July 2022, there was increased activity in BBBY and HDK on WSB. Both stocks increased by approximately 250% and 22,000% at their highest, respectively (Yahoo! Finance, 2022b, 2022a).

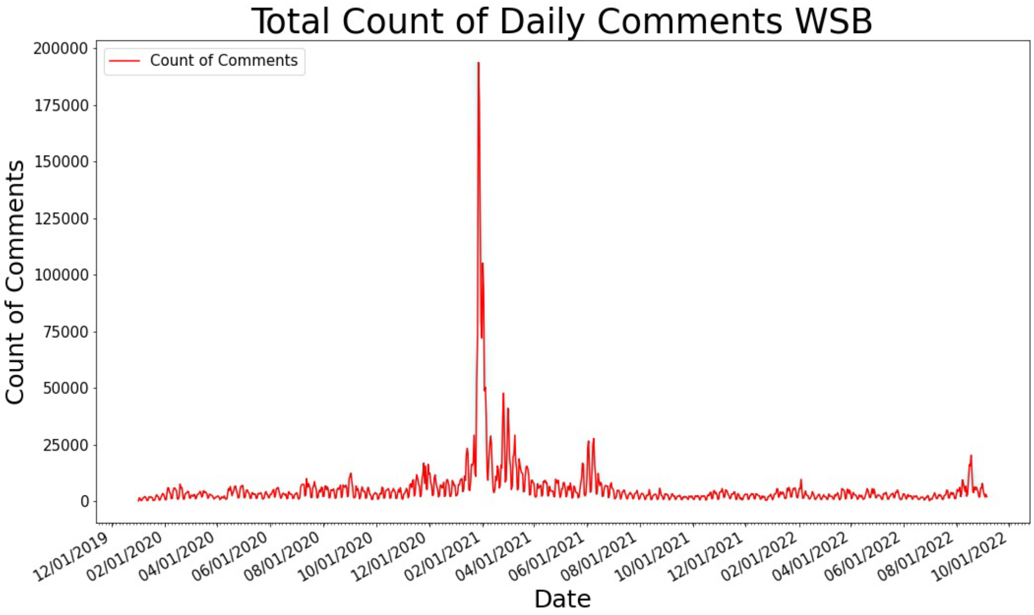

Figure 4: Daily Comments on WSB Concerning Meme Stock Craze

2 Literature Review

Stock movements have been a topic of interest among researchers for many years. Since the beginning of the 21st century, the emergence of various social media platforms has sparked debate on how activity in online communities can potentially impact stock returns. This thesis contributes to the literature by analyzing the aftermath of the meme stock craze in 2021 and studying whether WSB activity provides viable trading signals in a factor strategy. The following part examines the relevant literature regarding our research questions. First, there is a section that describes the power of social media. Then, a review of the existing literature on the effect of WSB's activity on stock prices follows. Finally, the last section reviews the literature on trading algorithms and their potential fallacies.

2.1 Power of Social Media

Social media has experienced tremendous growth since the introduction of the internet and the evolution of smartphones and applications. Four social media platforms are in the top seven most frequently visited web pages worldwide (Semrush, 2022). Millions of users produce a massive amount of data, making it possible to analyze real-world problems by utilizing the available information. At the beginning of the 21st century, scientists started processing data from the internet to predict future outcomes. For example, one paper examines weblog content on movies to predict whether they would succeed (Mishne & Glance, 2006). The results of the study indicate that "sentiment might be effectively used in predictive models or sales" (Mishne & Glance, 2006). Ginsberg et al. (2009) track influenza symptoms by analyzing Google search history. With the intelligence of millions of Google users, the search history "can provide one of the most timely, broad-reaching influenza monitoring systems available today" (Ginsberg et al., 2009). Werner Antweiler and Murray Z. Frank are among the first to use data from social media to predict the stock market. In 2004, they published a paper using Yahoo! Finance's message board to predict the stock market's volatility (Antweiler & Frank, 2004). The paper did not find evidence that message boards can predict returns. However, some of the magnitudes observed are fairly large compared to other relevant features (Antweiler & Frank, 2004).

As the position of social media has grown tremendously, this topic has increasingly become the subject of researcher's interest. Kalampokis et al. (2013) analyze all previous papers on the predicting power of social media. Their paper concludes that social media has a predicting power on the stock market (Kalampokis et al., 2013). As it became evident that social media had predicting power on the stock market, people started to analyze the effect of social media further. In recent years, the Twitter platform has been a favorite among researchers. The paper "Trade the Tweet" examines the movement in stocks listed on the S&P 500 index and the activity of stocks mentioned on Twitter. The results provide a trading strategy in which they maximized the Sharpe ratio, and this strategy performs better than the S&P 500 (Sun et al., 2016).

Another paper that provides a trading strategy on Twitter activity is "Trump, Tweet and Trade" (Cochrane, 2019). Cochrane introduces a sentiment analysis to rank tweets as hostile, positive, or neutral. On the basis of this ranking approach, the author presents a trading strategy based on the activity of Donald Trump's tweets. The strategy yields a better return than the S&P 500 index, and the author reaches the following conclusion: "Keep an eye on Donald Trump's Twitter feed" (Cochrane, 2019). Even though papers identify the predicting power of social media, the trends located cannot necessarily predict the future but may indicate further developments in the market.

2.2 WallStreetBets Stock Movement

After the meme stock craze in 2021, WSB became an intriguing topic among researchers. With an increased interest in the WSB community and the possibility that the number of mentions could affect the stock market, Bradley et al. (2021) published the paper "Place Your Bets". Their paper examines the mentions on WSB and analyzes the due diligence (DD) reports by Redditors in the community. Evidence from their paper identifies that recommendations from the DD reports provide a positive return in the following month. The method this paper uses finds a correlation between stock returns and DD reports. Unfortunately, this significant effect disappeared during the meme stock craze in 2021 (Bradley et al., 2021).

The notable meme stock craze rattled the finance world as the event highlighted the strength of coordinated retail investors. Financial researchers have tried to explain the stock returns given the sentiment from posts on WSB. In the paper "Predicting $GME

Stock Price Movements Using Sentiment From Reddit r/wallstreetbets," Charlie Wang and Ben Lou (2021) try different experiments to predict the movement of GME stock during the meme stock craze. Their results were unable to establish a strong relationship between the sentiment from the post regarding GME and price movements (Wang & Luo, 2021). However, in combination with other features, the sentiment "produces the overall best performance, so sentiment does appear to have some positive impact" (Wang & Luo, 2021).

Another paper that attempts to explain how the activity of WSB affects stock returns is "How online discussion board activity affects stock trading: the case of GameStop" by André Betzer and Jan P. Harris (2022). The regression results indicate that Betzer and Harris could not establish a causal relationship but identified a "significant effect on trading volume, but not on abnormal returns" (Betzer & Harries, 2022). Instead of analyzing the most frequently mentioned WSB stock during the meme stock craze, Anshul Gupta (2022) exploits two of the most frequently mentioned stocks. The results in relation to Tesla (ticker TSLA) were promising, but GME did "not mark a statistical improvement over the random baseline" (Gupta, 2022). Furthermore, the paper's conclusion indicates that overfitting may be a reason for the promising results.

2.3 Trading Algorithms

Algorithms accounted for more than 50 % of the trading activity in US equity as far back as 2012 (Hu et al., 2015). While the purpose of a trading algorithm can vary, it can generally be defined as "the use of sophisticated computer algorithms to automatically make certain trading decisions in the trading cycle, including pre-trade analysis, trading signal generating, and trade execution" (Hu et al., 2015). One of the most significant advantages of using algorithms is the ability to process massive amounts of data in a few seconds and to generate an automatic trading recommendation.

The paper by Chris Peeters (2018) tests the performance of different factor portfolios over 54 years of data on US equity. The findings indicate that the investor will obtain the highest risk-adjusted returns mixing a variety of factor portfolios following a long-short strategy (Peeters, 2018). The combination of portfolios has a diversification benefit, and all the factors tested in this paper provide a positive cumulative return. However, an investor could sustain “considerable losses if they would have stepped in at the wrong time” (Peeters, 2018), due to the volatility in the long-short portfolios.

Furthermore, some biases can affect the output when building and tuning a trading algorithm. First, look-ahead bias occurs when including data not known at the current time to fit the model. Research indicates that look-ahead bias will lead to spurious results (Horst et al., 2001). The second bias that could affect the algorithm negatively is selection bias. This bias is a real challenge, as it could occur if the decisions to include the data are non-random and not observable (Tucker, 2011). In other words, the data does not reflect reality. Tail risk is a further problem with a trading algorithm and many other financial strategies. Tail risk occurs when the value of a portfolio deviates from the current value over three standard deviations (Taleb, 2007). This happens as the returns in the market often violate normality assumptions (Pazarbasi, 2013). The tail risk phenomenon refers to deviation in both directions, but investors are primarily concerned about the downside.

3 Methodology

This section presents the methodology utilized in this thesis. It begins with a description of the theoretical framework for several asset pricing models and goes on to introduce the selected five-factor model. Then, an introduction to the theory for abnormal returns follows. Finally, this section presents the method for this thesis' trading algorithm.

3.1 Capital Asset Pricing Model and Fama–French

At the beginning of the 1960s, Sharpe (1964), Treynor (1962), Linter (1965a, 1965b), and Mossin (1966) introduced the capital asset pricing model (CAPM). The CAPM revolutionized finance as it was the first asset pricing model to explain the expected return of an asset and the related risk. Equation (1) below presents the CAPM model and how the systematic risk will affect the expected returns on a given asset, as follows:

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f) \tag{1}$$

where:

$$R_f = Risk-free\ rate\ measured\ by\ 1-month\ Treasury\ Bill$$

$$\beta_i = Market\ beta\ of\ asset\ i$$

$$E(R_m) = Expected\ return\ of\ market\ portfolio$$

$$\beta_i(E(R_m) - R_f) = Stock-specific\ risk\ multiplied\ by\ the\ market\ risk\ premium$$

Sixty years after the introduction of the CAPM, it still occupies a prominent position in the finance literature. Even though the model explains the correlation between risk and return, Eugene F. Fama and Kenneth R. French find evidence that CAPM does not explain the average returns of assets when sorted by price ratios (Fama & French, 2004). To extend the explanation of cross-sectional average returns, Fama and French introduce the three-factor model (Fama & French, 2004). The model includes the two new factors size (*SMB*) and book-to-market equity (*HML*) to explain average portfolio returns (Fama & French, 1993). After it was introduced, the three-factor model rapidly proved itself to be an essential asset-pricing model. Novy-Marx (2012) and Titman et al. (2004) prove statistically that the three-factor model does not capture all the variations in the portfolio returns. More accurately, the model does not capture how a company's profitability and investment activities affect the portfolio returns. The five-factor model includes investment (CMA) and profitability (RMW) to better explain the cross-sectional average returns of portfolios (Fama & French, 2015).

3.2 Fama French Five-Factor Model

The five-factor model is an extension of the three-factor model, and the two new factors should capture companies' investment activity and profitability. The first Fama-French factor is *HML*. Based on the stocks' book-to-market value, the factor divides the stocks into three groups. The first group is the lowest (L) 30% of the companies, while the middle (M) group constitutes the following 40% of the companies. Finally, the last group constitutes the highest (H) 30% of the book-to-market companies on a stock exchange. Earlier research provides evidence that *HML* will substantially impact the stock returns compared to *SMB* (Fama & French, 1992). The second factor ranks the companies by size and uses median market capitalization, dividing the firms into big (B) and small (S). They merge the aforementioned five groups by constructing the six portfolios S/L, S/M, S/H, B/L, B/M, and B/H (Fama & French, 1993).

To create the *SMB* factor, Fama and French use the six constructed portfolios. First, they reduce the influence of the book-to-market effect and isolate the effect of small and big stocks on the average return (Fama & French, 1993). To isolate the impact of size, they use the "difference between the simple average returns on the three small-stock portfolios S/L, S/M, and S/H and the simple average returns in the big-stock portfolios B/L, B/M, and B/H" (Fama & French, 1993). Hence, the factor will explain the effect of size on a diversified portfolio.

Fama and French make use of the factor *HML* to isolate the effect of book-to-market equity on a diversified portfolio without the influence of size (Fama & French, 1993). To mimic the average returns, they examine the difference between the high and low book-to-market portfolios to provide a return with a minimal effect on the size factor. They examine the difference between the two high book-to-market portfolios, S/H and

B/H, and the low book-to-market portfolios S/L and B/L (Fama & French, 1993).

The last two factors in the model capture the companies' investments and profitability (Fama & French, 2015). Research points towards a correlation between the three factors book-to-market (B/M), operating profitability (OP), and investment (INV)(Fama & French, 1995). Due to the correlation between the factors, the portfolios constructed will not isolate the effect of investment and profitability on average returns. For example, high B/M companies tend to have low profitability and small investment activity (Fama & French, 2015). Hence, they break down the size factor and combine it with book-to-market equity, profitability, and investment to isolate the effects. The new portfolios are SMB/INV, SMB/(B/M), and SMB/OP, removing the problem with correlating factors (Fama & French, 2015).

The objective is to make use of the *RMW* factor to capture the effect on average returns of companies classified with robust operating profitability as against companies with weak profitability (Fama & French, 2015). Furthermore, the companies are categorized on the basis of their size to identify the effect of small robust and big robust as against small weak and big weak. Hence, the model can diversify the effect on the average returns within the factor based on size. The last Fama-French factor *CMA* involves categorization on the basis of how conservative the investment activity of the companies is as against an aggressive approach (Fama & French, 2015). With this method, Fama and French can isolate the effect of investment activity on the average portfolio return. They categorize the companies in terms of small conservative and big conservative companies as against small aggressive and big aggressive companies. The five-factor model predicts the cross-sectional average returns better than the three-factor model (Fama & French, 2015), and the equation (2) below presents the model.

$$E(R_{it}) - R_{Ft} = a_i + b_i(E(R_{Mt}) - R_{Ft}) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{it} \tag{2}$$

3.2.1 Momentum

Fama and French (2011) present momentum (MOM) as an additional factor to explain more of the cross-sectional average returns. With the method of ranking the stocks in terms of size and profitability, the momentum factor explains the difference in average returns between the small and big stocks with a high prior return and the small and big stocks with the lowest prior return (Fama & French, 2011). As a breakout point for the stocks, they use the bottom 30% and the top 30% for the purposes of ranking.

3.3 Model of Choice

As the five-factor model captures most of the cross-sectional average returns of portfolios, it is the model choice for this thesis. Furthermore, the momentum factor is included as it strengthens the model. Therefore, the model selected for the regressions in this thesis is the five-factor model + momentum.

3.4 Abnormal Return

Abnormal return refers to the measurement that explains profits or losses above the expected return. The finance literature identifies the two methods cumulative abnormal return (CAR) and the buy-and-hold abnormal return ($BAHR$) for calculating an asset's abnormal return (Ritter, 1991). The CAR method is favorable for shorter periods, while $BAHR$ is preferable for more extended time frames. Another difference between the two methods is the use of arithmetic average for CAR, while $BAHR$ utilizes a geometric average. The equation (3) below presents the calculation of the abnormal returns.

$$AR_{it} = R_{it} - E(R_{it}) \tag{3}$$

where:

$$R_{it} = Return\ of\ asset\ i\ at\ time\ t$$

$$E(R_{it}) = Expected\ return\ of\ asset\ i\ at\ time\ t$$

The equation (4) below identifies how to summarize the *CAR* method. From the equation, the *CAR* is the sum of all the abnormal returns during the given time frame. Ritter (1991), Barber and Lyon (1997) find evidence that the *CAR* is affected by measurement bias when utilized for extended periods. Hence, the *CAR* method is more accurate when used to express abnormal returns over a short period. Fama (1998) finds evidence that the *CAR* method is statistically better for predicting monthly returns. Hence, using the method to calculate monthly and daily abnormal returns is a popular choice.

$$CAR_{i\tau} = \textstyle\sum_{t=1}^{\tau} AR_{it} \tag{4}$$

As this thesis examines the daily returns of stocks on WSB, empirical evidence demonstrates that the best option is to use *CAR*. Therefore, the method used in this paper is cumulative abnormal returns.

3.5 Optimal Strategy WallStreetBets

3.5.1 Building the Algorithm

There has been a change in activity on WSB in the period following the meme stock craze (Vargas & McCabe, 2022). Therefore, the selected sample period for the algorithm is between mid-2021 and mid-2022. Maximizing the WSB stock universe reduces look-ahead bias, as it is unclear which stocks will be popular in the future at given dates in the sample period. To extend the size of this universe, the algorithm selects stocks from the top mentioned tickers from 2020 to mid-2022. This selection process leads to a total universe size of 41 stocks. The portfolio is equal-weighted, as this is one of the most popular allocation methods used by private investors (Bessler et al., 2021). The portfolio rebalances every day and is long only. There will be no consideration of transaction costs, which in high turnover strategies significantly affect returns (Becker et al., 2007). Finally, the strategy aims to maximize the risk-adjusted return, also known as the *Sharpe ratio* (Sharpe, 1994). Equation (5) below visualizes the Sharpe Ratio.

$$Sharpe\ Ratio = \frac{E(R_p)-R_f}{\sigma_p} \tag{5}$$

where:

$$E(R_p) = Expected\ return\ of\ portfolio$$

$$R_f = Risk{-}free\ rate$$

$$\sigma_p = Standard\ deviation\ of\ porftolio`s\ excess\ return$$

The algorithm builds on a factor strategy that selects the optimal stocks to buy based on how well they are ranked on the basis of several features. First, our algorithm ranks the stocks on each selected feature. The stocks rank in comparison to each other, where the stocks with a relatively better feature rank receive a better score than others. Subsequently, each feature score is assigned a weight of importance. The weighing makes it possible to assign individual informational power to each factor, with a view to improving the strategy's performance (Dicthl et al., 2019). Lastly, the sum of the weighted feature scores gives the individual stock's composite score for each day, given by equation (6). As the stocks with the lower rank are preferred, the algorithm selects the stocks with the lowest numerical composite score at time t. Figure 5 below illustrates the process of the algorithm:

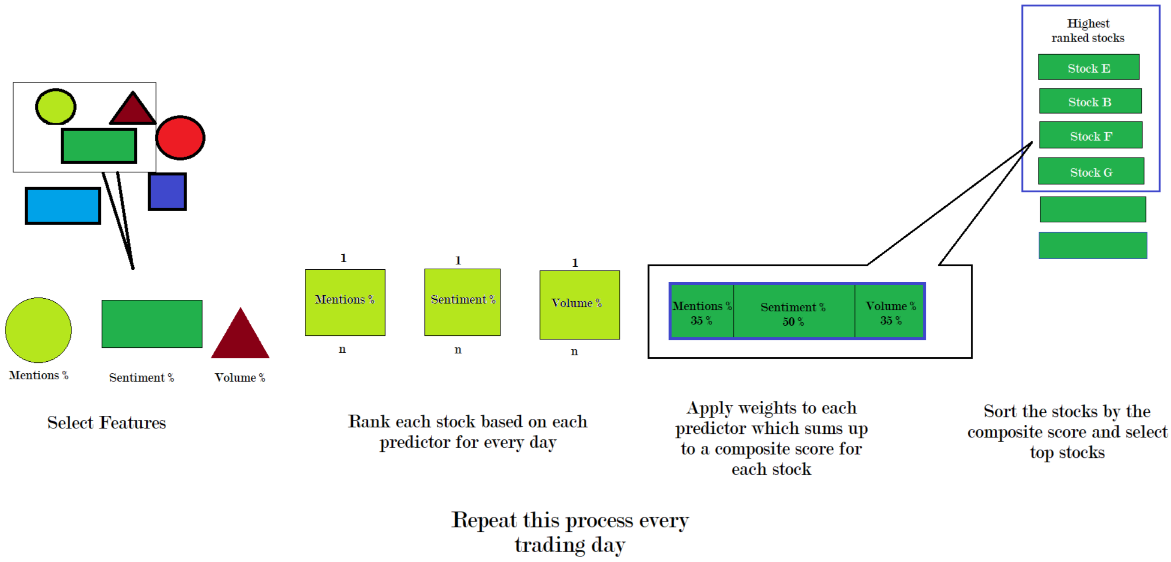
Highest ranked stocks
Stock E
Stock B
Stock F
Stock G
1
1
1
Mentions %
Sentiment %
Volume %
n
n
n
Mentions % 35 %
Sentiment % 50 %
Volume % 35 %
Mentions %
Sentiment %
Volume %
Select Features
Rank each stock based on each predictor for every day
Apply weights to each predictor which sums up to a composite score for each stock
Sort the stocks by the composite score and select top stocks
Repeat this process every trading day

Figure 5: Algorithm Stock Selection Process

$$Composite\ Score_{it} = \sum_{n=1}^{n} featurerank_{ift} * weightfeature_{f} \tag{6}$$

where

$$i = Individual\ stock$$

$$t = Date$$

$$f = Individualfeature$$

$$n = Number\ of\ features$$

3.5.2 Preparing the Data

The data in the algorithm is the same as that utilized by the regressions. An essential problem in quantitative strategies is overfitting the model to historically known data. Therefore, dividing the sample period is a prudent measure to avert overfitting. The training period accounts for 80% of the sample, and the testing period constitutes the remainder. However, the sample period is only one year, which may pose limitations in using the model in the future.

3.5.3 Selecting Features

For the strategy, it is necessary to define variables assigned to every stock. The regression results locate several significant variables that may explain returns in our defined WSB universe. For the trading algorithm, it is possible to lift the constraints regarding multicollinearity and whether the selected features are the best linear unbiased estimators.

3.5.4 Optimizing Weights and Number of Assets in Portfolio

The next step is optimizing each feature's weight and the number of assets in the portfolio. With an example of eight variables to optimize, the number of possible combinations is tremendously high. Therefore, the algorithm utilizes the optimizing function "gp_minimize" from the scikit-optimize library. "gp_minimize" takes a predefined function that returns a target value. Then, along with other arguments such as starting points, boundaries, and number of iterations, it searches for the optimal weights for the given problem. Hence, the function minimizes the negative of the objective value, which in our problem maximizes the Sharpe ratio. Accordingly, it is necessary to solve the following minimizing problem.

$$x^* = arg\min_x f(x) \tag{7}$$

under the assumptions that

- f is not known
- the evaluations of f are expensive
- the evaluations of $y = f(x)$ may contain noise

The minimizer follows a Bayesian optimization loop that utilizes a Gaussian process (GP) regression to search for all possible true functions. The algorithm further utilizes the following acquisition functions that decide the next explored sample of the variables x^*.

Expected Improvement:

$$-EI(x) = -E[f(x) - f(x_t^+)] \tag{8}$$

Lower Confidence Bound:

$$LCB(x) = \mu_{GP}(x) + \kappa * \sigma_{GP}(x) \tag{9}$$

Probability of Improvement:

$$-PI(x) = -P(f(x) \geq f(x_t^+) + \kappa \tag{10}$$

where x_t^+ is the optimal point achieved at time t.

The κ in the acquisition functions controls the exploration-exploitation trade–off, where a high value indicates a higher exploration preference in the problem. The default acquiring method is called "gp_hedge" in the latest version of "gp_minimize". "gp_hedge" is described as choosing the acquiring function probabilistically at each iteration. The acquisition functions are optimized individually and propose each candidate points, and out of these, a SoftMax function chooses the best points to explore. For more information about the optimization process, please refer to example A.1 in the Appendix.

The optimizer identifies the optimal weights on the features and the number of assets in the portfolio. Further, the algorithm assumes a constant value of the weights and the number of stocks in the portfolio. It is important to note that the optimizing process attempts to find the optimal weights per definition, but it is highly uncertain whether these are the globally optimal weights. As described, the number of combinations of the weights is exceptionally high. Therefore, the weights found are assumably locally optimal.

4 Data Section

4.1 Retrieving the Data

This thesis combines data from several sources: Reddit data, stock data, and Fama-French factors. Based on the available data sources and the time frame of the WSB phenomenon, the chosen sample period ranges from 01-01-2020 to 08-01-2022. Our sample period consists of three different sub-periods. The first is pre-2021 and ranges from 01-01-2020 to 12-31-2021. The second is the meme stock craze during the first half of 2021, from 01-01-2021 to 06-01-2021. Finally, the second half of 2021 until 08-01-2022 is the post-meme stock craze period. Due to data availability, the end date of the last period is 1. August 2022.

There is no direct way to download and analyze WSB activity from the Reddit website. As a result, the WSB postings and comments in this thesis originate from Jason Baumgartner's big-data storage and analytics project, pushshift.io. This website allows users to download large quantities of Reddit data for further analysis. Due to the size of the data, the website's API's downloading time is relatively slow for extended periods. Fortunately, a moderator of the pushshift subreddit called Watchful1 has optimized the collection process of the WSB data and made it available as open source (Watchful1, 2022). As a result, all the WSB activity from its inception in 2013 to mid-2022 is available in a single data dump. After downloading this dump, only one month was lacking, which was retrieved directly from the pushshift API at pushshift.io.

There are a few caveats with the retrieval of the pushshift dumps. Historically, the project collects the data at the time of submission. The retrieval process leads to potentially popular comments and posts receiving a lower score. It also limits the opportunity of removing unpopular comments not agreed upon by the community. Finally, the information about the pushshift scraping process was posted three years ago and is subject to change. Therefore, this thesis does not include Reddit scores to analyze activity consistently. The total dataset size from pushshift is approximately 7 GB and includes the following variables.

- created_utc – Publishing date and time of the comment/post in UTC-time.
- body – Text of the comment posted.

- title – Text of title of the post posted.

4.1.1 Timing of the Reddit Activity

The time zone for the WSB activity is UTC, while the stocks in our universe primarily trade at the stock exchanges in New York. Therefore, it is necessary to convert the timestamps of the comments and posts to the US Eastern time zone, which coincides with New York time. The stock data analysis uses close-to-close prices to calculate returns. With the timing of the close price, the data points end daily at 4 p.m. However, this implies that the effect of comments and posts published after close will not show up in the data until the following day. Therefore, the comments and posts published after the close of each trading day are moved to the following day to capture after-hours activity after the previous day's close. Moving the after-hours comments and posts to the next day removes potential look-ahead bias, where the after-hours activity will affect the current day's returns. The timing of the data also reduces the possibility of capturing a relationship between trading from other time zones and the activity on WSB.

After retrieving the variables, the final data set merges posts and comments and sorts by time. Figure 6 below reveals significantly higher activity during trading hours in New York. The next step identifies the top tickers mentioned in the data to find the stocks of interest for the analysis. For this purpose, it is reasonable to use the tickers of the S&P 500 and popular Reddit stocks found on various overviews and news sites to count the number of mentions[3].

[3] Due to processing limitations, the number of tickers searched for in the data set had to be constrained.

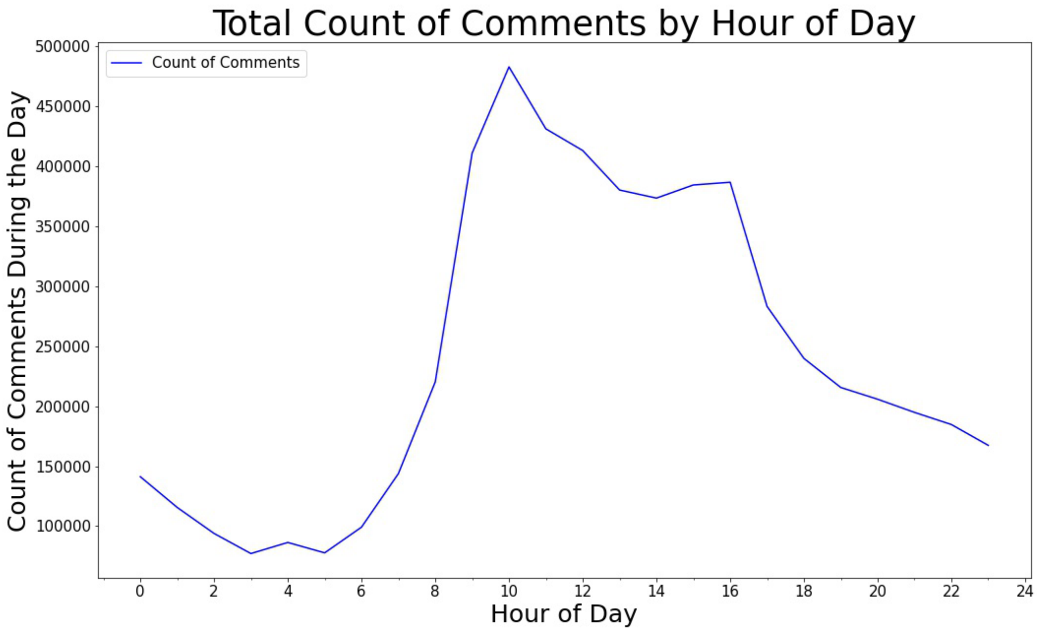

Figure 6: Activity on WSB by Hours of the Day

4.1.2 Selection of Tickers

The ticker list has an initial length of 544. However, the search function for tickers in the text does not differentiate between tickers and what represents a word itself. This problem also applies when combinations of strings are part of a word, and the function assumes it is a ticker. Therefore, it is necessary to omit several tickers which poses problems when counting the occurrence of the tickers in the text. Good examples of this problem are 'EXPE' being part of the word 'expectation', 'WISH' being the word 'wish', and 'SCHW' being part of the often-discussed founder of the World Economic Forum, Klaus Schwab. An important note on the last example is that even though the ticker is not part of a common word, mentioning of specific words or unknown names on WSB may take place. Subsequently, when mentions of words in another context occur, the search function may count the words as mentions for a specific ticker. Therefore, it is prudent to conduct a corroboration of all chosen tickers of interest with the Quiver Quantitative website at quiverquant.com, which among a variety of activities, analyzes WSB mentions of tickers.

Initially, the tickers, comments, and posts turn into lower-case letters to account for users varying between lower- and upper-case typing. Then, after counting the mentions of the tickers and identifying the stocks of interest, it is possible to reduce the data set in such a manner that it contains only rows with the stocks of interest. This process reduces the count of rows from 43 million to approximately 6 million. To reduce the potential problem of stock-specific effects but maintain a proper level of variation, the regression universe in this thesis includes the 30 most frequently mentioned stocks in the sample period of each regression. The universe expands to the top stocks from each sample period for the trading algorithm, and the total size for the latter universe consists of 41 stocks. The analysis also includes the tickers HKD and BBBY as stocks of interest but they are absent in the regressions and algorithms due to the lack of data. Tables A.2.1 and A.2.2 in the Appendix summarize the chosen stocks for the regressions and the algorithm's universe, respectively.

4.2 Sentiment Analysis

To conduct the sentiment analysis, it is advantageous to preprocess the comments and posts in the following manner:

1. For every post and comment, the tokenizing process splits the words into individual elements in a list. The tokenizing process also removes all punctuation.
2. Remove stop words in the English language, such as "on", "at", and "the", as they do not provide any valuable information to the sentiment of the text.
3. The last step lemmatizes the words. Lemmatization means transforming words into their original form. For example, the word "cars" turns into "car". The preprocessing of the data makes it easier for the sentiment package to interpret the sentiment in the text. Emojis are widespread on WSB, and it is important to note that these symbols do not disappear in this process.

The main goal of sentiment analysis is to analyze a text's feelings toward a subject. This thesis analyzes sentiment to explore whether it impacts stock data. Each text receives a sentiment score and attaches it to the stock mentioned. An aggregation of daily sentiment scores for each stock follows. The chosen sentiment analyzer is the Valence Aware Dictionary for Sentiment Reasoning (VADER) model from the Natural Language Toolkit (NLTK) python library (Hutto & Gilbert, 2014). The VADER model takes both polarity and intensity of emotion into account. For clarification, if a sentence consists of a polarized text, it consists of both positive and negative sentiment elements. The VADER model uses a dictionary of words, weighs these words depending on their sentiment intensity, and finally takes the sum of the words of a text to create the final sentiment score. In addition, as the language on WSB is unique (Gendron et al., 2022), custom words and sentiment intensity are added to the dictionary used to analyze the sentiment on WSB. Examples of these custom words are YOLO[4], rocket emoji, and moon. The model returns a negative, neutral, positive, and aggregated compounded text score for each row. For further analysis, the selected variable is the compound sentiment score. Figure 7 below visualizes the aggregated daily sentiment score for all the stocks in the defined universe.

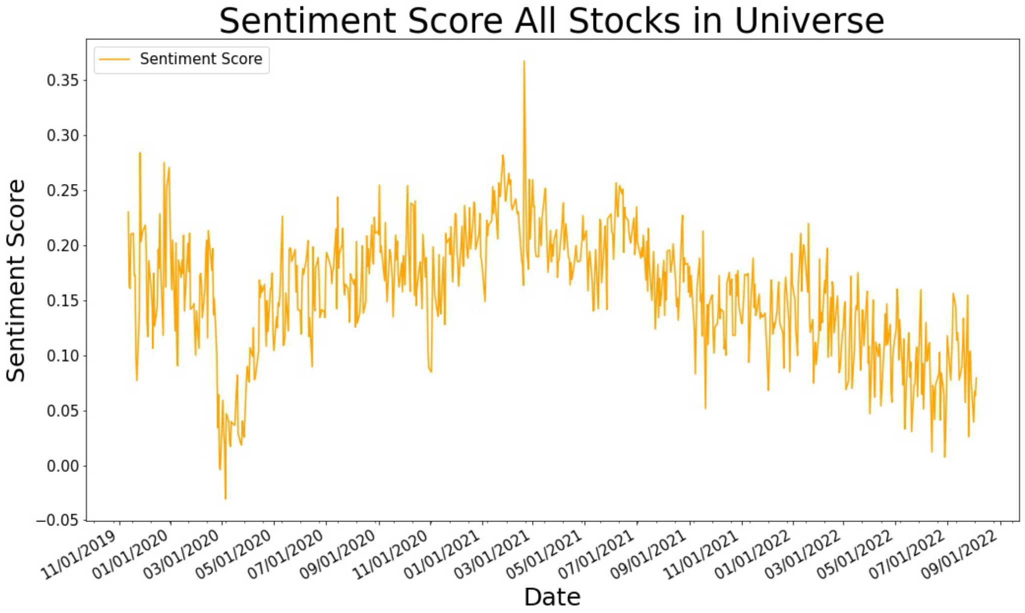

Figure 7: Sentiment Score on WSB Stocks in Universe

4.3 Stock Data

Yahoo! Finance is the source of the individual companies' stock data. The retrieval process collects the daily adjusted close prices, volumes, market capitalizations, and P/E ratios and stores them in a data frame in a long format. Each row in the data frame represents an observation of the tickers in the WSB universe for all trading days in the sample period.

[4] YOLO is an abbreviation for "you only live once" and is a popular term used by generation Z and millennials.

The calculation of the daily percentage return utilizes the daily adjusted closing price. Further, the process collects volumes for call and put options, short interest percentage of float, and implied volatilities from the Bloomberg Terminal and concatenates this with the data from Yahoo! Finance.

4.4 Non-Stationarity in Variables

Conducting regression on a non-stationary time series may lead to spurious regression results (Hill et al., 2001). It is therefore necessary to analyze the variables and adjust for potential non-stationarity. A time series is non-stationary if it contains seasonal effects, inconsistent summary statistics, or structures dependent on time (Woolridge, 2021). The non-stationarity analysis requires a daily aggregation of all the tickers in the universe. This aggregation function provides only one observation for each trading day. Initially, the analysis examines the graphs graphically. Then, if necessary, the augmented Dickey-Fuller (ADF) test indicates whether the time series is non-stationary.

4.4.1 Analysis for Sentiment

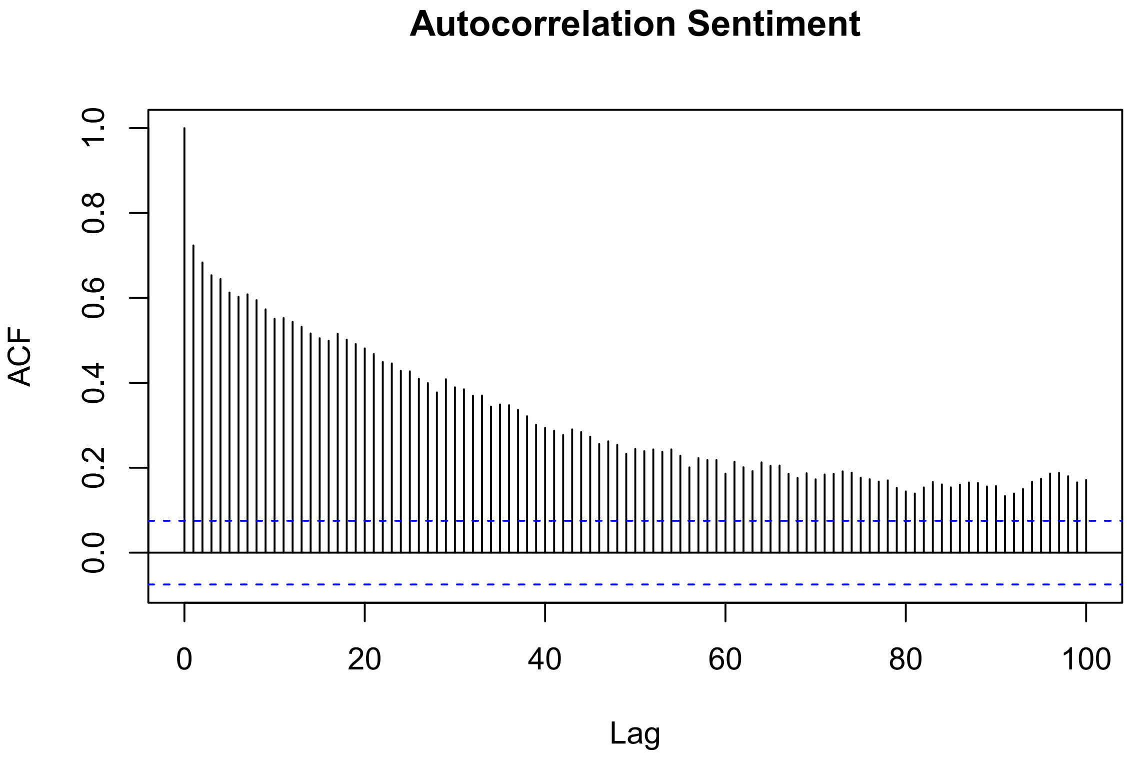

Figure 8: Autocorrelation Plot for Sentiment

4.4.1.1 Autocorrelation As Figure 8 above indicates, the autocorrelation of the sentiment variable is persistent over an extended period. It is therefore necessary to analyze this variable further. In addition, a significant problem in the sentiment score is the inconsistency in the variance when the number of observations changes, specifically when the number is low.

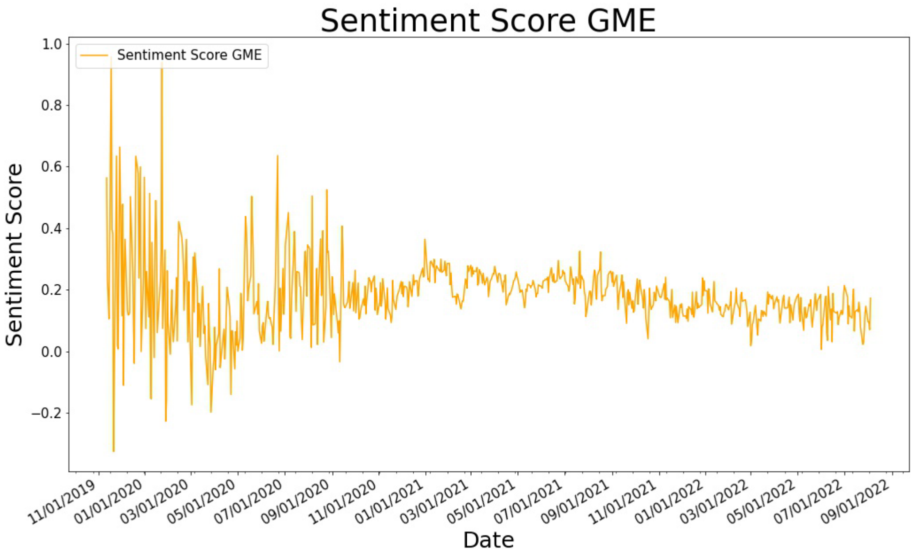

Figure 9: Sentiment Score for GME

4.4.1.2 Augmented Dickey-Fuller Sentiment The ADF test for the sentiment variable gives the following results.

Table 1: ADF Test Sentiment

Test Result	Test Statistic
ADF Statistic	-2.9702
p-value	0.1676

The ADF output for the sentiment series in Table 1 reveals a test statistic of -2.97 and a relatively high p-value, where it is impossible to reject the null hypothesis of a non-stationary series. Therefore, a possible solution for making the series stationary is to integrate the series of first order. Rerunning the test after converting the sentiment to the first difference yields the results presented in Table 2 below.

Table 2: ADF Test Sentiment FD

Test Result	Test Statistic
ADF Statistic	-12.019
p-value	0.01

The ADF statistic improves, and the following analysis in this thesis will examine the first difference of the sentiment variable for each ticker. In conclusion, the sentiment variable is integrated of order one.

4.5 Variable Overview

4.5.1 Dependent Variables

Table 3 below presents the variables that the regressions attempt to explain.

Table 3: Dependent Variables

Variable	Definition
$ExcessRet_{it}$	The logarithm of 1 + the daily excess return. Daily excess return is on a close-to-close basis where the risk free rate is subtracted from the daily % return.
$Volume_{it}$	Individual stock volume at time t.
$PutVol_{it}$	Daily contracts of put options purchased for the individual stock.
$CallVol_{it}$	Daily contracts of call options purchased for the individual stock.
$ImpVol_{it}$	The ten-day implied volatility derived from call options for the stock.

4.5.2 Explanatory Variables

Table 4 below presents the explanatory variables' of the regressions which attempt to explain the variation in the dependent variables. The analysis also includes lagged versions of all the variables except for the FF-factors.

Table 4: Explanatory Variables

Variable	Definition
RF_t	The daily % risk free rate.
$Mkt.RF_t$	The daily % excess return on the Fama-French market portfolio. Can also be expressed as $Rm - Rf$.
SMB_t	The daily return on the Small-Minus-Big Fama-French(FF) portfolio.
HML_t	The daily % return on the High-Minus-Low FF portfolio.
RMW_t	The daily % return on the Robust-Minus-Weak FF portfolio.
CMA_t	The daily % return on the Conservative-Minus-Aggressive FF portfolio.
MOM_t	The daily % return on the Momentum FF portfolio.
$Sentimentscore_{it}$	The first-differece of the compound sentiment score on a daily aggregated basis for each stock.
$Mentions_{it}$	Mentions on an aggregated daily basis for each stock.
$TotalPutCall_t$	The daily put vs. call mentions ratio. The procedure counts and compares the words “put” and “call”. A higher value is, in theory, bearish.
$D(SI100_{it})$	A dummy variable indicating whether or not the daily short interest % of float is above 100 % for each individual stock.
$D(stock_i)$	A dummy variable for each stock in the selected sample universe.

4.5.3 Algorithm Features

Based on initial research, the algorithm utilizes the features presented in Table 5 below.

Table 5: Selected Features in Optimizing Problem

Variable Name	Definition
Sentiment	Daily first-differenced sentiment.
Sentiment %	Daily percentage change in first-differenced sentiment compared to the previous trading day.
Mentions	Daily mentions.
Seven-Day Mentions	Seven-day rolling mean of the mentions. Calculates the mean of the mentions over the last week.
Mentions %	Daily percentage change in mentions compared to previous trading day.
Volume %	Daily percentage change in volume compared to previous trading day.

4.5.4 Overview of Companies

The total universe sample consists of 41 stocks. Therefore, providing a brief overview of the company data may assist in providing a better overview of the companies in this analysis. Figure 10 reveals a small basket of mega-cap stocks, such as Apple and Microsoft driving up the average market capitalization. At the same time, it is evident that a majority of the stocks have a market cap below US$200 billion.

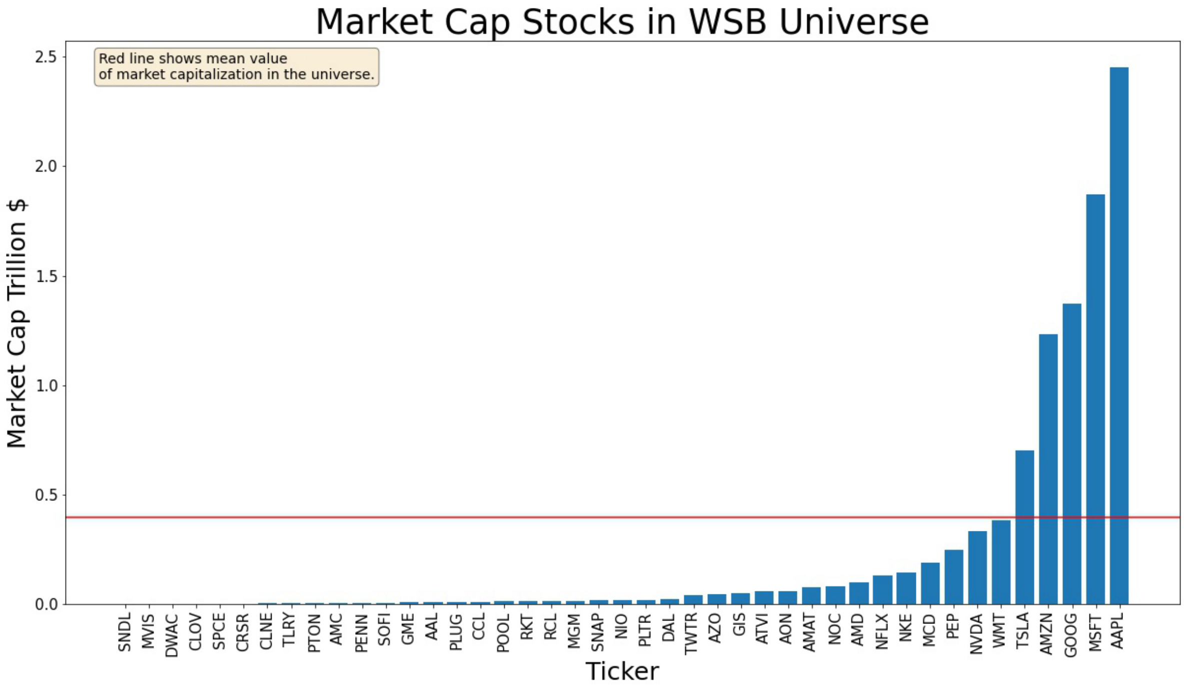

Figure 10: Market Cap for Stocks in WSB Universe

The sector distribution in Figure 11 indicates that most stocks belong predominately to Information Technology, whereas Communication Services follows in the WSB universe.

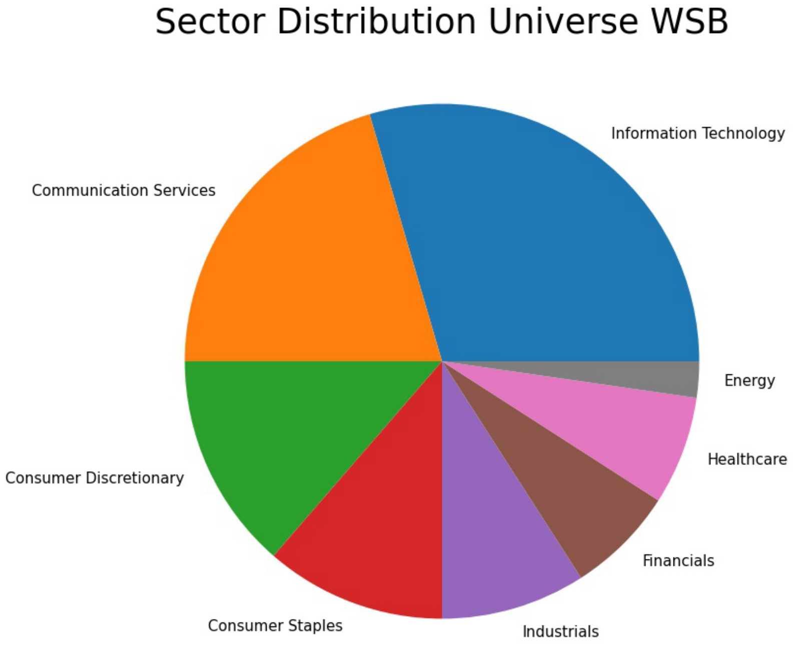

Figure 11: Sector Distribution WSB Universe

The valuation of the companies in the universe is more balanced. As Figure 12 indicates, there is a relatively equal distribution between high and negative P/E stocks with no clear tendency regarding which direction the WSB community favors. The graphs generally remain the same when conducting the same analysis for the sub-periods. Therefore, this brief analysis includes only the total sample universe.

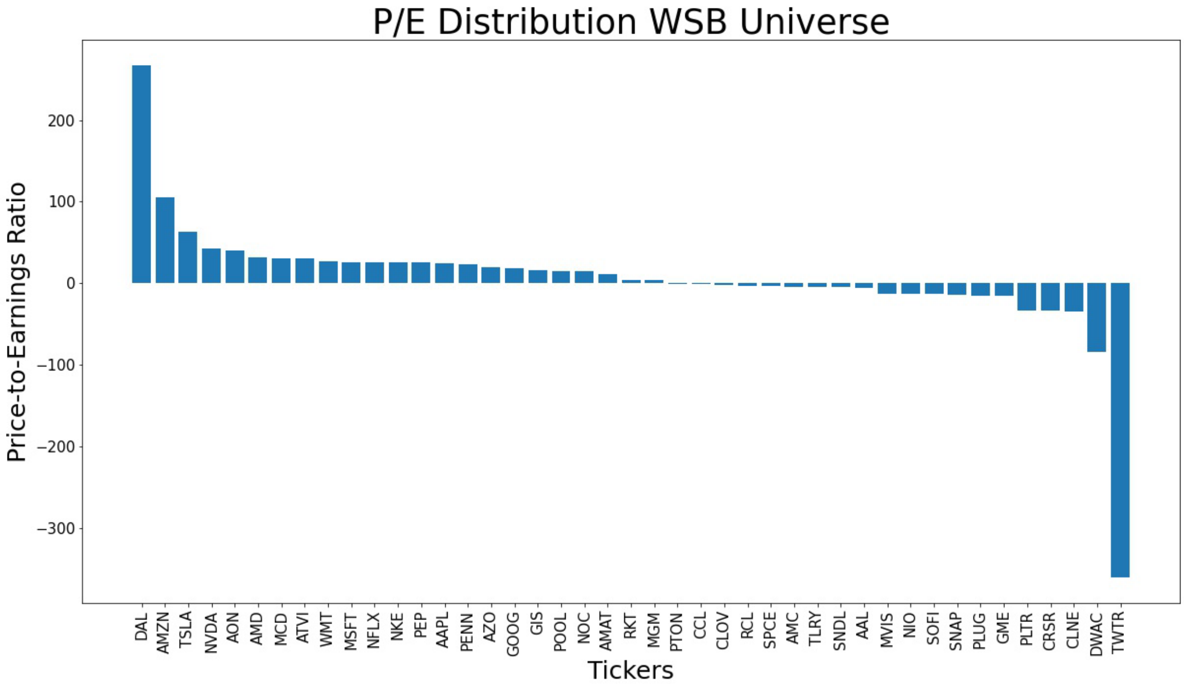

Figure 12: P/E Ratios WSB Universe

4.6 Model: Effects of WSB Activity on Stock Data

This thesis utilizes ordinary least square (OLS) regressions to explain the variation of the defined dependent variables. It is also necessary to include Fama-French factors for the excess returns regressions to account for the risk that derives from these factors. Equation (11) indicates the first regression specification for the analysis. The remainder of the regressions appears in Appendix A.4.

$$\begin{gathered} ExcessRet_{it} = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it} + \\ Mentions_{it} + TotalPutCall_t + D(SI100_{it}) + D(stock_i) \end{gathered} \tag{11}$$

The regressions include a dummy variable for each ticker in the universe to account for individual-specific effects. The omission of Apple Inc. (ticker AAPL) from these dummies leads to this ticker acting as a baseline to avoid multicollinearity issues. There is also a dummy variable that indicates whether the short interest of float is above 100%. The reason for the inclusion of this dummy variable is the known contribution of the short interest of float to the GME short squeeze, as Figure 2 indicates.

The analysis involves regressions for each sample sub-period and the entire sample period. The stocks in each sample period are the 30 most frequently mentioned stocks on WSB, while the entire sample universe consists of 41 stocks. However, the full sample regressions only include the 30 most mentioned stocks from the entire sample period. The sample size for the different sub-periods varies. For example, for the periods before 2021 and after the 2021 craze, the sample size is approximately 7,250 observations. However, the 2021 craze consists only of 3,314 observations. In total, the full sample has 17,285 observations. With regard to the activity on WSB, Figures 1 and 4 indicate that the first half of 2021 is an outlier period. Therefore, the results from this period will likely differ from the other sample periods. Including robust standard errors in the regressions accounts for potential non-constant standard deviations of the predicted variables. The function vcovHC in R produces the necessary robust standard errors.

5 Empirical Results of WSB Activity

The empirical results analysis of WSB activity consists of two sections. The first part examines the WSB effects on stock data and changes in these variables, while the second part analyzes the trading algorithm, given the significant findings from the regressions.

5.1 Excess Returns

5.1.1 Same-Day Activity

	ExcessReturn			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
Mkt.RF	1.19 (0.03)***	1.09 (0.03)***	0.74 (0.04)***	1.19 (0.20)***
SMB	0.70 (0.08)***	0.45 (0.08)***	0.57 (0.09)**	0.25 (0.24)***
HML	−0.42 (0.08)***	−0.08 (0.06)	−0.68 (0.11)***	−0.08 (0.25)
RMW	−0.58 (0.07)***	0.23 (0.07)**	−1.18 (0.12)***	−0.61 (0.31)***
CMA	0.17 (0.18)*	−0.75 (0.11)***	2.60 (0.16)***	−0.33 (0.67)***
MOM	−0.11 (0.04)***	−0.04 (0.05)	0.34 (0.10)**	−0.16 (0.12)***
Sentimentscore	0.60 (0.20)***	0.58 (0.19)**	2.48 (0.21)**	0.74 (0.67)***
Mentions	0.0002 (0.0002)***	0.0005 (0.001)***	0.0001 (0.0003)***	0.0002 (0.0002)
TotalPutCall	−0.001 (0.001)***	−0.001 (0.0004)	−0.001 (0.001)	0.001 (0.001)
I(SI100)	1.02 (0.86)***	−0.52	10.47 (1.02)***	(6.64)
N	16,500	7,222	3,315	7,291
R^2	0.18	0.26	0.11	0.31
Adjusted R^2	0.18	0.26	0.10	0.30
Residual Std. Error	4.80 (df = 16476)	4.03 (df = 7198)	7.23 (df = 3291)	3.27 (df = 7268)

Notes: ***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

lm() function

vcovHC(type = 'HC0')-Robust SE

The first regression of the analysis has the stocks' excess returns as the predicted variable. An important finding in our thesis is that the mentions and sentiment variables are significantly different from zero for all sample periods, except the mentions variable after the 2021 craze. The significant variables give an apparent relationship between the activity on WSB and the excess returns. Moreover, the direction of the variables seems plausible. When the count of mentions increases or the sentiment increases, the excess returns increase. The effect of the sentiment variable changes over time. According to the model, the sentiment has a significant impact on the returns prior to 2021. During the meme stock craze, this effect increases by approximately 400 % before retracting to a 30 % increase compared to before 2021.

Moreover, according to the model, the magnitude of the mentions variable changes over time. Prior to 2021, the effect of mentions on returns is higher than during the meme stock craze. A probable reason for the decrease during 2021 may be the surge in mentions and the subsequent decreasing marginal effect. After the meme stock craze, the mentions variable increases slightly but is no longer significant. When the dummy variable for the short interest of floating stocks is above 100%, it demonstrates a significant effect during the 2021 craze. There is no variation in the dummy variable in the sample period after the 2021 craze. Specifically, there are no occurrences where the short interest of float exceeds 100 % in the data for this period. Hence, the regressions omit the dummy variable for this period. However, the period before 2021 indicates no significant effect. The positive impact during the meme stock craze is likely due to the short squeeze in the GME stock in early 2021, as Figure 2 illustrates.

The Fama-French factors also reveal some interesting changes regarding the exposure of the stocks in the different sub-samples. The beta is relatively close to the market at 1.09 before 2021, but decreases to 0.74 during the meme stock craze. This decrease may be due to the extreme movements in the prices of the most frequently mentioned stocks during the meme stock craze. The factor SMB also indicates a decrease in the exposure to small stocks, where large-cap stocks are relatively more popular on WSB after the meme stock craze. There is also a change in the RMW factor. From discussing robust stocks prior to 2021, there is a substantial decrease during the meme stock craze and thereafter, to more weak and unprofitable companies. The shift to weaker companies may be due to Redditors' motivation to buy heavily shorted stocks to "stick it to the man". Some exciting movements in the *CMA* factor also occur. For example, prior to 2021, there is greater exposure to aggressive companies, while during the meme stock craze, there is a shift to more conservative companies. However, after the meme stock craze, there is again a shift towards more conservative companies and a more balanced exposure between conservative and aggressive. An interesting shift in the momentum factor takes place. There is a preference for momentum stocks during the meme stock craze. This changes in the aftermath of the craze, when the preference shifts to stocks with negative momentum. However, this effect may be due to poor market returns in 2022. The adjusted R^2 for the regressions starts at 0.26 prior to 2021 and decreases to 0.10 during the meme stock craze. However, this decrease was only temporary, as the R^2 increases to 0.30 after the 2021 craze.

5.1.2 Lagged Activity

		ExcessReturn		
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
Mkt.RF	1.19 (0.03)***	1.10 (0.03)***	0.69 (0.05)***	1.19 (0.22)***
SMB	0.70 (0.08)***	0.45 (0.08)***	0.58 (0.09)***	0.26 (0.25)***
HML	−0.43 (0.08)***	−0.09 (0.06)	−0.69 (0.11)***	−0.08 (0.24)
RMW	−0.58 (0.07)***	0.24 (0.07)**	−1.10 (0.12)***	−0.62 (0.33)***
CMA	0.18 (0.18)**	−0.73 (0.11)***	2.53 (0.16)***	−0.33 (0.61)***
MOM	−0.12 (0.05)***	−0.04 (0.05)	0.38 (0.10)**	−0.16 (0.13)***
lag(Sentimentscore, 1)	0.21 (0.26)	0.35 (0.22)	−0.10 (0.26)	0.02 (0.90)
lag(Mentions, 1)	0.0001 (0.0002)***	−0.0002 (0.0004)	−0.0000 (0.0002)	−0.0002 (0.0002)
lag(Sentimentscore, 2)	0.28 (0.27)	0.14 (0.22)	0.56 (0.28)	0.12 (0.87)
lag(Mentions, 2)	−0.0003 (0.0002)***	−0.0002 (0.0003)	−0.0003 (0.0002)***	0.0001 (0.0002)
lag(TotalPutCall, 1)	0.0002 (0.0004)	−0.001 (0.0003)	0.001 (0.001)***	0.0000 (0.001)
lag(TotalPutCall, 2)	−0.0001 (0.0004)	−0.0002 (0.0003)	0.0001 (0.001)	0.0003 (0.001)
lag(SI100, 1)	2.78 (5.63)	−1.81	17.15 (1.22)***	(7.30)
lag(SI100, 2)	−1.81 (5.37)	1.86		
N	16,498	7,220	3,313	7,289
R^2	0.19	0.26	0.15	0.30
Adjusted R^2	0.19	0.26	0.14	0.30
Residual Std. Error	4.79 (df = 16470)	4.03 (df = 7192)	7.09 (df = 3288)	3.28 (df = 7263)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

From the lagged regression output, the only significant variable for the one-day lags is mentions for the full sample period. There is no significant effect of the mentions variable for the various sub-periods. These insignificant effects make it problematic to reject a theory of reverse causality. The only indication to reject this theory for the mentions variable is the significant positive effect for the sample period as a whole. The lagged sentiment variables exhibit no significant effects on today's excess return. This insignificant relationship weakens the possibility of rejecting a theory of reverse causality between sentiment and excess return.

In conclusion, the same-day mentions and sentiment variables significantly and positively affect the excess return variable. However, both variables may be prone to reverse causality. This problem arises when the return movements may also affect the sentiment and mentions variable. The mentions variable loses significance after the 2021 meme stock craze but remains positive. Meanwhile, the sentiment score has reduced since the first half of 2021 but remains elevated above pre-2021 levels.

5.2 Volume

5.2.1 Same-Day Activity

	Volume			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
Sentimentscore	−1,634,421.00	−665,574.10	−6,930,601.00	1,366,699.00
	(1,173,419.00)	(1,158,886.00)	(1,696,599.00)	(4,547,753.00)
Mentions	6,498.69	45,817.63	4,511.01	57,079.50
	(2,149.39)***	(7,244.71)***	(8,422.17)***	(1,710.52)***
TotalPutCall	−1,159.66	2,745.03	4,335.05	−846.62
	(4,124.23)	(2,255.88)	(5,893.83)	(4,356.59)
I(SI100)	20,620,828.00	−303,174.60	136,552,265.00	
	(5,875,759.00)***		(3,119,389.00)***	(61,421,576.00)
N	16,500	7,222	3,315	7,291
R^2	0.36	0.53	0.42	0.51
Adjusted R^2	0.36	0.53	0.42	0.51
Residual Std. Error	42,353,247.00 (df = 16482)	40,718,500.00 (df = 7204)	50,409,564.00 (df = 3297)	23,452,871.00 (df = 7274)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

For all the sample periods, the coefficients of the sentiment score are insignificantly different from zero. Furthermore, the regressions indicate a change in the magnitude and direction of the sentiment variable. For example, before the 2021 craze, the sentiment score is negative and continues to decrease during the meme stock craze. However, this effect reverses in the period after the craze. Nonetheless, as the sentiment coefficients are insignificant, they will not impact our analysis.

However, the mentions variable continues to be strongly significant throughout the sample period. Interestingly, our model indicates a drop during the meme stock craze, while the numbers are fairly equal before and after the 2021 craze. A link between the reduction in the mentions coefficient and the increased activity on WSB is likely. The increased activity, which Figures 1 and 4 indicate, may lead to a decrease in the marginal effect of each mention. Subsequently, the coefficient of the mentions variable for volume drops.

The dummy variable for the short interest above 100 % stands out with some interesting findings. First, the regression results indicate a strongly significant dummy variable during the craze. The significance seems plausible as the short interest on the GME stock resulted in increased trading activity during the closing of short positions. Furthermore, there is no significant effect before the craze. The adjusted R^2 is 0.53 and 0.51 before and after the meme stock craze. However, during the 2021 craze, it drops to 0.42.

5.2.2 Lagged Activity

The lagged regressions in Appendix A.6 also yield some interesting results. All of the lagged sentiment variables' coefficients are insignificant, indicating a possible reverse causality problem between sentiment and volume. However, all the lagged variables have significant coefficients for the mentions variable. The significant variables weaken the theory of a reverse causality problem for mentions for all periods. Further, the magnitude and direction of the effects follow the same pattern for the same-day mentions variable, where the coefficients are all positive and drop during the meme stock craze. The only exception in this regard is the two-day lagged mentions variable before the meme stock craze.

In conclusion, the model finds significant effects between the mentions and the volume variables. The output indicates evidence of rejecting the reverse causality theory for the mentions variable. Further, there are no significant effects for the sentiment variables for the sample periods. However, the regression output fails to reject the theory of reverse causality for the sentiment variable.

5.3 Call Volume

5.3.1 Same-Day Activity

	CallVol			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
Sentimentscore	−6,010.90	1,068.81	46,265.36	−4,215.56
	(14,360.39)	(20,823.91)	(18,006.48)	(96,517.13)
Mentions	27.09	552.63	11.37	635.93
	(8.86)***	(82.60)***	(122.24)***	(4.04)***
TotalPutCall	−56.18	−105.18	0.46	−9.29
	(27.57)*	(35.74)	(84.11)	(53.37)
I(SI100)	−99,535.79	−65,661.16	453,930.00	
	(27,403.97)**		(32,027.90)**	(238,093.80)
N	16,500	7,222	3,315	7,291
R^2	0.64	0.68	0.18	0.70
Adjusted R^2	0.64	0.68	0.17	0.70
Residual Std. Error	591,464.00 (df = 16482)	612,454.70 (df = 7204)	796,881.50 (df = 3297)	466,364.00 (df = 7274)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

The call volume regressions output yields results that are similar to the volume variable. The impact of the mentions variable changes drastically during the sample periods. During the 2021 craze, the effect is approximately 2 % of the effects before and after the meme stock craze. Again, this drop in effect is likely due to the increase in activity on WSB during the meme stock craze, where the marginal effect of one mentions drops.

For the sentiment variables, there are no significant effects in the regressions. However, an interesting direction change occurs during the meme stock craze. The direction of the variables before and after is negative but shifts to a high positive number during the meme stock craze. In addition, the short interest of floating stocks follows the same pattern for call volume as for the two previous dependent variables. The only period the variable is significant is during the meme stock craze, where the variable is strongly significant at the 1% level. The adjusted R^2 for the regressions starts at 0.68 before 2021 and drops to 0.17 during the meme stock craze. Finally, the adjusted R^2 increases to 0.70 in the period after the craze.

5.3.2 Lagged Activity

The lagged regressions in Appendix A.7 exhibit certain significant effects of the mentions variable on the call option volume. For the one-day lagged mentions variables, all are significantly different from zero, except for the meme stock craze period. The only significant variable for the two-day lag is after the 2021 craze at a 5% level. In other words, the regression results are inconclusive regarding the theory of reverse causality.

To summarize, the mentions variable exhibits a significant positive effect on the call option volume. However, the lagged mentions variables only demonstrate certain significant relationships, making rejection of the theory of reverse causality problematic. Finally, the sentiment variables are insignificant.

5.4 Put Volume

The output in Appendix A.8.1 for the put volume yields results similar to those for the call volume. Still, the mentions variable is strongly and significantly different from zero, with a drop in magnitude during the meme stock craze. The sentiment variable remains insignificant for all sample periods. The adjusted R^2 starts at 0.68 and falls to 0.25 during the meme stock craze, before rising to 0.78 in the following period.

However, the lagged regressions in Appendix A.8.2 yield higher significance levels than the call volume regressions. Both the one-day and two-day lagged mentions variables are significantly different from zero. Interestingly, the two-day lag mentions variable is strongly significant during the meme stock craze, but only significant at the 10 % level before and after the 2021 craze.

In conclusion, the results identify significant effects of the mentions variable on the put volume. In comparison to the call volume, the put regressions provide stronger evidence for rejecting the theory of reverse causality for the mentions variable. Further, all sentiment coefficients are insignificant.

5.5 Implied Volatility

5.5.1 Same-Day Activity

	ImpVol			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
Sentimentscore	−1.34	0.25	−1.33	−0.10
	(3.48)	(1.74)	(2.33)	(7.26)
Mentions	0.01	0.02	0.01	0.02
	(0.001)***	(0.003)***	(0.003)***	(0.001)***
TotalPutCall	0.01	0.02	0.02	−0.004
	(0.004)***	(0.002)***	(0.01)***	(0.004)*
I(SI100)	−2.71	−14.39	−1.90	
	(4.33)		(7.03)	(27.58)
N	16,500	7,222	3,315	7,291
R^2	0.34	0.26	0.63	0.44
Adjusted R^2	0.33	0.26	0.63	0.44
Residual Std. Error	45.63 (df = 16482)	37.95 (df = 7204)	47.85 (df = 3297)	24.68 (df = 7274)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

The implied volatility regressions yield a number of significant results regarding the changes on WSB. Foremost, as in the case of the options' regressions, the mentions variable indicates a significant effect for all sample periods. There are no changes in the magnitude between the periods before and after the 2021 craze. As in the other regressions, the effect of mentions decreases during the meme stock craze. Further, the total put/call mentions ratio turns significant for all the sub-periods. The effect is positive for the periods before and during the meme stock craze, indicating an average increase in implied volatility as the relative mentioning of the word put compared to call on WSB increases. This effect reverses into slightly negative territory after the meme stock craze.

The adjusted R^2 starts at 0.26 before increasing to 0.63 during the craze. In the post-craze period, it drops to 0.44.

From the lagged regressions in Appendix A.9, it is clear that all the mentions variables are significant and positive. The significant results contribute to rejecting a theory of reverse causality. The put/call mentions ratio is significant for all one-day lagged variables. For the two-day lagged regressions, the put/call ratio is significantly different from zero only in the period before the meme stock craze. This result provides evidence for rejecting the reverse causality theory between the total put/call ratio and the stocks' implied volatility. All the put/call ratio variables experience a shift from a positive to a negative effect after the meme stock craze. This change means that when the word "call" receives relatively more mentions than the word "put" on WSB, the implied volatility for the stocks rises on average. However, the market downturn in 2022 may impact the direction of this result, as volatility measures increase in the event of market downturns (Whaley, 2008). The market volatility may contribute to noise in the error term, leading to the negative coefficient on the put/call mentions variable.

In conclusion, the implied volatility regressions provide evidence that points towards rejecting a theory of reverse causality for both the mentions and total put/call variables. In addition, the total put/call variable also yields significant results with a questionable direction of effect after the meme stock craze.

5.6 Stock-Specific Events

Trying to explain several companies stock returns using regressions may pose problems regarding stock-specific effects. For example, Figure 13 indicates that the highest movements in returns for the most volatile stocks in the data occur around the meme stock craze. These movements impact the regression results significantly as OLS measures average effects for the different periods. In other words, the spikes in volatility during this period are likely to affect the whole sample period and the period of the meme stock craze. Another problem may arise when sudden company news affects stock returns and volatility. These events, which were not captured by the model, may lead to endogeneity problems.

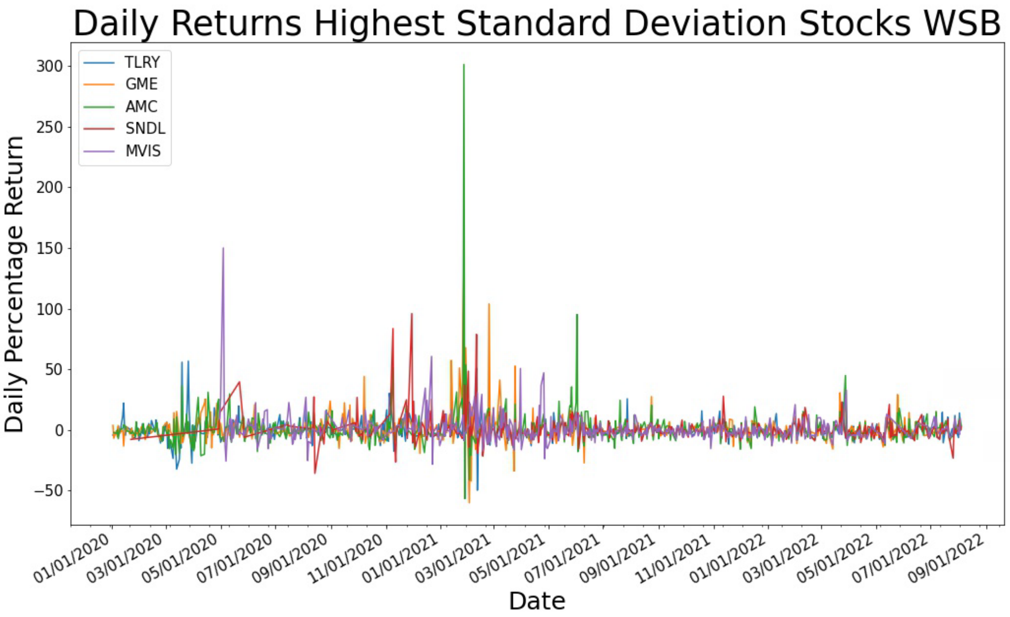

Figure 13: Daily Returns Most Volatile WSB Stocks.

As Figure 13 indicates, the highest spikes evolve around the meme stock craze. However, there are notable spikes before 2021 as well. For the ticker MVIS, a spike of approximately 150 % occurred around May 2020. According to rumors, Microsoft (ticker MSFT) was planning a buyout of the company, leading to a surge in the stock price (Mack, 2020). Moreover, SNDL experienced a spike in late 2020. The company SNDL Inc. holds a portfolio of cannabis producers and experienced high volatility during the 2020 elections. As Joe Biden was declared winner in the American elections in early November, cannabis stocks surged on the hope of cannabis reforms by the new administration. However, as Donald Trump subsequently disputed the election outcome, cannabis stocks experienced high volatility (Kilgore, 2020).

Sudden movements in stock prices occurring due to factors unrelated to WSB may lead to spurious regression results. However, omitting the top five most volatile stocks, including those that experienced the spikes referred to above, does not change the conclusions of the analysis. It is important to note that volatile stocks are a popular discussion topic on WSB. Removing the top five volatile stocks from the sample may lead to important effects being missed. Therefore, the experiment above will only contribute to the discussion of whether stock-specific events may lead to spurious regression results.

5.7 Concluding Remarks on the WSB Activity

The regression results identify different relationships with regard to how WSB activity evolved in relation to stock data. In addition, the main regressions explaining the excess returns display interesting effects of mentions and sentiment. However, the analysis fails to reject a reverse causality theory, especially for the sentiment variable. There are considerable differences for the different sub-periods. A common observation is the outlier period of the meme stock craze. For the excess return, the sentiment effect increases drastically during this period, which later reduces to a level above the period prior to 2021. The mentions variable experiences the same pattern, only in the opposite direction. The marginal effect of mentions appears to drop during the 2021 craze, most likely due to the surge in activity on WSB. The mentions variables also display more robust effects in the period after the meme stocks craze than in the period before 2021.

For the excess returns, the exposure to the Fama-French factors also changes. As the beta of the returns increases to 1.19 after the meme stock craze, a more risky WSB universe is evident. The relatively lower *SMB* coefficient for the period after the meme stock craze indicates a higher preference by Redditors for larger companies. Lastly, the *RMW* variable yields different results for all periods. For the period before 2021, the preference is for robust companies. However, this effect reverses to weak companies during the meme stock craze to -1.18 and decreases by 50% in the subsequent period.

In conclusion, there are significant relationships for both the mentions and sentiment variables towards the dependent variables. The mentions variable primarily provides evidence for rejecting reverse causality problems. However, the regressions yield varying results for the sentiment variable and fail to reject reverse causality issues. Overall, the regressions paint an exciting picture of how WSB activity evolves in relation to stock data. Therefore, it is viable to include the sentiment and mentions variables in the following trading strategy, based on the results of this analysis.

5.8 WSB Strategies

This section presents the results of the custom-made algorithms based on the findings in the activity analysis. Table 6 presents a summary of the common traits of the strategies. First, each section clarifies the assumptions for the algorithms before a summary of the optimal weights follows. Lastly, each section provides the algorithm's Sharpe Ratio and the *CAR*. When analyzing the returns for the chosen sample period, it is essential to note the poor overall market performance in 2022. Therefore, identifying a strategy with a positive return will most likely result in a positive abnormal return, but may display a poor Sharpe ratio. This section concludes with a discussion of the algorithm's drawbacks. Please note that the following results are not investment advice.

Table 6: Summary of Strategy

Strategy	Rebalancing	Sample Period	Long/Short	Test Period
Multi-Factor	Daily	1-7-2021 to 1-8-2022	Long Only	Final 20% of Sample Period

5.8.1 Same-Day Information

The following strategy utilizes the daily information from WSB and rebalances the portfolio the same day. This strategy assumes that the data from one day is constant throughout the day, and it is possible to live-monitor the activity on WSB. For example, this means the relative count of comments for each stock in the universe is constant throughout the day. These assumptions likely violate look-ahead bias. Table 11 in Appendix A.10 presents a summary of the weights in this strategy.

As Figure 23 in Appendix A.10 indicates, the returns for the sample period are strikingly high and most likely unrealistic to replicate. However, it offers an interesting correlation between the weighted selected features and the returns of the selected stocks. As previously discussed, the correlation is due either to the WSB activity or to the price movements themselves. Nevertheless, based on the flawed assumptions regarding the timing of the data, this strategy serves only as an illustration of the correlation between activity on WSB and stock returns.

5.8.2 Adjusted for Look-Ahead Bias

The following strategy applies the signals from one day at the end of the trading day and holds the stocks until the next day's close. It should be recalled that the WSB data after-hours are moved to the next day to avoid look-ahead bias. As a result, this process significantly reduces the bias. Table 7 below provides the optimal weights for this strategy.

Table 7: Summary Weights Strategy Adjusted for Look-Ahead Bias

Feature	Weight	% of Sum Weights
Sentiment Score	49	0.7
Sentiment %-Change	462	6.57
Rolling Mean Seven-Day Sentiment	3387	48.14
Mentions Score	2051	29.15
Mentions%-Change	59	0.84
Rolling Mean Seven-Day Mentions	163	2.31
Volume %-Change	860	12.22
Stocks in Portfolio	3	

The most important features in the strategy are the rolling mean of the seven-day sentiment and the mentions score. These two variables account for almost 80 % of the strategy's weights and exhibit remarkable importance in predicting the following day's return. In short, stocks receiving a positive sentiment the past week and a relatively high number of mentions one day may help predict positive returns the following day. Figure 14 indicates that the returns appear more realistic and still quite impressive. Table 8 presents a performance summary with an abnormal return of 0.4 % and an annualized Sharpe ratio of 1.16. With these findings, it is possible to locate a strategy that performs well during the training and testing periods. However, the optimizer only includes three stocks in the portfolio daily, adding risk due to low diversification.

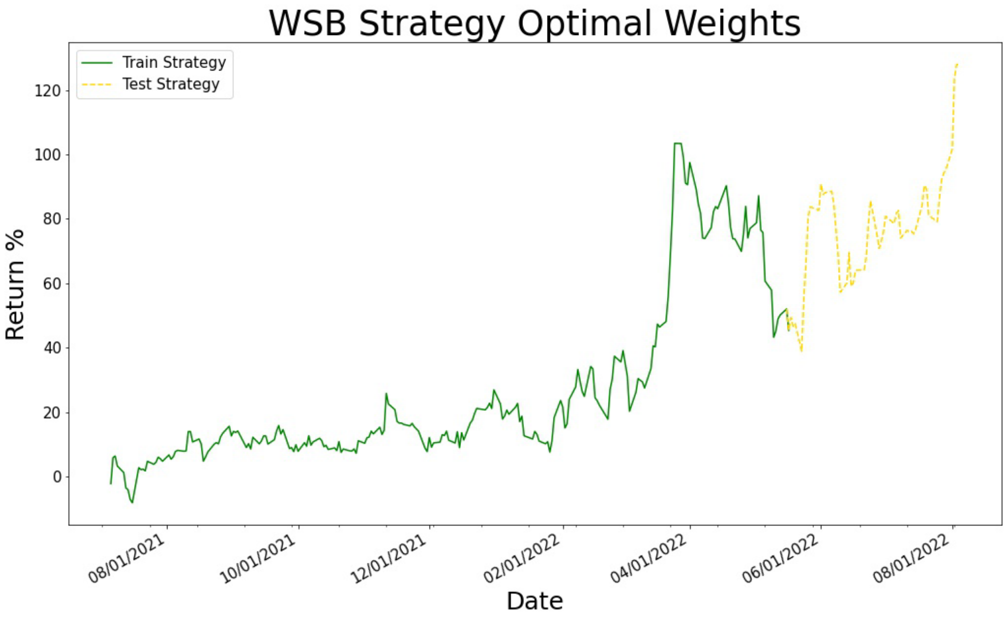

Figure 14: Optimal Strategy Adjusted for Look-Ahead Bias. Signals are Applied the Following Day

Table 8: Summary Performance Adjusted for Look-Ahead Bias

Metric	Value
Sharpe Ratio	1.16
Cumulative Abnormal Return	110%
Cumulative Percentage Return	127.95%
Max Drawdown	32%

Figure 15 presents the periods of drawdowns in the strategy. The worst drawdown occurred in the first half of 2022, when markets experienced a general decline. The S&P 500 experienced a downturn of approximately 24 % in the same sample period as the strategy. Therefore, the strategy's potential downturn is close to that of the general market. Finally, Figure 16 presents the most popular stocks in the strategy.

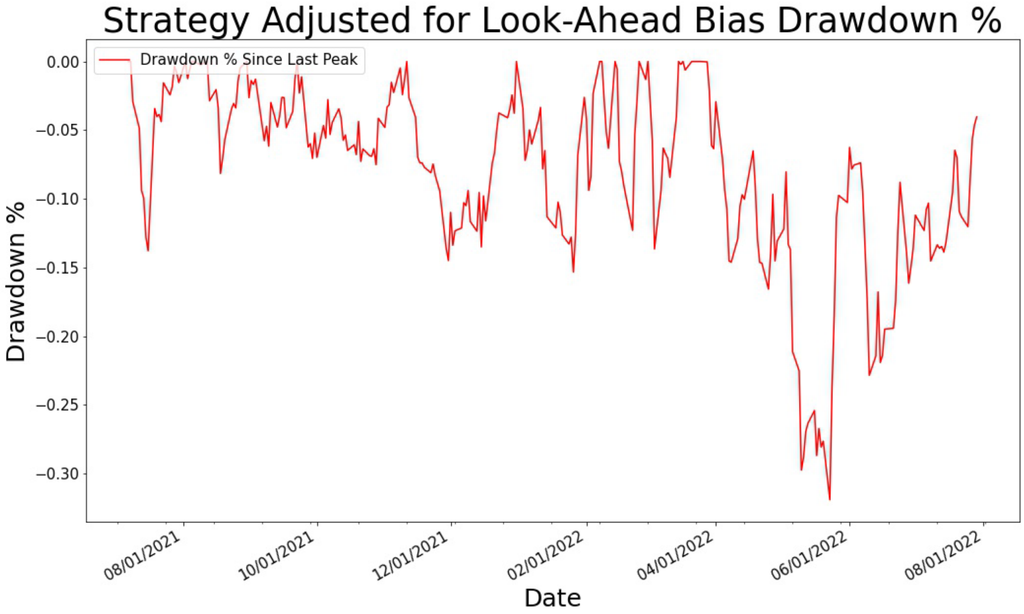

Figure 15: Maximum Drawdown since Last Peak. Strategy: Adjusted for Look-Ahead Bias

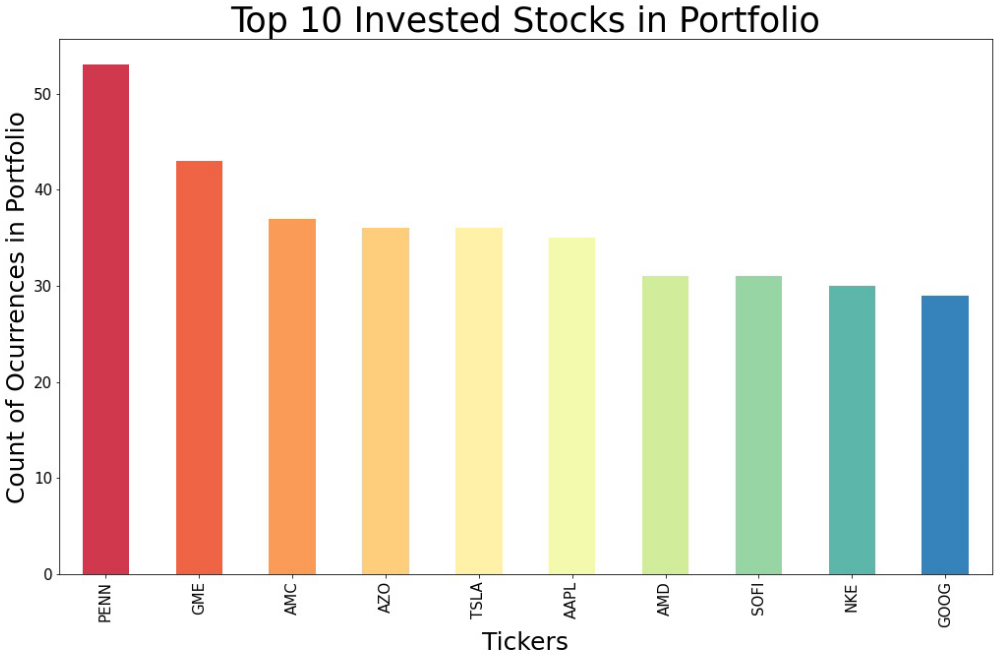

Figure 16: Top Selected Stocks in Strategy Adjusted for Look-Ahead Bias

5.8.3 Drawbacks of the Algorithms

The strategy adjusted for look-ahead bias exhibit a fairly attractive result for risk-seeking investors. However, it is important to note that the trading algorithm has several drawbacks. The first problem involves transaction costs. As described earlier, transaction costs in high turnover strategies significantly impact returns. With our strategy that rebalances daily, it is often the case that the algorithm sells and buys three stocks every day. These transaction costs will significantly impact the returns and depend on the cost structure of brokerage firms. Second, the sample period for the algorithm is short. Therefore, fitting a model to a single year and assuming it will perform well in the subsequent year may be a critical mistake. As demonstrated in the regression results, the effect of WSB activity on returns changes. Therefore, one should be cautious about extrapolating the findings of the algorithms into the future.

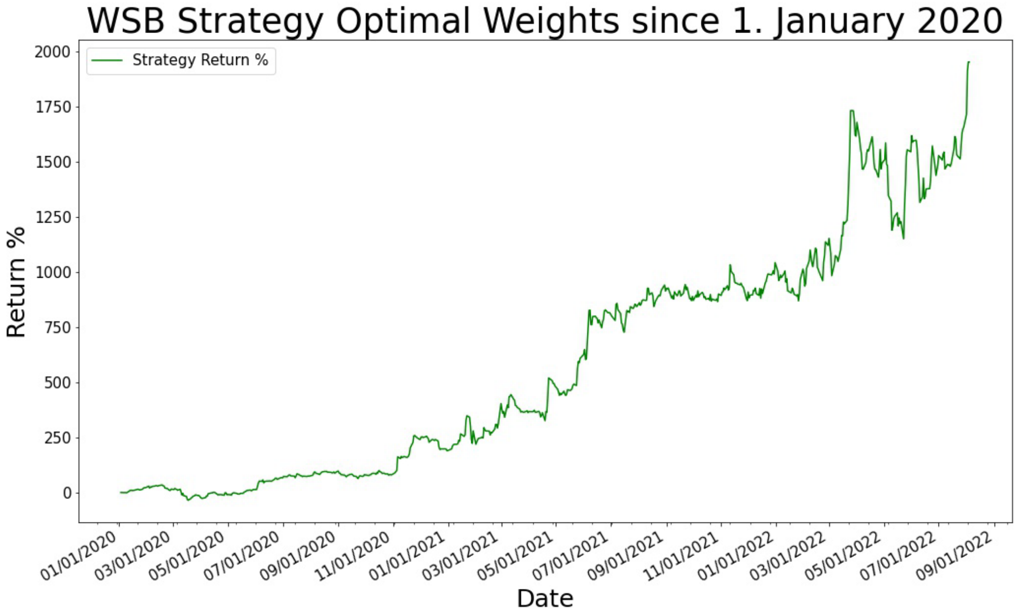

Figure 17: Optimal Strategy Adjusted for Look-Ahead Bias for Full Sample Period

As Figure 17 illustrates, the returns of the optimal strategy adjusted for look-ahead bias have been profitable since 2020. The results indicate that the strategy performs well outside the algorithms sample period, but that it may have a number of drawbacks. First, to implement this strategy, one would need to be familiar with the WSB phenomenon in advance to reap the benefits of this opportunity. Second, one needs to know what stocks to look for and include in the universe. Therefore, implementing this dynamic universe would require constant monitoring of WSB, where one would add the most popular stocks in the algorithm's universe. Third, the universe size of the algorithm is relatively small. Hence, the algorithm is prone to selection bias with a size of 41 stocks, selecting only the average top-mentioned stocks on WSB. Therefore, new tickers experiencing spikes in mentions may be absent in the algorithms universe. Consequently, live tracking the mentions and including new tickers in the universe will improve the algorithm's dynamic properties. A final problem with this algorithm is tail risk. The risk that extraordinary events impose may be a concern when market sentiment deteriorates. For example, Figure 17 indicates a sharp decline at the beginning of 2022, which coincides with the general market decline in the same period. If the model is fit until 1-4-2022, one may believe that the strategy will rise, but instead experience a more than 30 % downturn. This is an excellent example of the tail risk investors face following our trading algorithm.

The strategy displays promising results in the training and testing periods. Given the scenario of the same performance in the future, the strategy yields attractive opportunities for risk-seeking investors. However, the trading algorithm poses several risks and drawbacks that require careful consideration before deeming the strategy viable.

6 Concluding Remarks

This thesis examines the changes in the effect of WSB activity on stock data and investigates how to identify an optimal strategy on the basis of these findings. In line with the discussion of the podcast on WSJ (Vargas & McCabe, 2022), the results of this thesis indicate a remarkable change between the meme stock craze and the following period. The activity on the forum is back to pre-2021 levels, but there are indications of a higher effect on stock data. However, it is important to note that the analysis does not test the changes in these differences directly. Many of the observed variables' effects are within the standard error of the variables from the previous periods. This crucial fact prohibits us from concluding that a significant increase or decrease in WSB effects has occurred.

Further, based on the results, the meme stock craze is an outlier period in the data, with large spikes in several stocks in the sample. The results support the hypothesis of a significant effect of WSB activity on stock data. Tickers' mentions on WSB are the most evident metric from the subreddit to explain variation in the defined dependent variables. Nevertheless, the sentiment variable may also have an effect and yield significant results, especially for the same-day variables. However, the possibility of a reverse causal relationship between the activity variables on WSB and the stock data persists. The evidence for such a relationship is stronger for the sentiment variable, while the mentions variable displays differing results.

Further, of the defined variables, the final algorithm identifies the rolling seven-day mean of the sentiment, daily mentions, and volume %–change as the best means to forecast returns the following day. Notably, the strategy assumes the weights that the algorithm finds are locally and not globally optimal. The algorithm performs strikingly well for the training, test, and the sample period as a whole. However, as with many other trading algorithms, there are caveats following such a strategy. Most importantly, the portfolio size of only three stocks, tail risk, and transaction costs are a cause of concern for investors.

7 Further Research

The results of this thesis identify significant effects of the mentions and sentiment variable from WSB on stock data. As the trading algorithm provides an impressive return in a bear market, it could be interesting to replicate the same techniques on the Norwegian stock market. One example of a similar online society discussing stocks is *Finansavisen-forum*.

In the newspaper *Finansavisen*, different brokers constantly recommend stocks to buy or sell and provide valuations of the companies concerned. As the brokers usually work for an investment bank, they have a dual role when making recommendations, as they probably have an interest in the stocks they mention. Therefore, an interesting approach would be to analyze the different recommendations provided by the brokers and how they affect the Oslo Stock Exchange (OSEX) stock return and include the results in a trading algorithm.

Ellen Chang (2022) introduces her news article with the following words: "Reddit and Twitter are in a hurry to announce the collapse of Credit Suisse". Since the beginning of October 2022, there has been speculation in social media regarding the potential fall of Credit Suisse (Chang, 2022). In the following two months, Quiver Quantitative (2022) picked up an increase in Credit Suisse mentions on WSB, and in the same period, the stock decreased by approximately 40% (Yahoo! Finance, 2022c). Adjusting the algorithm to pick up mentions providing poor sentiment and including these in a short strategy could offer an interesting extension to the algorithm.

The new Netflix series *Eat the Rich* could result in an increase in WSB activity. As the platform receives attention worldwide, analyzing the potential impact on stock data becomes even more interesting. Given the algorithm's short testing period, a new era with more activity could improve the robustness of the strategy as it would include more data with higher variability.

References

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59(3)*(1259-1294). https://www.researchgate.net/publication/4992674_Is_All_That_Talk_Just_Noise_The_Information_Content_of_Internet_Stock_Message_Boards

Barber, B. M., & Lyon, J. D. (1997). Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. *Journal of Financial Economics*, *43*(3), 341–372. https://www.sciencedirect.com/science/article/pii/S0304405X96008902

Becker, Y. L., Fox, H., & Fei, P. (2007). *An empirical study of multi-objective algorithms for stock ranking. Genetic programming theory and practice.* https://ssrn.com/abstract=996484

Bessler, W., Taushanov, G., & Wolff, D. (2021). Factor investing and asset allocation strategies: A comparison of factor versus sector optimization. *Journal of Asset Management*, *22*, 488–506. https://doi.org/https://doi.org/10.1057/s41260-021-00225-1

Betzer, A., & Harries, J. P. (2022). How online discussion board activity affects stock trading: The case of GameStop. *Financial Markets and Portfolio Management*, *36*, 443–472. https://link.springer.com/article/10.1007/s11408-022-00407-w

Bloomberg. (2022). *GameStop corp. (GME).*

Bradley, D., Jr, J. H., Jame, R., & Xiao, Z. (2021). Place your bets? The market consequences of investment research on reddit‘s wallstreetbets. *SSRN*, 1–30. https://ssrn.com/abstract=3806065

Chang, E. (2022). Embattled credit suisse faces social media wrath. In *The Street.* https://www.thestreet.com/investing/embattled-credit-suisse-faces-social-media-wrath

Clement, J. (2021). *Most popular websites worldwide as of november 2021, by total visits.* https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/

Clement, J. (2022). *Worldwide visits to reddit.com from december 2021 to may 2022.* https://www.statista.com/statistics/443332/reddit-monthly-visitors/

Cochrane, N. (2019). Trump, tweet and trade. *Towards Data Science.* https://towardsdatascience.com/trump-tweets-and-trade-96ac157ef082

Dicthl, H., Drobetz, W., Lohre, H., Rother, C., & Vosskamp, P. (2019). Optimal timing and tilting of equity factors. *The Financial Analysts Journal*, *75*, 4, 84–102.

Espiner, T. (2022). *Elon musk warns twitter deal stuck without fake account proof.* https://www.bbc.com/news/business-61432483

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, *49*(3), 283–306. https://www.sciencedirect.com/science/article/pii/S0304405X98000269

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465. https://doi.org/https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stock and bonds. *Journal of Financial Economics*, *33*(1), 3–56. https://doi.org/https://www.sciencedirect.com/science/article/pii/0304405X93900235

Fama, E. F., & French, K. R. (1995). Size and book-to-market factors in earnings and returns. *The Journal of Finance*, *50*(1), 131–155. https://doi.org/https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1995.tb05169.x

Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, *18*(3), 25–46. https://pubs.aeaweb.org/doi/pdfplus/10.1257/0895330042162430

Fama, E. F., & French, K. R. (2011). Size, value, and momentum in international stock returns. *Tuck School of Business Working Paper*, *85*, 1–36. https://www.sciencedirect.com/science/article/abs/pii/S0304405X12000931

Fama, E. F., & French, K. R. (2015). A five factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22. https://doi.org/https://www.sciencedirect.com/science/article/pii/S0304405X14002323

Gendron, Y., Madelaine, A., Paugam, L., & Stolowy, H. (2022). Alternative expertise in financial markets: An analysis of due diligence post on WallStreetBets. *SSRN*, 1–70. https://ssrn.com/abstract=4234609

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influeza epidemics using search engine query data. *Nature*, *458*(1012-1014). https://www.nature.com/articles/nature07634

Gobler, E. (2021). *What is a meme stock?* https://www.thebalancemoney.com/what-is-a-meme-stock-5118074

Grape, M. (2022). *Ny reddit favoritt har steget 21400 prosent paa to uker.*

https://www.finansavisen.no/nyheter/bors/2022/08/03/7908787/ny-reddit-favoritt-har-steget-21.400-prosent-pa-to-uker

Gupta, A. (2022). Wall street vs r/wallstreetbets: Exploring the predictive power of retail investors on equity prices. In *Department of Computer Science, Stanford University.*

Hill, C. R., Griffiths, W. E., & Judge, G. G. (2001). (2nd ed.). Wiley.

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables. Central european labour studies institute (CELSI).* https://CRAN.R-project.org/package=stargazer

Horst, J. R. ter, Nijman, T. E., & Verbeek, M. (2001). Eliminating look-ahead bias in evaluating persistence in mutual fund performance. *Journal of Empirical Finance*, 345–373. https://www.researchgate.net/publication/2521878_Eliminating_Look_Ahead_Bias_in_Evaluating_Persistence_in_Mutual_Fund_Performance

Hu, Y., Liu, K., Zhang, X., Zu, L., Ngai, E., & Liu, M. (2015). Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review. *Applied Soft Computing*, *36*, 534–551. https://daneshyari.com/article/preview/494833.pdf

Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.*

IPOScoop. (2022). *AMTD DIGITAL INC.* https://www.iposcoop.com/ipo/amtd-digital-inc

Kalampokis, E., Tambouris, E. A., & Tarabanis, K. (2013). Understandig the predictive power of social media. *Internet Res.*, *23*, 544–599. https://www.semanticscholaRorg/paper/Understanding-the-predictive-power-of-social-media-Kalampokis-Tambouris/a92b886b952de51518948febe2701cf92da77e75

Kilgore, T. (2020). Cannabis stocks surge after trump administration sets formal transition in motion. In *MarketWatch.* https://www.marketwatch.com/story/cannabis-stocks-surge-after-trump-administration-sets-formal-transition-in-motion-2020-11-24?mod=mw_quote_news_seemore

Linter, J. (1965a). Security prices, risk and maximal gains from diversification. *The Journal of Finance*, *20*(4), 587–615. https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1965.tb02930.x

Linter, J. (1965b). The valuatuion of risky assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, *47*(1), 13–37. https://www.jstoRorg/stable/1924119#metadata_info_tab_contents

Mack, J. (2020). MicroVision (MVIS): Rumors of buyout by microsoft (MSFT) up 238 percent. In *StockPence*. https://stockpence.com/stocks/microvision-mvis-rumors-of-buyout-by-microsoft-msft-up-238

Marsh, B. (2021). *A brief history of reddit*. https://www.thefactsite.com/reddit-history/

Mishne, G., & Glance, N. (2006). Predicting movie sales from blogger sentiment. *ResearchGate*. http://www.aaai.org/Library/Symposia/Spring/2006/ss06-03-030.php

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica*, *34*(4), 768–783. https://www.jstoRorg/stable/1910098#metadata_info_tab_contents

Novy-Marx, R. (2012). Is momentum really momentum. *Journal of Financial Economics*, *103*(3), 429–453. https://doi.org/https://econpapers.repec.org/article/eeejfinec/v_3a103_3ay_3a2012_3ai_3a3_3ap_3a429-453.htm

Otani, A. (2021). *WallStreetBets founder reckons with legacy amid stock-market frenzy*. https://www.wsj.com/articles/wallstreetbets-founder-reckons-with-legacy-amid-memes-loss-porn-and-online-threats-11611829800

Pazarbasi, A. (2013). *Tail risk literature review*. https://caia.org/sites/default/files/2013-aiar-q1-tail-risk_1.pdf

Peeters, C. (2018). *Factor investing strategies: An assessment on performance in u.s. Equity markets*.

Quiver Quantitative. (2022). https://www.quiverquant.com/

Ritter, J. R. (1991). The long-run performance of initial public offerings. *The Journal of Finance*, *46*(1), 3–27. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1991.tb03743.x

r/wallstreetbets. (2022). *R/wallstreetbets*. https://www.reddit.com/r/wallstreetbets/

Semrush. (2022). *Top websites*. https://www.semrush.com/website/top/

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425–442.https://onlinelibrary.wiley.com/doi/10.1111/j.15 6261.1964.tb02865.x

Sharpe, W. F. (1994). The sharpe ratio. *The Journal of Portfolio Management*. http://web.stanford.edu/~wfsharpe/art/sr/sr.htm subredditstats. (2022). *R/wallstreetbets stats* (Vol. 322). https://subredditstats.com/r/wallstreetbets

Sun, A., Lachanski, M., & Fabozzi, F. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, *48*(issue c), 272–281. https://econpapers.repec.org/article/eeefinana/v_3a48_3ay_3a2016_3ai_3ac_3ap_3a272-281.htm

Taleb, N. (2007). *The black swan: The impact of the highly improbable.* Random House Publishing Group.

Thompson, C. (2013). *Twitter trading: 8 tweets that moved market.* https://www.cnbc.com/2013/04/25/Twitter-Trading:-8-Tweets-That-Moved-Markets.html

Titman, S., Wei, K. J., & Xie, F. (2004). Capital investments and stock returns. *The Journal of Financial and Quantitative Analysis*, *39*(4), 677–700. https://doi.org/https://doi.org/10.1017/S0022109000003173

Treynor, J. (1962). Towards a theory of market value of risky assets. *SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=628187

Tucker, J. W. (2011). Selection bias and econometric remedies in accounting and finance research. *Journal of Accounting Literature*, *29*, 31–57. https://ssrn.com/abstract=1756911

u/bighomie69. (2022). *Hit me BBBY one more time.*https://www.reddit.com/r/wallstreetbets/comments

Vargas, L., & McCabe, C. (2022). A year later, reddit‘s WallStreetbets isn‘t the same. In *The Wall Street Journal.* https://www.wsj.com/podcasts/google-news-update/a-year-later-reddit-wallstreetbets-isnt-the-same/90a9e751-4541-4be1-b44d-27f37314c386

Wang, C., & Luo, B. (2021). *FinNLP*, 22–30. https://aclanthology.org/2021.finnlp-1.4

Watchful1. (2022). https://github.com/Watchful1/PushshiftDumps

Whaley, R. E. (2008). *Understanding VIX.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1296743

Wise, J. (2022). *How many people use reddit in 2022? (New statistics).* https://earthweb.com/how-many-people-use-reddit

Woolridge, J. M. (2021). *Introductory econometrics* (7e ed.). Wiley.

Yahoo! Finance. (2022a). *AMTD DIGITAL, INC, HDK.* https://finance.yahoo.com/quote/HKD/history?p=HKD

Yahoo! Finance. (2022b). *Bed bath and beyond inc. (BBBY).* https://finance.yahoo.com/

quote/BBBY?p=BBBY&.tsrc=fin-srch

Yahoo! Finance. (2022c). *Credit suisse group AG, CS.* https://finance.yahoo.com/quote/CS?p=CS&.tsrc=fin-srch

Yahoo! Finance. (2022d). *GameStop corp. (GME).* https://finance.yahoo.com/quote/GME/history?p=GME

A Appendix

A.1 gp_minimize example

The documentation page for the "gp_minimize" function provides the following example with a noisy function to explain the optimizing process.

$$f(x) = sin(5x_0) * (1 - tan(x_0^2)) + \epsilon_t \tag{12}$$

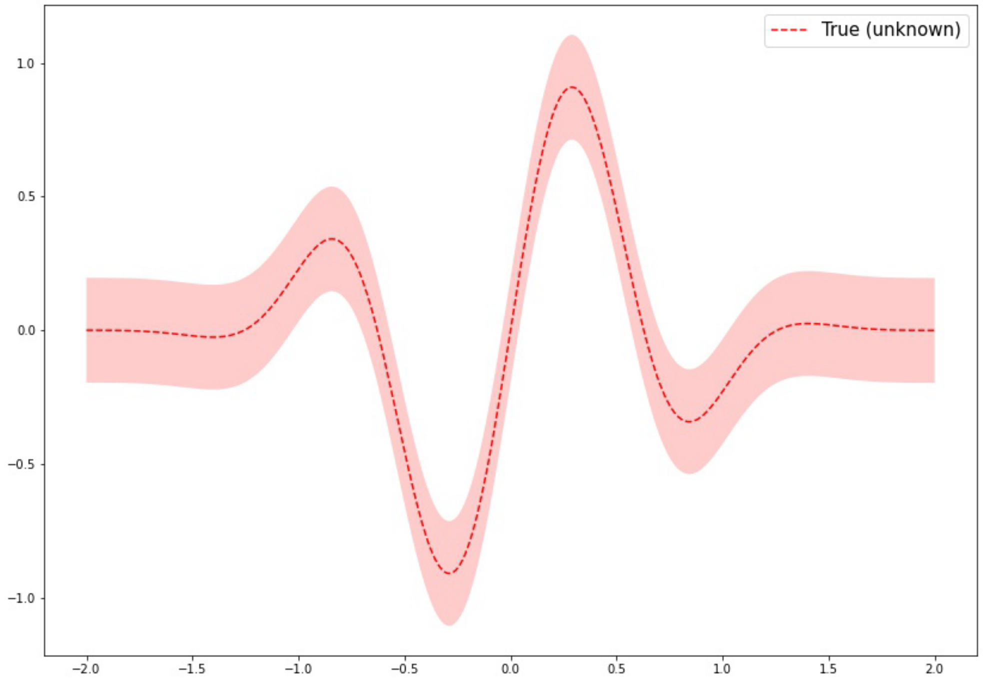

Figure 18: Example Noisy Function.

First, it demonstrates how the GP model fits to replicate the original function with five iterations and how the acquiring function selects the next points to examine. Each iteration chooses the five points randomly.

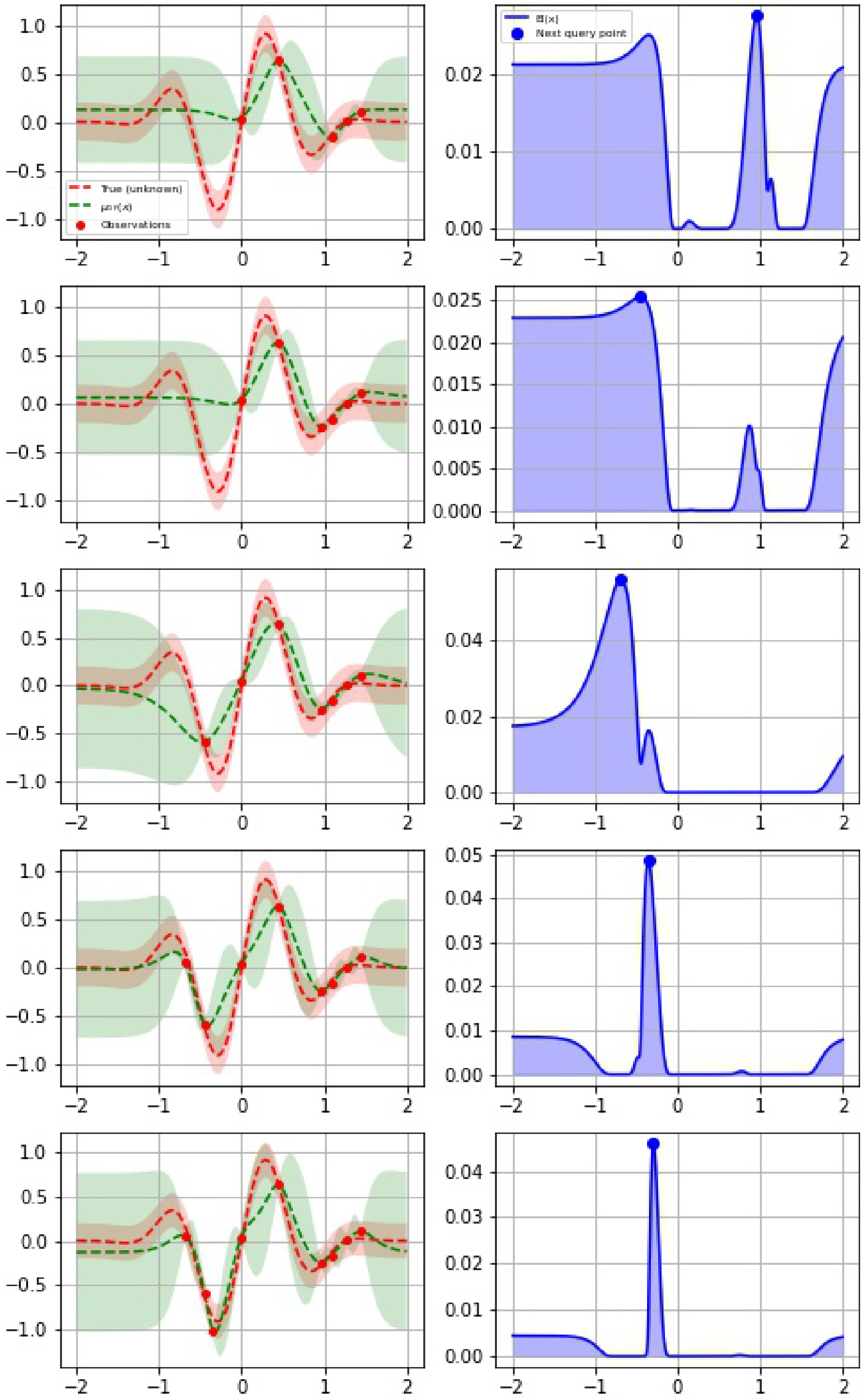

Figure 19: Approximation of the Original Function and Determination of Next Query Points. Five Iterations.

The left-hand column indicates how the true function in red is trying to be approximated by the Gaussian process model in green and the confidence bound of the function following the approximation. The right-hand column illustrates how the acquiring function determines the next query point after every surrogate model fits. Note how the random points for x assist the approximation of the true function. As demonstrated in these iterations, it is important to note that the minimizer may achieve a local instead of a global minimum. In the example, when increasing the number of iterations to 15, the minimizer clusters around the global minimum because there are no gains in the further search for optimal points.

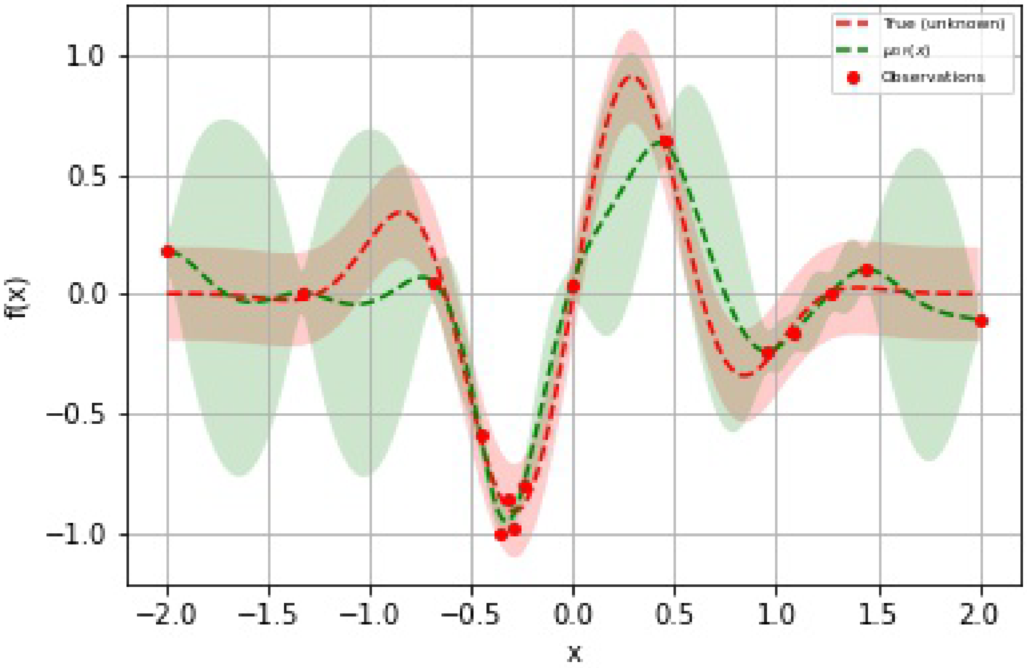

Figure 20: Approximation of the True Function. 15 Iterations

It is also possible to plot the number of iterations needed to achieve convergence in the minimized result of the objective function.

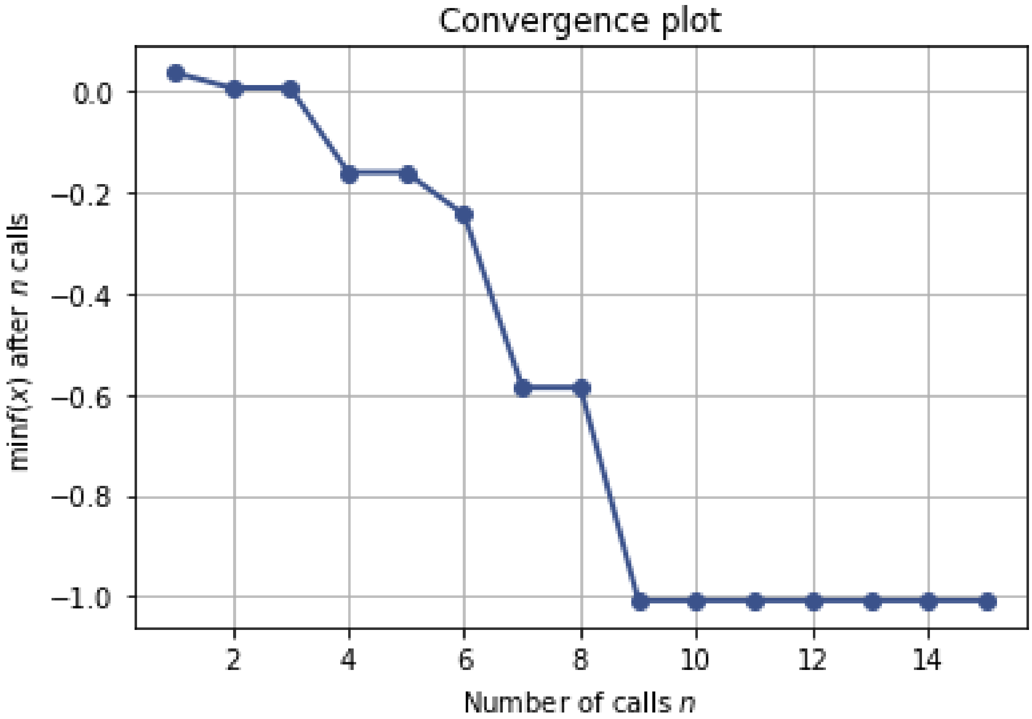

Figure 21: Convergence Plot

This thesis attempts to solve a multidimensional problem to which the abovementioned dynamics also apply.

A.2 Stocks of Interest

A.2.1 Top Mentioned Stocks Sample Periods

	Before Craze 2021	Craze 2021	After Craze 2021	Whole Sample Period
1	TSLA	GME	GME	GME
2	NIO	AMC	BBBY	AMC
3	PLTR	PLTR	TSLA	TSLA
4	AAPL	NOK	AMC	NIO
5	MSFT	TSLA	GOOG	PLTR
6	AMD	NIO	NKE	AMD
7	GME	CLOV	NIO	AAPL
8	AZO	NKE	GOOGL	GOOG
9	GOOG	RKT	AMD	NKE
10	SPCE	GOOG	AZO	GOOGL
11	AMZN	AMD	NVDA	AZO
12	GOOGL	AAPL	AAPL	NOK
13	NKE	GOOGL	PENN	MSFT
14	PENN	PENN	SNAP	PENN
15	RKT	TLRY	AMZN	AMZN
16	SNAP	SNDL	AON	BBBY
17	PTON	AZO	MSFT	RKT
18	NVDA	CLNE	SOFI	SNAP
19	WMT	MVIS	PLTR	NVDA
20	AAL	AMAT	NFLX	CLOV
21	MGM	SOFI	GIS	SPCE
22	DAL	SPCE	NOC	TLRY
23	CCL	NVDA	TWTR	AMAT
24	RCL	PLUG	PTON	PTON
25	ATVI	CRSR	MCD	GIS
26	GIS	AMZN	TLRY	DAL
27	NFLX	GIS	DAL	SNDL
28	PEP	NOC	RCL	PLUG
29	PLUG	MSFT	AMAT	NOC
30	NOC	SNAP	CLOV	SOFI

Table 9: Stocks of Interest in Different Sample Periods

A.2.2 Total Universe

	Ticker	Company Name
1	GME	GameStop Corp
2	AMC	AMC Entertainment Holding Inc.
3	TSLA	Tesla Inc.
4	NIO	NIO Inc.
5	PLTR	Palantir Technologies Inc.
6	AMD	Advanced Micro Devices Inc.
7	AAPL	Apple Inc.
8	GOOG	Alphabet Inc.
9	NKE	Nike Inc
10	AZO	AutoZone Inc.
11	MSFT	Microsoft Corporation
12	AMZN	Amazon.com Inc.
13	BABA	Alibaba Group Holdings Ltd
14	PENN	PENN Entertainment Inc.
15	RKT	Rocket Companies Inc.
16	SNAP	Snap Inc.
17	NVDA	NVIDIA Corporation
18	CLOV	Clover Health Investments Corp.
19	SPCE	Virgin Galactic Holdings Inc.
20	TLRY	Tilray Inc.
21	PTON	Peloton Interactive Inc.
22	AMAT	Applied Materials Inc.
23	SNDL	SNDL Inc.
24	GIS	General Mills Inc.
25	DAL	Delta Air Lines Inc.
26	PLUG	Plug Power Inc.
27	NOC	Northrop Grumman Corporation
28	SOFI	SoFi Technologies Inc.
29	RCL	Royal Caribbean Cruises Ltd.
30	CLNE	Clean Energy Fuels Corp.
31	NFLX	Netflix Inc.
32	MCD	McDonalds Corporation
33	CRSR	Corsair Gaming Inc.
34	MVIS	Microvision Inc.
35	AAL	American Airlines Group
36	WMT	Walmart Inc.
37	ATVI	Activision Blizzard Inc.
38	MGM	MGM Resorts International
39	CCL	Carnival Corporation
40	AON	AON PLC
41	BBBY	Bed Bath & Beyond Inc.
42	TWTR	Twitter Inc.
43	HKD	AMTD Digital Inc.

Table 10: Stocks of Interest

A.3 HKD and BBBY Activity

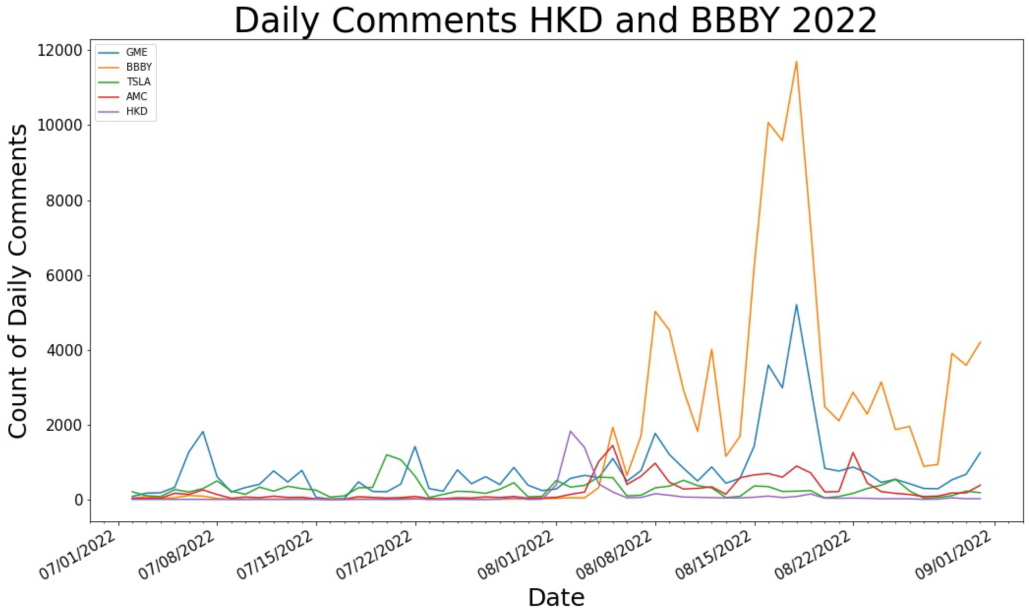

Figure 22: HKD and BBBY Activity in Recent Months

A.4 Regression Specifications

The following lists the remainder of the conducted regressions for all selected sample periods. Note that the sample period will change in the regressions, but the regression specification will remain the same.

$$ExcessRet_{it} = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it-1} + Mentions_{it-1} + TotalPutCall_{t-1} + D(SI100_{it-1}) + Sentimentscore_{it-2} + Mentions_{it-2} + TotalPutCall_{t-2} + D(SI100_{it-2}) + D(stock_i) \tag{13}$$

$$Volume_t = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it} + Mentions_{it} + TotalPutCall_t + D(SI100_{it}) + D(stock_i) \tag{14}$$

$$Volume_t = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it-1} + \\ Mentions_{it-1} + TotalPutCall_{t-1} + D(SI100_{it-1}) + Sentimentscore_{it-2} + \\ Mentions_{it-2} + TotalPutCall_{t-2} + D(SI100_{it-2}) + D(stock_i) \tag{15}$$

$$CallVol_{it} = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it} + \\ Mentions_{it} + TotalPutCall_t + D(SI100_{it}) + D(stock_i) \tag{16}$$

$$CallVol_{it} = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it-1} + \\ Mentions_{it-1} + TotalPutCall_{t-1} + D(SI100_{it-1}) + Sentimentscore_{it-2} + \\ Mentions_{it-2} + TotalPutCall_{t-2} + D(SI100_{it-2}) + D(stock_i) \tag{17}$$

$$PutVol_{it} = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it} + \\ Mentions_{it} + TotalPutCall_t + D(SI100_{it}) + D(stock_i) \tag{18}$$

$$PutVol_{it} = Mkt.RF_t + SMB_t + HML_t + RMW_t + CMA_t + MOM_t + Sentimentscore_{it-1} + \\ Mentions_{it-1} + TotalPutCall_{t-1} + D(SI100_{it-1}) + Sentimentscore_{it-2} + \\ Mentions_{it-2} + TotalPutCall_{t-2} + D(SI100_{it-2}) + D(stock_i) \tag{19}$$

A.5 Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Median	Max
RF	25,290	0.001	0.002	0.000	0.000	0.007
Mkt.RF	25,290	0.056	1.568	−12.000	0.110	9.340
SMB	25,290	0.011	0.883	−4.560	0.010	5.730
HML	25,290	0.017	1.366	−5.000	−0.040	6.740
RMW	25,290	0.035	0.723	−2.150	0.010	4.210
CMA	25,290	0.024	0.614	−2.260	0.000	2.460
MOM	25,290	−0.010	1.547	−14.370	0.090	5.930
Sentimentscore	25,290	−0.0002	0.230	−1.841	−0.0003	1.949
Mentions	25,290	212.859	1,986.319	1	45	184,466
PutVol	25,290	183,911.700	566,348.600	0	20,135.5	9,854,760
CallVol	25,290	286,402.800	820,468.200	0	42,652.5	11,543,560
Volume	25,290	27,186,098.000	45,888,825.000	70,460	10,666,350	1,222,342,500
TotalPutCall	25,290	57.898	160.349	0.609	19.364	2,431.828
ImpVol	25,290	64.863	52.269	7.571	51.082	964.777
SI100	25,290	0.010	0.100	0	0	1

A.6 Volume Lagged Explanatory Regressions Results

	Volume			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
lag(Sentimentscore, 1)	−1,080,228.00	175,378.80	2,440,285.00	−1,507,144.00
	(1,488,987.00)	(1,511,664.00)	(2,111,785.00)	(6,165,106.00)
lag(Mentions, 1)	4,160.71	31,636.69	2,066.06	32,065.90
	(1,949.54)***	(5,525.72)***	(8,841.43)***	(1,487.40)***
lag(Sentimentscore, 2)	−22,795.92	−97,789.07	6,378,178.00	−505,345.30
	(1,464,197.00)	(1,522,238.00)	(2,328,272.00)	(5,909,573.00)
lag(Mentions, 2)	1,308.68	2,672.24	1,062.89	12,690.84
	(1,203.86)***	(3,587.49)	(5,146.98)***	(810.75)***
lag(TotalPutCall, 1)	−60.25	6,091.98	4,782.40	−661.08
	(2,648.95)	(2,148.41)	(6,694.52)	(3,446.91)
lag(TotalPutCall, 2)	−5,616.90	−4,422.23	−4,523.05	−1,237.59
	(1,833.58)**	(2,381.92)	(6,198.97)	(2,481.34)
lag(SI100, 1)	138,751,797.00	1,728,649.00	171,565,184.00	
	(92,147,611.00)***		(4,990,714.00)***	(62,932,832.00)
lag(SI100, 2)	−119,622,164.00	−1,660,183.00		
	(91,712,031.00)***			
N	16,498	7,220	3,313	7,289
R^2	0.33	0.50	0.32	0.45
Adjusted R^2	0.33	0.49	0.32	0.45
Residual Std. Error	43,447,262.00 (df = 16476)	42,383,332.00 (df = 7198)	54,532,136.00 (df = 3294)	24,885,177.00 (df = 7269)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

A.7 Call Volume Regression Lagged Activity

	CallVol			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
lag(Sentimentscore, 1)	7,856.06	23,409.22	−49,717.84	−1,773.60
	(17,009.73)	(23,681.95)	(21,991.23)	(114,991.30)
lag(Mentions, 1)	16.65	312.19	3.91	310.77
	(7.67)***	(65.38)***	(104.22)	(3.96)***
lag(Sentimentscore, 2)	18,198.24	21,971.51	25,331.68	−1,257.10
	(16,799.54)	(23,623.89)	(21,750.26)	(103,805.70)
lag(Mentions, 2)	4.68	23.22	4.16	70.92
	(5.11)	(39.78)	(65.47)	(2.73)**
lag(TotalPutCall, 1)	−5.77	−123.18	31.05	37.94
	(30.99)	(35.79)	(83.33)	(75.00)
lag(TotalPutCall, 2)	−53.92	−351.46	−33.13	42.91
	(24.62)*	(39.74)***	(88.69)	(58.99)
lag(SI100, 1)	415,068.50	−18,041.85	580,670.50	
	(362,082.90)*		(37,310.81)***	(253,947.20)
lag(SI100, 2)	−520,466.80	−21,960.52		
	(360,807.90)**			
N	16,498	7,220	3,313	7,289
R^2	0.64	0.66	0.06	0.69
Adjusted R^2	0.64	0.65	0.05	0.69
Residual Std. Error	593,077.20 (df = 16476)	634,706.80 (df = 7198)	853,144.10 (df = 3294)	477,587.10 (df = 7269)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

A.8 Put Volume Regressions Results

A.8.1 Same Day Results

	PutVol			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
Sentimentscore	−23,129.08	−17,847.41	7,079.05	−26,281.35
	(9,141.04)	(12,336.95)	(12,620.16)	(65,734.87)
Mentions	31.96	284.68	21.34	400.65
	(8.13)***	(60.78)***	(64.94)***	(5.52)***
TotalPutCall	−37.31	95.80	−0.41	−5.72
	(20.80)**	(19.48)	(82.83)	(34.66)
I(SI100)	−50,220.55	−44,948.64	514,031.90	
	(26,229.96)*		(18,459.38)***	(276,552.90)
N	16,500	7,222	3,315	7,291
R^2	0.69	0.68	0.25	0.78
Adjusted R^2	0.69	0.68	0.25	0.78
Residual Std. Error	383,391.50 (df = 16482)	421,315.50 (df = 7204)	532,759.00 (df = 3297)	272,129.60 (df = 7274)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

A.8.2 Lagged Activity

	PutVol			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
lag(Sentimentscore, 1)	−9,968.37	−2,534.85	−36,297.45	−9,724.49
	(10,601.68)	(12,910.76)	(15,157.67)	(75,737.44)
lag(Mentions, 1)	17.97	180.56	7.49	201.30
	(7.21)***	(45.77)***	(62.15)**	(4.82)***
lag(Sentimentscore, 2)	640.58	1,878.75	−1,455.97	−772.76
	(10,714.27)	(14,208.17)	(14,993.37)	(67,428.47)
lag(Mentions, 2)	10.21	−31.83	10.94	29.86
	(4.25)***	(24.46)*	(38.61)***	(3.12)*
lag(TotalPutCall, 1)	−11.32	41.55	14.12	7.64
	(23.16)	(19.78)	(60.22)	(53.33)
lag(TotalPutCall, 2)	−42.18	−125.80	−20.08	10.77
	(18.08)**	(21.45)	(59.76)	(41.46)
lag(SI100, 1)	446,930.60	−5,229.56	772,217.80	
	(337,494.50)***		(23,215.59)***	(335,417.60)
lag(SI100, 2)	−489,521.70	−21,537.44		
	(334,660.70)***			
N	16,498	7,220	3,313	7,289
R^2	0.68	0.67	0.09	0.77
Adjusted R^2	0.68	0.67	0.09	0.77
Residual Std. Error	385,864.30 (df = 16476)	430,676.40 (df = 7198)	586,332.60 (df = 3294)	280,665.70 (df = 7269)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

lm() function
vcovHC(type = 'HC0')-Robust SE

A.9 Implied Volatility Lagged Activity

	ImpVol			
	Full Sample Period	Before 2021	During 2021 Craze	After 2021 Craze
	(1)	(2)	(3)	(4)
lag(Sentimentscore, 1)	−0.42	1.55	1.62	−0.41
	(3.94)	(2.03)	(2.74)	(8.74)
lag(Mentions, 1)	0.005	0.01	0.004	0.02
	(0.002)***	(0.003)***	(0.004)***	(0.001)***
lag(Sentimentscore, 2)	−0.37	2.61	0.51	−0.23
	(4.02)	(2.05)	(2.74)	(8.63)
lag(Mentions, 2)	0.002	0.01	0.001	0.01
	(0.001)***	(0.002)***	(0.003)***	(0.001)***
lag(TotalPutCall, 1)	0.01	0.02	0.01	−0.005
	(0.003)**	(0.002)***	(0.01)***	(0.004)**
lag(TotalPutCall, 2)	0.004	0.02	0.004	−0.003
	(0.003)*	(0.002)**	(0.01)	(0.004)
lag(SI100, 1)	110.39	−9.85	50.49	
	(81.26)***		(16.62)***	(35.27)
lag(SI100, 2)	−110.88	−5.77		
	(81.32)***			
N	16,498	7,220	3,313	7,289
R^2	0.32	0.26	0.57	0.44
Adjusted R^2	0.32	0.26	0.57	0.44
Residual Std. Error	46.18 (df = 16476)	37.98 (df = 7198)	51.72 (df = 3294)	24.74 (df = 7269)

Notes: ***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

lm() function

vcovHC(type = 'HC0')-Robust SE

A.10 Same-Day Algorithm

Table 11: Summary Weights Same-Day Information

Feature	Weight	% of Sum Weights
Sentiment Score	440	31.6
Sentiment %-Change	89	6.3
Rolling Mean Seven-Day Sentiment	18	1.29
Mentions Score	462	33.2
Mentions%-Change	92	6.6
Rolling Mean Seven-Day Mentions	177	12.7
Volume %-Change	110	7.9
Stocks in Portfolio	3	

Table 12: Summary Performance Same-Day Information

Metric	Value
Sharpe Ratio	6.37
Cumulative Abnormal Return	464%
Cumulative Percentage Return	8239%
Max Drawdown	25.5%

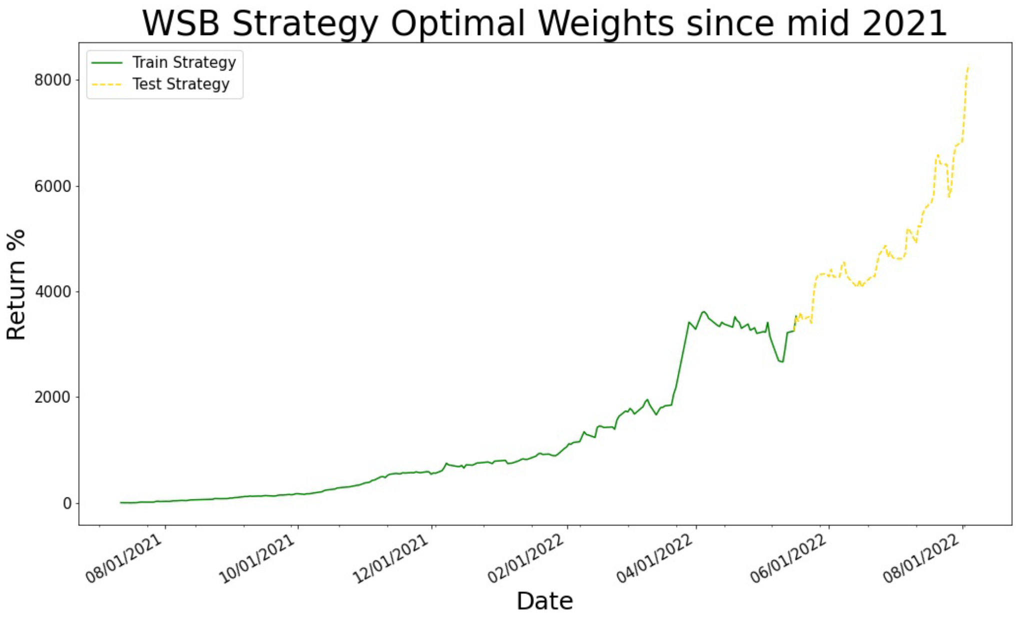

Figure 23: Optimal Strategy Same-Day Information.