Handelshøyskolen BI

GRA 19703 Master Thesis

Thesis Master of Science 100% - W

Predefinert informasjon

Startdato:	16-01-2022 09:00	**Termin:**	202210
Sluttdato:	01-07-2022 12:00	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	T		
Flowkode:	202210||10936||IN00||W||T		
Intern sensor:	(Anonymisert)		

Deltaker

Navn:	Kristoffer Vold Hytten og Dennis Holmvik

Informasjon fra deltaker

Tittel *:	Trading Social Media: Just Noise or Predictive Power? (An Empirical Analysis)
Navn på veileder *:	Kjell Jørgensen

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	123
Andre medlemmer i gruppen:	

Trading Social Media: Just Noise or Predictive Power?
An Empirical Analysis

Master Thesis

by
Dennis Holmvik and Kristoffer Vold Hytten
MSc in Finance

Supervisor:
Kjell Jørgensen
Department of Finance, BI Norwegian Business School

Oslo, June 30, 2022

ABSTRACT

We study the relationship between two large social media forums on Reddit and the stock market. We look at whether it is possible to create profitable trading strategies based on data collected from the subreddits and whether social media discussions have an effect on individual stock trading volume and returns. We find that it is possible to create profitable strategies and that the forum members seem to be better at identifying profitable long opportunities rather than short opportunities, and that more frequent trading seems to be the most profitable. We also find that increased social media attention affects trading volume for all companies in our sample, as well as returns for some.

This thesis is a part of the MSc programme at BI Norwegian Business School. The school takes no responsibility for the methods used, results found, or conclusions drawn.

Acknowledgements

We would like to thank our supervisor, Kjell Jørgensen, for his guidance and feedback throughout the process of developing this thesis. His input, feedback and assistance was of great value.

Contents

List of Figures

List of Tables

1 Introduction and motivation

The world is becoming increasingly digitalized, and the technological advances seen in the last 20 - 30 years are one of the largest in history. For example, computing power has become so accessible that basically everyone has access to a laptop that would have been considered a supercomputer 20 years ago. The development also appears to not be slowing down as we are still being presented with new technology to this day, self-driving cars being one of the most prominent.

How we communicate with each other has changed a lot as well. With the rise of social media platforms such as Facebook, Snapchat, Twitter, and Instagram, communicating and sharing content with friends, family and even people you don't know is easier and more efficient than ever before. There is essentially no area of our lives that has not been affected by technology in some form, and finance is no exception. The financial industry has moved from traders physically being present on a trading floor to almost every transaction happening through computers and algorithms. These technological advances have made the industry more efficient and made trading, investing, and financial information increasingly available to the general public. Setting up a brokerage account takes minutes and instantly provides an individual access to trading from their personal computer or smartphone. In some cases, even allowing for the trading of complex financial derivatives. In addition to this, a quick google search will give large amounts of information such that the investor can make informed decisions.

The increased availability has also led to increased interest, which in the last years has led to a rapidly increasing number of retail investors. In combination with the growth of social media, we have seen forums related to the financial market arising on numerous platforms such as Twitter and Reddit over the last few years. These groups/forums are venues where retail investors discuss various finance-related topics and share information. The degree of professionalism varies greatly among these forums, where some focus mostly on bragging about profits, joking about 5-6 figure losses, and placing very large and risky positions. Investors sharing information is nothing new, but the way they do it has changed rapidly over the last few years, with social media becoming the main channel for most retail investors. Social media makes the information available to anyone looking for it, and researching the effects social networks have on decision-making becomes easier (Pedersen (2021)).

The financial markets through 2021 were interesting, especially at the beginning of the year when one of the finance forums on Reddit named WallStreetBets attempted to short squeeze hedge funds who were shorting GameStop stock by collectively buying a large number of the outstanding shares. The short squeeze was initially motivated by a dislike for Wall Street and its large financial institutions, probably stemming from the 2008 financial crisis. However, most of the investors were not buying for this reason. They were simply buying because of the attention it had and a belief that joining in would make them a lot of money. This belief was largely driven by the group members discussing extreme potential valuations. The events led to a large increase in trading volume, price and volatility. The price peaked at around 400 dollars before it fell. The stock price as of the end of June 2022 sits at around 120 dollars, which still is approximately 20 times more than the price before the squeeze. We believe this event illustrates the powerful networking effects of social media and how investors may influence each other into making decisions, both in terms of less extreme events and events like the GameStop case.

This thesis will analyze the social media networking effects, its effect on stock prices, and trading volume, and if this is something rational investors potentially could utilize to identify profitable trading opportunities. By analyzing data from large social media platforms, we believe it will be possible to measure the sentiment of a large (and growing) part of the investor base and investigate whether there is a connection between social media trends and stock market movements. We will then use this sentiment to construct portfolios and determine whether social media sentiment can be traded upon and explain changes in other characteristics of a stock.

2 Literature Review

With the growth in the number of retail investors and social media forums related to finance over the last decade, we have also seen a growth in the amount of research conducted on the topic. In this chapter, we will be reviewing some of the most relevant existing literature and discussing approaches, theories and results.

2.1 DeGroot Model

Research conducted on how individuals behave in groups and is influenced by other members is nothing new. However, as the way we communicate with each other changes, how we analyze networking effects will have to be altered as well.

Pedersen (2021) presents a model of how investment ideas can spread through social networks and affect behavior and prices. He did so by analyzing financial markets using a standard DeGroot (1974) model. The standard DeGroot model captures how individuals with different subjective probability distributions for the unknown value of some parameter (e.g., price of some stock) is affected by other members' subjective beliefs. In addition, the model examines how the group reaches a common probability distribution by pooling their individual opinions (DeGroot (1974), p. 2). In a standard DeGroot model, it is assumed that all participants are naive and update their views based on a subset of people within the group which they for different reasons (e.g., the perception that they are more informative than the rest) choose to devote a larger part of their attention to. This phenomenon was researched by Demarzo, Vayanos and Zwiebel in 2003 under the term "persuasion bias" (DeMarzo et al. (2003)). They argued that an individual's way of selective hearing and failing to account for repetition in information they receive could explain how a person may be influenced to act in a certain way. For example, if a particular politician is given more time to speak in a debate than another, this could influence an individual's opinion to be more in line with the politician speaking the most. This is considered irrational behavior, as airtime would not affect the degree of influence on a rational person (Chen et al. (2014)).

Pedersen (2021) attempts to describe the networking effects behind the GameStop events and how market participants interact and bubbles form in general. He did this by building on the standard DeGroot model and introducing rational investors, to capture professional investors in the market. These rational investors are characterized by initially listening to everyone (not selecting some to listen to), and updating his/her view until all the available information has been processed. At this point, the rational agent stops listening to other members of the group if the information they are presenting is not new. The rational investor essentially differs from the naive in two ways: By not valuing some opinions more than other initially, and by accounting for repetition of the same information. This means that the naive investors over time will be affected by the repetition of the stubborn opinions coming from the rational agents. This should in theory make the market more rational. However, some irrational, fanatic and stubborn agents are also present in the groups and may influence the naive agents over time if they are willing to listen to them. This means that over time the common view of the group will be dominated by the views of the two different types of stubborn participants, and the naive agents will differ in whether they are weighing the rational or irrational views heavier (Pedersen (2021), p. 9-10).

The model Pedersen presents sheds light on the dynamics behind the GameStop frenzy and may explain how irrational views repeatedly presented may over time lead to bubbles and be a driver of (retail) investor behavior in general. The results from Pedersen (2021) helps shed light on the psychological processes behind what happens on social media platforms. This model creates a foundation of some psychological theory which we will build our analysis on. It helps us understand why people act in a certain way and how an individual's social network affects him/her.

2.2 Social media overtaking traditional media

Individuals presented with new information may change their views, opinions, and how they act, as discussed in the subchapter above. Traditionally, a lot of the information investors consumed came through more conventional channels such as newspapers or news presented on television. This has changed drastically over the last decade, as we briefly discussed in section 1. Now, social media is the primary source of information for a large part of the investor base. A lot of the information on social media is of poor quality, but regardless of this it will affect the decision-making of an individual collecting and believing in this information. More traditional media is however still present, and information can be obtained through these channels as well.

Yu et al. (2013) aim to determine how social media and conventional media affect short-term fund performance, which of the two has the highest relative importance, and whether the two can interrelate and create an amplifying effect (Yu et al. (2013), p. 1). The researchers use a large dataset of daily media content from both social media outlets and traditional outlets for 824 publicly traded companies. They use the stock performances of these companies as a proxy for the fund performance, mostly because information on stock performance is more available than a lot of the other metrics on how a firm is doing. They employ what they call an advanced sentiment analysis, considering more than simply the number of mentions to analyze the overall sentiment of the media source towards a company. Sentiment analysis is described in section 3 of this thesis. The researchers then examine the return and risk of the firm's stock to examine performance related to the sentiment of the media source.

As we suspected, the researchers concluded that social media has a larger effect on stock performance than traditional media, while both have a strong interacting effect.

This paper sheds light on the fact that the media has an effect on stock performance and that this effect over the last years has shifted to social media being the most prominent when compared to the traditional media sources. There are no signs of this development slowing down or reversing, indicating that social media most likely will play an even bigger role in influencing human decision-making in the years moving forward.

2.3 Social media platforms

Research conducted on the different social media platforms' effect on equity prices has increased along with the increase in the number of platforms themselves. Chen et al. (2014) looked at the most popular social media platform in the U.S. for investors at the time, Seeking Alpha. They employed a textual analysis, which looked at the frequency of negative and positive words used in an article to capture the sentiment of the report. This was done on the article and the comments it received from other users.

To calculate the returns, they created quintiles, based on the fraction of negative words across the articles and comments. They went short the most negative quintile and longed the least negative quintile. Looking at abnormal returns, they were able to see that an increase in negative words (as a fraction) would decrease the returns over different holding periods, controlling for various characteristics. The authors assumed that the overall tone of the articles could be determined by the frequency of negative words. The larger the fraction of negative words suggests an increasingly negative view of the company among the users.

Zhang et al. (2011) conducted a similar analysis on Twitter. Over a six-month period, they collected a subsample of data from the platform, analyzed the sentiment for each day and looked at its correlation with the major stock indices. The authors selected a six-month period in 2009 and collected a small, randomized subsample, which corresponded to approximately 1% of all the tweets over the period. They used a simple approach for deciding the sentiment of the tweets; counting single words like "fear" and "hope" as the emotions of a tweet, categorizing them as negative and positive emotions respectively. With this, they created ratios of each emotional tweet, and checked what percentage of all tweets contained that emotion (e.g., "hope"). After having obtained the ratios, they looked at the market returns on the following day (t + 1) with each respective emotion and calculated the correlations. The authors unveiled a negative correlation, meaning that an increase in positive emotions and the frequency of emotional words led to negative returns the following day. The authors concluded that this was because periods of economic uncertainty increased the frequency of emotional words, which was when the stock indices did badly.

A more recent study on the social platform we will analyze is done by Buz and de Melo (2021), which studied the popular subreddit "WallStreetBets". They collected data from January 1st, 2019, to April 4th, 2021, to perform their analysis. In contrast to the other authors, they not only tried to create buy and sell signals based on the sentiment of equities, but also created portfolios based on the popularity of individual stocks and considered whether the attention came before an increase in price or following a price increase. The authors created two portfolios, where the simple one was constructed by selecting the tickers which have consistently over the three-year period been the most mentioned stocks in the community. While the second portfolio dived into the submissions and counted transaction-related words related to an individual stock ticker. The buy and sell signal were then created by counting the frequency of each "buy" and "sell" related word, and whichever were the highest defined whether that particular day was regarded as a buy or sell signal. Combining this signal with the prices retrieved through Yahoo finance, they were able to create a hypothetical portfolio that invested based on the signals. They found that both portfolios outperformed the market over the short and long term. The buy signals for the second portfolio were quite successful beating the market overall, while also having an accuracy of 70% after three months. However, their sell signals were very unsuccessful, and further analysis, and potentially a larger dataset would be needed to draw conclusive conclusions.

Our analysis will be somewhat similar to the ones described above, but our analysis will be conducted on a more recent dataset, looking at a larger time horizon, and not only analyzing price development. We do not aim to measure the skill or competence of the participants in the social network, but rather utilize knowledge on social networking effects to see whether this can be traded upon and used to explain returns and trading volume.

3 Hypothesis and methodology

This thesis aims to further analyze the relationship between social media networking effects and stock market movements. Specifically, we will focus a lot on the forecasting ability data collected from social media platforms potentially present. Can social media networking effects influence certain financial market dynamics and can this be capitalized upon to generate returns? In this section we will describe our approach to attempt to shed light on these issues as well as presenting our hypotheses. This thesis aims to answer two research questions:
Can a sentiment analysis of forums/discussions on large social media platforms present opportunities to create profitable trading strategies and portfolios?

Can social media attention be used to predict returns and trading volume?

3.1 Sentiment analysis

In order to shed light on the above-mentioned research questions, a form of sentiment analysis will be deployed. Sentiment Analysis (SA) is a tool which utilizes natural language processing to analyze and extract desired information from data presented in text form. It enables researchers to analyze opinions and emotions towards a topic, language complexity as well as other subjective characteristics of the writer. This makes the tool very well suited for our purpose, as our data is largely in text form.

Many different sentiment analysis techniques exists, see figure 1 (Medhat et al. (2014), p 1095), and the area can generally be split into two different main approaches: The machine learning approach and the Lexicon-based approach. This thesis focuses on the dictionary-based approach within the lexicon-based subgroup. The machine learning approach was deemed unsuitable for this thesis, mainly because it requires a pre-labeled training dataset to learn from.

We did not have access to a labeled dataset suitable for our purposes, and it is not certain that training the program on one dataset makes it suitable for the actual dataset we conduct the analysis on (Borg and Boldt (2020)). In addition to this, research conducted by Islam and Zibran (2017), suggests that the use of a well-defined dictionary in the lexicon-based approach will lead to higher accuracy than many of the techniques under the machine learning approach. They studied the low accuracy of SA in software engineering texts and concluded that this inaccuracy was due to the lack of a tailored dictionary being used to study the particular domain (Islam and Zibran (2017), p. 478). This applies not only to software engineering, but any area where sentiment analysis may be applied.

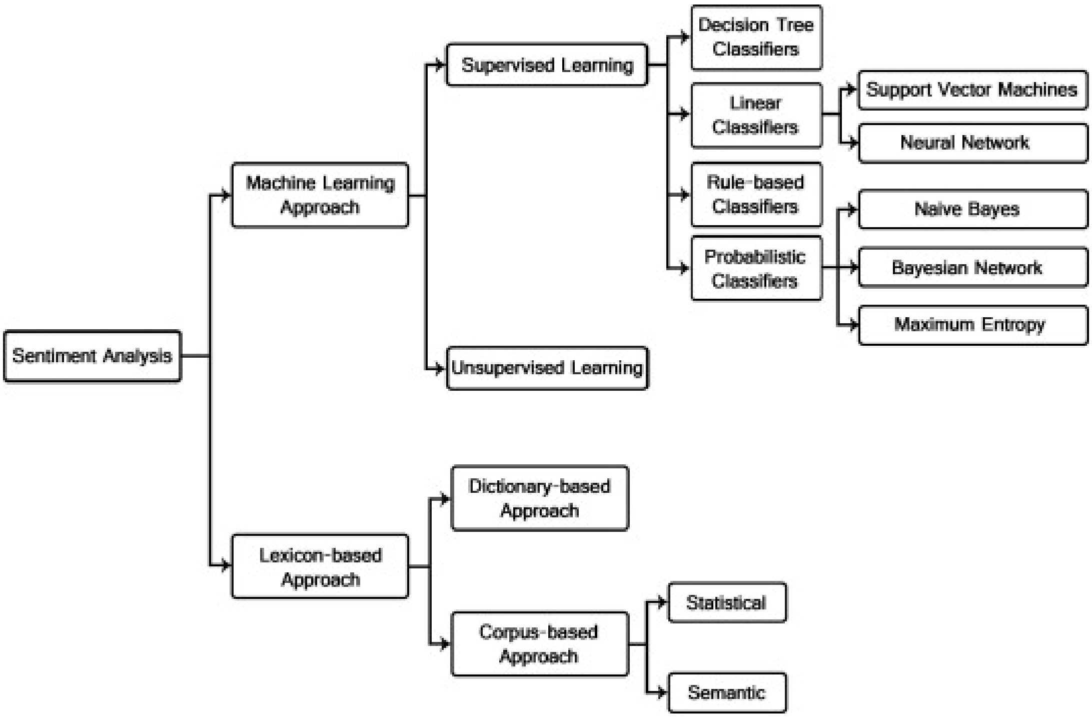

Figure 1: Different sentiment analysis techniques

3.1.1 Valence Aware Dictionary for Sentiment Reasoning (VADER)

There exist a lot of different models with various dictionaries to conduct sentiment analysis under the lexicon-based approach. As proven by Islam and Zibran (2017), using an effective lexicon tailored specifically for the purpose is crucial to obtain accurate results. The sentiment analysis in this thesis is done by using the VADER (Valence Aware Dictionary sEntiment Reasoning) sentiment model, which is a rule- and lexicon-based technique for sentiment analysis. VADER is especially suited for our analysis as the framework is developed to work well on social media content (Hutto and Gilbert (2014)). This is mainly because VADER can account for emoticons, slangs and conjunctions.

VADER utilizes a sentiment lexicon to determine the sentiment of a social media post or comment. The sentiment lexicon is essentially a list of words with a sentiment valence (intensity) score attached. The sentiment valence score of each word ranges from -4 to +4, with -4 indicating the most negative words, +4 indicating the most positive words and 0 being neutral. The valence scores allow the model not only to classify the general attitude of the group towards a particular company, but also quantify the strength of the sentiment. The framework outputs four different classification metrics on each post or comment: A negative, positive, neutral and a compound score. All of these scores range between -1 and +1, where -1 is most negative and +1 is most positive. The positive, negative and neutral metrics indicate what proportion of the words in the comment or post falls in the positive, neutral or negative category and the scores ranging between - 4 and + 4 are normalized to be between -1 and +1. The positive, neutral and negative metrics always sum up to one. The compound score is the metric of interest for our purpose, as it is the best stand-alone metric for gauging the overall sentiment. The score is computed by summing the intensity score of each word in the post or comment and then normalizing such that it falls between -1 and +1 (Hutto (2014)).

The framework also allows us to easily update the dictionary with additional words and phrases we would like the program to account for, such that it provides a more accurate measure of sentiment. As social media forums often use slang and abbreviations, being able to add to the existing dictionary is important. Updating the lexicon with relevant words and phrases is a time-consuming process as you have to manually look through and identify commonly used words and phrases that indicates a certain opinion and score the word based on how positive or negative it is. To make this process more efficient, we utilized an existing list created by Julian Klepatch for his own real time sentiment analysis of WallStreetBets (Klepatch (2021)). This list was available in his GitHub repository and contained a lot of words commonly used in discussions on WallStreetBets. These words are more commonly used in more recent times, and the earlier parts of the sample will contain less of these expressions. However, they will improve the accuracy of VADER overall. See appendix B for the full list of words.

3.2 Portfolio construction

The compound scores obtained from the sentiment analysis will be used to construct long/short sentiment portfolios. This portfolio construction is similar to how the Fama and French (2015) factor- portfolios are constructed. The long portion consists of stocks with the highest sentiment scores and the short portion consists of the stocks with the lowest sentiment scores, implying that a higher sentiment intensity score should lead to higher returns while a lower sentiment intensity score should lead to lower/negative returns. Specifically, the portfolio is constructed in the following way:

- We identify the most discussed stocks within a specific year, which makes up the investment universe for the following year. For example, the 20 most discussed stocks in 2018 are the ones we track the sentiment of through 2019.

  For the majority of the years in our sample period, the portfolio consists of 20 stocks. We did not use a higher number of companies as we had little data of the less popular companies, especially for the earlier parts of the sample. For example, having only 20-30 posts or comments regarding a company throughout a year is not sufficient for a good analysis, hence we elected to include a smaller number of stocks in our portfolio. For this reason, we also have fewer stocks included for the first few years of the sample. See appendix A for full list of included stocks per year.

- The dataset is filtered to extract all submission text and comments where these specific stocks are mentioned either by name or ticker. We are aware that some discussions where a certain company is mentioned could be unrelated to the stock. For example, someone may say ”search this up on google”, which has no relation to the google stock. For this reason, we exclude some company names, and search only for tickers for the most obvious instances. However, we will not be able to completely filter out all these cases, but seeing as we are collecting data solely from stock-related forums, it is reasonable to assume that the majority of the discussion is regarding stocks and company performance, and not whether a hat ordered from Amazon looks good. We believe our results are not affected by the few cases where a company is mentioned for other reasons.

- Sentiment scores are obtained for the posts and comments related to the specific companies by running VADER, which allows us to track the sentiment of all the discussions regarding the company throughout the year. The created_utc attribute[1] of the post/comment helps with this.

- We aggregate the sentiment scores by averaging to obtain monthly scores. Then the stocks are ranked from highest to lowest sentiment score within a specific month, and the portfolio for the next month is based on this. The long portfolio consists of the 10 companies with the highest sentiment scores and the short portfolio consists of the 10 companies with the lowest sentiment scores[2]. All stocks are equally weighted.

- These positions are held for a month, and the portfolio is then rebalanced and reconstructed at the start of a new month to reflect the new sentiment scores for the previous month (the month we just held the previous positions through).

- At the start of a new year, the investment universe is updated and the process is repeated. For example, if a stock was among the 20 most discussed in 2018, this was a part of the portfolio for 2019. If this stock was not among the most discussed through 2019, it is not included in the portfolio for 2020. If the company is still among the 20 most discussed, it is still included in the portfolio and the sentiment will be tracked throughout the year. This update is done at the start of every new year.

The portfolio construction is done in this way to ensure the strategy does not benefit from clairvoyance in terms of using information that is not yet available at the time of trading. The goal is to backtest a trading strategy that could be practically implemented. The portfolio could also be constructed to trade in a

[1] This attribute will be explained in section 4.1.3 below.

[2] As a robustness check, we removed the stock with the lowest sentiment score in the long bucket and the stock with the highest sentiment score in the short bucket (removing stock 10 and 11 when ranked from highest to lowest). This caused the difference in sentiment score for the bottom stock in the long bucket and the top stock in the short bucket to widen. Potentially increasing the return difference of the two buckets and enhancing the long-short portfolio performance. The changes in results were minimal.

stock as soon as the sentiment for that specific stock is identified, thus reducing the risk that we "miss out" on the effect of increased social media attention by trading only monthly. However, to reduce the implications of trading costs if implemented in practice, the portfolio is rebalanced on a monthly basis. The second part of the analysis will cover a more real time approach to the effect of social media on stock returns and trading volumes (see section 3.3).

3.2.1 Performance evaluation

In order to evaluate the portfolio performance, we compute traditional performance measures such as mean, standard deviation and Sharpe Ratio. In addition to this the portfolio returns are regressed on the Fama-French Five-Factor model (FF5-model) including the momentum factor. The FF5-factor model is one of the most notable asset pricing models in modern finance. It builds on the traditional CAPM- and Fama-French three-factor model (Fama and French (1993)) and aims to explain the excess returns of our portfolio by exposures to five factors. The regression looks the following:

$$\begin{aligned} R^e_{p,t} = & \beta_1 + \beta_2 \quad R^e_{M,t} \quad + \beta_3 \text{ SMB}_t + \beta_4 \text{ HML}_t \\ & + \beta_5 \text{ RMW}_t + \beta_6 \text{ CMA}_t + \beta_7 \text{ MOM}_t + \epsilon_t \end{aligned}$$

where $R^e_{p,t}$ is the excess return on the sentiment portfolio, $R^e_{M,t}$ is the excess return of the value weighted market portfolio, SMB (Small Minus Big) is the return on a portfolio of small stocks minus the return on a portfolio of large stocks (measured in market cap), HML (High Minus Low) is the return on a portfolio of high B/M-ratio (Book-to-Market) stocks minus the return on a portfolio of low B/M stocks, RMW (Robust Minus Weak) is the difference in returns of portfolios consisting of firms with robust and weak profitability, CMA (Conservative Minus Aggressive) is the difference in returns of portfolios consisting of firms with aggressive and conservative investment policies, MOM is the return on the momentum portfolio, which consists of stocks with high and low prior returns (12 month period). The ϵ denotes the residual of the regression. Seeing as the portfolio returns are on monthly basis, so will all the factors.

3.3 Stock regressions

In addition to creating long-short portfolios, we also want to further investigate whether increased social media attention of an individual stock influences some of the characteristics of the security, such as trading volume or returns. In order to assess this, we will be performing regressions of individual stock returns or trading volumes on a *number of mentions* - factor as well as a sentiment factor. Unlike for the portfolio construction, we disregard a lot of the implications of a practical implementation as the goal here is not to create a strategy but rather investigate the predictive capabilities of social media attention on individual stock characteristics.

As this is an analysis of individual companies, and it is important that the factor is constructed such that it has as most datapoints available, we have filtered out the five stocks that are consistently the most discussed across the entire sample-period. The stocks selected for the analysis are the following: Amazon (AMZN), Tesla (TSLA), Apple (AAPL), Meta (FB), and Microsoft (MSFT). All of these companies are large and are often covered in traditional media, so it is not surprising that these companies are the most discussed in the subreddits as well (See appendix D for graph of total number of comments and submissions for each of the five stocks).

The number of mentions factor is constructed by simply collecting the number of times the ticker or name of the company is mentioned throughout a specific 24-hour period. Every post and comment containing the specific company name or ticker is counted if it is published from the US Market close time (day t) to the US market close time the following day (day t+1). This is the number of mentions for this particular day (t+1). The reason for this is that any mention of the company after the US Markets close on day t can not be traded upon until the market opens the next day (t+1), hence can only affect the trading volume or price on the following day (t+1). If a post or comment is published while the market is open, this can be immediately traded upon and may have an effect on the same-day trading volume or returns, and is therefore included in the count for this particular day. The sentiment variable is the vector of sentiment scores obtained from the sentiment analysis for the specific stock within the same timeframe as the NoM-factor.

These factors will be combined in a regression with the Fama-French five-factor model as well as the momentum factor to determine whether the number of mentions and stock sentiment has an effect on stock returns or trading volume. It is also interesting to lag the number of mentions variable to see whether same day discussions or previous day discussions has the largest effect. The return-model we seek to estimate looks the following:

$$\begin{aligned} R^e_{c,t} = \beta_1 + \beta_2 \quad & R^e_{M,t} \quad + \beta_3 \quad \mathrm{SMB}_t \quad + \beta_4 \qquad \mathrm{HML}_t \\ & + \beta_5 \ \mathrm{RMW}_t + \beta_6 \quad \mathrm{CMA}_t \quad + \beta_7 \qquad \mathrm{MOM}_t \\ & + \beta_8 \ \mathrm{NoM}_{c,t} + \beta_9 \ \mathrm{NoM}_{c,t-1} + \beta_{10} \ \mathrm{Sentiment}_{c,t} + \epsilon_t \end{aligned}$$

where $R^e_{c,t}$ is the excess return of the particular company, and the new variables for the specific companies are NoM_t which is the number of mentions factor, NoM_{t-1} is the lagged number of mentions factor and Sentiment is the sentiment factor. The regressions for all the different companies will have the same structure. For these regressions, the factor dataset is in daily format.

The trading-volume model we seek to estimate looks the following:

$$Volume_{c,t} = \beta_1 + \beta_2 \ \mathrm{NoM}_t + \beta_3 \ \mathrm{NoM}_{t-1} + \beta_4 \ \mathrm{Sentiment}_t + \epsilon_t$$

where $Volume_{c,t}$ is the daily trading volume of the particular company, NoM, NoM_{t-1} and Sentiment is as described above. The regressions for all the different companies will have the same structure. For these regressions, the factor dataset is in daily format. Note that the FF5-factors are not included in the trading volume regression, as these factors are constructed to explain returns and not trading volume.

These regressions could also potentially give an indication of whether the discussion in the forums drives returns and trading volume, or returns and trading volume drives the discussion in the forums.

3.4 Hypothesis development

Based on the above discussions we present three hypotheses. The first part of the analysis covers a practical portfolio approach to investigate whether portfolio construction based solely on Reddit-sentiment is profitable. The first hypothesis is then defined in the following way:

Hypothesis 1

H_0: *Constructing a long-short portfolio based on Reddit-sentiment is not profitable*

H_1: *Constructing a long-short portfolio based on Reddit-sentiment is profitable*

The second part of the analysis covers an empirical investigation of whether the activity (measured in number of mentions) surrounding an individual company on the two Reddit-forums has an effect on trading volume or security price. Based on this, we present two hypotheses:

Hypothesis 2

H_0: *Increased social media attention does not increase trading volume*

H_1: *Increased social media attention increases trading volume*

Hypothesis 3

H_0: *Increased social media attention does not affect returns*

H_1: *Increased social media attention does affect returns*

4 Data

This section explains the methods and tools used to collect the various types of data utilized in the analysis. We have used two different types of datasets to conduct our analysis: A dataset consisting of social media posts and comments as well as data consisting of financial data for selected stocks and factors. The below sub-sections will describe the characteristics of the different datasets and the collection-process in detail.

4.1 Reddit dataset

4.1.1 Subreddit selection

The dataset used for the sentiment analysis comes solely from the social media platform Reddit. Reddit is currently the largest online forum with more than 50 million daily users and over 13 billion posts and comments submitted since the forum was started in 2005. The site is built up of a collection of different sub-communities/forums called ”subreddits”, where registered users can join in to post content and contribute to discussions on specific topics such as politics, sports or finance. The different subreddits are referred to by writing ”r/” followed by the name of the subreddit. For example ”r/politics” will be a subreddit where users discuss politics.

For our research, the finance-related subreddits will be the ones of relevance and we have specifically chosen to gather our data from the subreddits named r/stocks and r/investing. We elected to collect our data from these subreddits as they are both among the largest finance-subreddits with a combined 6.1 million subscribed users (as of May 2022), not accounting for any users being a member of both subreddits. Both subreddits were started in 2008, and have sufficient data for the earlier part of our sample period.

Compared to WallStreetBets (WSB), the forums are much smaller in size. If measured in number of subscribers, r/investing and r/stocks combined amount to about half of r/WallStreetBets. We did however not elect to collect from WSB for a number of reasons:

- WSB was started in 2012 and had very limited activity compared to r/investing (which was the largest of the three in the beginning) for the first few years. It was not until the beginning of 2020 that the forum surpassed r/investing in number of subscribers. This means that r/WallStreetBets has less available data for the earlier part of our sample period. Appendix G, H and I shows the development in number of subscribers for the three different subreddits. r/stocks has also seen a very large growth over the last few years, and is currently the largest of the two subreddits we are collecting our data from.

- WSB has for the last few years seen a very large increase in activity (See appendix I for graph). Due to this large number of comments and posts over the last 1-3 years, we would only be able to collect data from a smaller time period due to the collection process becoming extremely time-consuming. This will be elaborated on in subchapter 4.1.2 below. As our goal is to examine the effect of social media on stock market movements over a longer time period, spending a lot of time collecting the more recent data and having less data from a few years back would not be the best approach.

- r/WallStreetBets has developed a very specific way of communicating, using a lot of specific terms not used in more ordinary communication. Sarcasm is also very common in WSB posts and comments. Both of these would make sentiment analysis a lot harder to conduct, as the software is bad at identifying when the author is being sarcastic and does not recognize a lot of the terms and expressions used. This would lead to a higher risk of falsely measuring the sentiment of a specific post or comment.

4.1.2 Data collection

When collecting the data, there were two different ways we could approach this. Either by utilizing Reddit's own API directly or by accessing the Pushshift Reddit Dataset (Baumgartner et al. (2020)). As Reddit's own API does not allow for data-collection from a specific time period in the past, using this for the data collection would not be possible and therefore accessing the Pushshift dataset through their web API is the optimal choice. Do however note that if one wants to potentially trade on social media sentiment in real time, using Reddit's own API would be the more efficient and better choice.

The Pushshift Reddit dataset is a project started in 2015, and is a database consisting of all submissions and comments from all existing subreddits spanning from June 2005 to today. The data is collected the second the post or comment is published. The database is updated monthly with the new data (Baumgartner et al. (2020)). By writing a Python-script accessing the web API we were able to collect the data we needed. This is also what Buz and de Melo (2021) did in their research article, where they aimed to evaluate the quality of investment advice on WallStreetBets by developing a trading strategy based solely on this.

Pushshift's API is open source and available to use by anyone. As a result of this the number of requests the API can handle in a specific timeframe is limited, making the data collection process highly time-consuming. Because of this, the data was collected in pieces (mostly year by year) by setting the start and stop parameters of the script and letting the program run until the data for this period was collected and downloaded. Then repeating the process for new start and stop parameters until data for the entire time-period was collected.

As expected, the most recent years had the most data, and downloading these parts of the dataset could take up to 7-8 hours. This was done firstly for the submissions of the posts and then for the comments. Naturally, there were more comments than submissions, and collecting all the comments was the most time-consuming part of the collection process.

Other research conducted on similar topics have treated submissions and comments separately and used the id-attribute to link comments to a specific submission. As our goal is to measure the overall sentiment of the investing communities towards specific stocks, whether a comment is related to a specific post is of no relevance to us and we have conducted sentiment analysis on the comments and submissions together. Buz and de Melo (2021) also chose to not include comments in their analysis and focused solely on submissions. Their reasoning for this was that submissions represent the main topic that a user encounters when browsing the WallStreetBets subreddit and that comments typically are only short replies to the statements made in the submission. They went on saying that the comments are very heterogeneous in terms of their informational value and therefore chose to exclude them (Buz and de Melo (2021)).

Our analysis incorporates comments as well as submissions, as we believe there to be quite a bit of informational value to be extracted from the comments. Both r/stocks and r/investing have daily general discussion posts, where users discuss companies, macroeconomic outlook and other general finance. A new post is created by the automoderator of the subreddit every day, and people discuss and reply to each other in the comment-section. These posts are among the posts with most activity every day and will therefore be among the most insightful when aiming to measure the overall sentiment of the community over time. In addition to this r/stocks has a discussion post updated quarterly named ”Rate My Portfolio” where users present their portfolios and receive advice regarding their allocations from other users. This is another example of a post where people display expressions towards certain stocks in the comments rather than in the submission itself and we believe some users will make changes to their portfolio based on others advice and that this is another example of comments providing insightful information.

4.1.3 Characteristics and cleaning

The collection resulted in multiple csv-files for both submissions and comments as we collected the data mostly year by year. The uncleaned datasets contains the text of the comment or submission (the attributes are called *body* for a comment and *selftext* for a submission) as well as different metadata such as time of posting, the username of the author, id-number, score, name of the subreddit it was posted to and the title of the post. The related metadata vary some whether we are collecting submissions or comments. E.g., if we collect submissions there will also be an attribute telling us how many comments the post has at the time of collection.

Most of this information is of no relevance to our analysis, and we have excluded most of the variables in our cleaned dataset. The only attributes of interest is the selftext or body of the comment/submission and the attribute called ”created_utc”. The ”created_utc” attribute represents the time of posting in Unix time. Unix time is a way of tracking time in number of seconds since Unix epoch, which is January 1st 1970. Essentially the ”created_utc” attribute displays the time of posting in number of seconds since 1. January 1970, which can easily be converted into a more understandable date format. Using this attribute gives us a very specific time of posting which allows for accurate tracking of sentiment over time.

When collecting historical data from an online forum, it is inevitable that some posts or comments have been removed or deleted from the forum. This may be because the user who wrote the post no longer wants it to be available and therefore deletes it, or because it violated some of the community guidelines and was removed because of that. In the Pushshift Reddit dataset, these posts and comments are displayed as *[deleted]* in the body or the selftext. These posts or comments were eliminated from the dataset as they add no value to the analysis. The database also disregards pictures or videos, and stores only the text accompanying the image or video, and this type of content is therefore not included in our dataset.

By running our python-script we have collected all submissions and comments posted on r/investing and r/stocks from January 1, 2014 to December 31, 2021. The finalized dataset consists in total of 6,713,525 posts and comments.

4.1.4 Weaknesses

A drawback of collecting data from social media platforms such as Reddit, is that the majority of activity has occurred over the last few years alone. Most of the subreddits on the site has accumulated the majority of it's userbase over the last 3 years alone. This implies that there will be less available data for the earlier part of our sample period, and we are less likely to identify connections between forum-activity and individual stock performance. If we do so, we will be careful with drawing conclusions as we are aware that the results could be due to other factors. In order to mitigate this weakness, data is collected from two subreddits where one was the largest for the earlier years. Our initial plan was to collect only from r/stocks. However, seeing as this subreddit has seen most of its growth since 2020, we quickly discovered that it existed small amounts of data from 2015 - 2017 primarily. Thus, we elected to include r/investing as well. The inclusion of r/investing gave access to more data for the earlier years which will make the conclusions stronger. Do however note that there is still substantially less data for the earlier years of our sample period than for the later years.

It is also worth noting that the dataset from the Pushshift API could deviate some from the original Reddit data. This is because the database is updated with submissions and comments as soon as they are posted to Reddit and some statistics for the posts may be incorrect. Most notably, the "score" of a post or comment may not be accurate. The "score" is a statistic reflecting the result from Reddit's ranking system where other users upvote or downvote a given post or comment based on whether they like it or not. The score is then computed as the number of upvotes minus the number of downvotes. Seeing as we don't use the score in our analysis it is not important to us whether this is accurate or not. In addition to this, the Pushshift database has had days where it has been offline, resulting in the service not being able to consistently collect data for the entire time period.

In order to validate that the python script collects accurate data, we randomly selected about 50 points in our dataset and manually looked them up on Reddit to control that the time of posting and the content of the post was the same as displayed in the dataset collected from Pushshift. All of the relevant information was accurate for the ones we controlled, and hence we conclude that the dataset contains accurate information for our purposes.

We are aware of these potential weaknesses in our dataset, but given that the service has only been offline for a few days over our sample period of 7 years, as well as our sample-test being accurate, we believe this will have no effect on our conclusions.

4.2 Financial data

The Reddit dataset was used in combination with financial data for the most mentioned stocks/tickers within a specific year. This data was needed to measure the historical performance of the long-short sentiment portfolio as well as providing information on trading volume. The financial dataset was collected from Yahoo! Finance and contained data on the individual stocks opening price, closing price, trading volume, highest price, lowest price and adjusted close price on a daily basis.

Data on the Fama-French factors was collected from the Kenneth French database, which again collects it's financial data from CRSP. Monthly as well as daily data on the market portfolio, the SMB, HML, RMW and CMA factors as well as the momentum portfolio was retrieved from this database.

5 Results

5.1 Long-Short sentiment portfolio

We create a long-short portfolio consisting of the 20 most discussed companies throughout a year and sort them in the long or short basket based on the sentiment valence scores. This portfolio is rebalanced monthly to reflect the updated sentiment scores of the companies and the included stocks are updated at the start of every year. An overview of the stocks included in the portfolio per year can be found in appendix A.

5.1.1 Monthly rebalanced portfolio

We have measured the performance of the portfolio on the entire sample-period spanning from 2015 to 2021 and collected performance statistics for the long basket, short basket and the combined long-short portfolio. The performance statistics of the portfolios is summarized in table 1 below:

Table 1: Summary statistics of the monthly sentiment-portfolios

The table displays the mean excess returns, the standard deviation and the Sharpe Ratio of the long, short and combined sentiment-portfolios over the sample period of 7 years. There is a total of 84 observations (months) included in the sample. All statistics are annualized. It is also worth noting that the statistics for the short portfolio is displayed as if it was a long-portfolio. The statistics are not taking into account that this portfolio's returns will be multiplied by -1.

	Mean excess return (μ_e)	Standard deviation (σ)	Sharpe ratio (SR)
Long	26.57%	22.09%	1.20
Short	22.26%	27.14%	0.82
Long-Short	4.31%	17.78%	0.24

From the table above we see that the long portfolio performs the best if measured in annual excess returns as well as Sharpe ratio for the sample period. The short portfolio performed only slightly worse than the long portfolio when it comes to excess returns, but it is also worth noting that the standard deviation of the short portfolio is higher, hence providing a worse SR. The combination portfolio performed the worst, driven by the high returns of the short portfolio. The long-short portfolio was however the least volatile portfolio of the three, but that did not weight up for the low returns, and the combination portfolio has the lowest SR of the three.

Both the long portfolio and the short portfolio outperforms the S&P500 - index, which has generated an average annualized return of 14.43% from 2015 to 2021. The long portfolio also outperforms the NASDAQ100 - index which have generated an average annual return of 22.42% for the sample period. The combination portfolio outperforms neither of the two indices.

The poor performance of the combination portfolio is due to the good performance of the short portfolio, which was surprising to us. However, when we looked into why the short portfolio performs so well, we realized this may partially be because of the way the portfolio is constructed. As described in section 3.2, we aggregate the sentiment scores by averaging all scores within a specific month to obtain monthly sentiment scores. One of the weaknesses of averaging, is that it smooths out the result. In our case the negative comments are essentially eliminated as we average to get monthly scores because the majority of the discussion is positive. This leads to the short portfolio not consisting of the stocks with the most negative sentiment, but the stocks with the lowest positive sentiment.

See appendix E for a graph of the cumulative returns of these portfolios.

5.1.2 Daily rebalanced portfolio

In order to investigate the effect of this averaging further, we construct a new long-short portfolio and evaluates its performance from 2018 to 2021.

The new portfolio consists only of the 5 stocks included in the individual stock regression. The stocks are equally weighted and this portfolio is rebalanced on a daily basis. Very few of the daily observations are negative, hence, we changed the rule for when we go long or short. Instead of simply going long the stocks with the highest sentiment and short the stocks with the lowest sentiment, the position in the specific stock is determined by whether the sentiment for that day is higher or lower than the average sentiment level. If the sentiment is above the average, we go long the particular company and opposite if the sentiment is below the average. The average sentiment level is computed over the entire sample, so we are disregarding the clairvoyance-constraint we imposed on the original portfolio. Do note that this could have been implemented differently by for example using the previous 12-month average or similar if it were to be practically implemented. The portfolio we constructed goes long and short in a specific stock approximately 50% of the time. The performance of the new long, short and long-short portfolios is displayed in the table below:

Table 2: Summary statistics of the daily sentiment-portfolios

The table displays the mean excess returns, the standard deviation and the Sharpe Ratio of the long, short and combined daily rebalanced sentiment-portfolios over the sample period of 4 years. All statistics are annualized. It is also worth noting that the statistics for the short portfolio is displayed as if it was a long-portfolio. The statistics are not taking into account that this portfolio's returns will be multiplied by -1.

	Mean excess return	Standard deviation	Sharpe ratio
	(μ_e)	(σ)	(SR)
Long	54.08%	19.20%	2.82
Short	10.67%	25.17%	0.42
Long-Short	43.40%	25.51%	1.70

Based on the above results, we see that the long portfolio performs very well while the short portfolio also generates some positive return (which is bad for the long-short portfolio). The combination portfolio performs much better for this daily rebalanced portfolio than for the monthly rebalanced portfolio. Do note that these portfolios are constructed differently and are therefore not directly comparable, but some of the findings are interesting.

The trend in this portfolio is similar to the one we saw for the monthly, you would be better off by simply going long. However, in this case the short portfolio performed not nearly as well as the long portfolio. Compared to the monthly rebalancing, the difference between the long and short portfolios is much larger leading to the long-short portfolio performing substantially better. The companies included in this portfolio are among the ones that since 2018 have generated very high returns and we find it quite impressive that the short bucket did not generate higher returns, leading to a similar decrease in the overall long-short portfolio performance as we saw in the monthly rebalanced portfolio.

By not directly comparing the returns of the different portfolios, but rather the relationship between the returns of the long- and the short-baskets in the monthly and daily portfolio it appears that the duration of the sentiment is much shorter than we expected and hence more frequent trading will be the most profitable. One drawback of this will be that more frequent trading will increase the trading costs, which we have not taken into account for this analysis.

See appendix F for a graph of the cumulative returns of these portfolios.

5.1.3 Portfolio construction summarized

In conclusion we see that simply investing in the long portfolio would be the better option over combining the long and the short portfolios. These findings are similar to those Buz and de Melo (2021) found in their article. Out of all the portfolios the authors created based on r/WallStreetBets posts, the one constructed by their sell-signals were highly unsuccessful while the portfolio constructed on the buy-signals outperformed the benchmark. The authors stated that further research would be required to draw any conclusive conclusions.

Our analysis, which is conducted on a dataset collected from different subreddits and covering a larger time-period, show the same trend. The long portfolio performs well, while the short portfolio performs worse, especially when averaging to obtain a monthly rebalanced sentiment score. We do not claim that this is conclusive evidence, but it is possibly indicating that the members of social media forums are better at identifying profitable long-opportunities, than they are at identifying profitable short-opportunities. A reason for this could be the short-constraint most retail investors are affected by. Most of them can't place short positions and will because of this mostly discuss companies they believe in, rather than companies they believe will perform bad. This is also supported by the low number of observations with negative sentiment scores in our dataset.

5.1.4 Monthly rebalanced portfolio regressions

We have performed regressions of the different monthly rebalanced portfolios' excess returns on the Fama and French (2015) five-factor model including the momentum factor in order to evaluate the portfolio performance in detail. The results of the regressions are displayed in table 3 below:

Table 3: Monthly sentiment-portfolio regression

The table shows the coefficients, t-statistics (in parentheses) and adjusted R^2 of the regressions of the excess returns of the long, short and long-short portfolios for the monthly rebalanced sentiment-portfolios. The dependent variable of all regressions are the excess returns of the portfolio. The independent variables are the Fama-French five factors including the momentum factor. Do note that the metrics for the short portfolio are presented such that this would be the exposure to the different factors if a long position is placed in the portfolio. Multiply all coefficients for this portfolio by -1 to obtain short exposures.

	Monthly Long (1)	Monthly Short (2)	Monthly Long-Short (3)
Constant	0.0075* (1.93)	0.0037 (0.79)	0.0038 (0.71)
Mkt-Rf	1.29*** (12.29)	1.41*** (11.06)	- 0.12 (- 0.84)
SMB	- 0.32* (- 1.91)	0.31 (1.52)	- 0.63*** (- 2.76)
HML	0.02 (0.15)	- 0.38** (- 2.15)	0.40** (2.03)
RMW	- 0.40* (-1.84)	- 0.70** (- 2.63)	0.30 (1.00)
CMA	- 0.42* (- 1.69)	0.13 (0.45)	- 0.55 (- 1.64)
Momentum	0.16 (1.30)	- 0.10 (- 0.68)	0.26 (1.56)
Adj. R^2	0.72	0.73	0.21

*p<0.1, **p<0.05, ***p<0.01

The output from the regression of the monthly rebalanced sentiment-portfolio on the Fama-French five factor model including the momentum factor shows that for the long portfolio (model 1 in table 3), the Mkt-Rf factor is statistically significant at $\alpha = 1\%$. This is also the case for the short portfolio (model 2 in table 3), while for the long-short portfolio (model 3 in table 3) this coefficient is not statistically significant. This is to be expected as a long-short portfolio consisting of stocks with similar market exposure aims to be approximately market neutral by having one part of the portfolio hedge the other.

When it comes to the SMB - factor (Small-Minus-Big), this is statistically significant at $\alpha = 10\%$ for the long portfolio, and at $\alpha = 1\%$ for the long-short portfolio. The factor is not statistically significant for the short portfolio. For both the long and the long-short portfolio the coefficient is negative, indicating that these two portfolios load on companies with larger market caps. This result is not surprising as the companies included in the portfolios are mostly large, well-established companies like Apple and Amazon (see appendix A for full list of included companies). The SMB-coefficient is more negative for the long-short portfolio, which indicates that by combining the long-portfolio with the short portfolio, the exposure to the underlying risk factors of large companies is even higher. This indicates that the short portfolio also consists of smaller companies, and when these are shorted, the exposure to the underlying risk factor associated with larger companies is amplified. The coefficient on the SMB-factor also shows that a 1% increase in the returns of the SMB-portfolio in a specific month should lead to a 0.32% decrease in the long-portfolio and a 0.63% decrease in the long-short portfolio for the same month.

The HML - factor (High-Minus-Low) is statistically significant for both the short portfolio and the long-short portfolio at $\alpha = 5\%$. The coefficient is negative for the short portfolio and positive for the long-short portfolio. This means that the short portfolio overweights low Book-to-Market (B/M) stocks and the exposure of the long-short portfolio is higher to the underlying risk factor of high B/M-stocks, which makes sense as the exposure of the long-short portfolio should be opposite of the short portfolio seeing as you are placing a short position (unless the long portfolio has a statistically significant HML factor with a positive coefficient larger than the absolute value of the negative coefficient of the short portfolio).

The RMW-factor (Robust-Minus-Weak) is statistically significant at $\alpha = 10\%$ for the long portfolio and at $\alpha = 5\%$ for the short portfolio. The factor is not statistically significant for the long-short portfolio. The coefficient is negative for both the long and the short portfolio, indicating that the portfolios load on companies with weak profitability. Going long and short the respective portfolios will then lead to a hedge, rendering the RMW-factor statistically insignificant for the combination portfolio.

The CMA-factor (Conservative-Minus-Aggressive) is only statistically significant for the long portfolio at $\alpha = 10\%$. The coefficient is negative, indicating that the portfolio overweights companies with low investment policies (conservative companies), exposing itself to the underlying risk factors related to this. This is however only at the 10% significance level.

The Momentum-factor is not statistically significant for any of the three portfolios, indicating that the portfolios do not load particularly on the companies with the highest performance over the last 12 months. We found this to be quite surprising, as we expected the subreddits to discuss stocks which have performed well in the recent past, hence causing a portfolio constructed solely on this discussion to load on the momentum-factor. However, the momentum portfolio changes a lot, leading to a lot of trading if you aim to track this factor. When looking into the development of the sentiment-portfolios and the included companies over the years we see that a lot of the included stocks are part of the portfolio for multiple years. This may be part of the reason for why the sentiment-portfolios do not load on this factor. As the momentum portfolio changes frequently while the included stocks in the sentiment-portfolios are updated on an annual basis, it is not very likely that the stocks included in our portfolio will be a part of the momentum portfolio for a very long time, making the exposure to this factor statistically insignificant.

The intercept of the long portfolio is statistically significant at $\alpha = 10\%$, while for the other two portfolios the intercept is statistically insignificant, indicating that the long portfolio is the only portfolio which seems to collect a risk premium not captured by the model. This means that the portfolio is the only one obtaining some form of outperformance relative to what is expected by the FF5+momentum model, indicating that you would be gettin a higher return on your money if you invested in the monthly rebalanced long portfolio rather than if you invested in all the different factor portfolios with weights equal to the respective beta-coefficients (the factor-benchmark portfolio). Another thing to consider is that investing in the factor portfolios in this case would require short positions, as the coefficients on some of the factors are negative. Simply going long the monthly rebalanced sentiment portfolio will not only yield a higher return but there is also no need for shorting, making this strategy suitable for short-constrained investors as well.

We also evaluate the models by their *adjusted* R^2, yo measure how good a particular model fits our dataset. We use the *adj.* R^2 and not the R^2 because the latter metric usually increases when adding more independent variables to the model, falsely indicating a better fit. Therefore, the *adj.* R^2 provides a better measure of the actual fit of the model. The *adj.* R^2 of the long portfolio and the short portfolio are similar at 0.72 and 0.73, while the combination portfolio has a somewhat lower *adj.* R^2 at 0.21. This shows that for the long and short portfolios, approximately 70% of the variation in excess returns are explained by the included factors. For the long-short portfolio, only 21% is explained by the factors. Meaning that the Fama-French five factor model including the momentum factor fits better for the long and the short portfolio than for the combination portfolio.

5.1.5 Daily rebalanced portfolio regressions

We have also performed the same regressions of the different daily rebalanced portfolios in order to evaluate the portfolio performance in detail. The results of the regressions are displayed in table 4 below:

Table 4: Daily sentiment-portfolio regression

The table shows the coefficients, t-statistics (in parentheses), and adjusted R^2 of the regressions of the excess returns of the long, short and long-short portfolios for the daily rebalanced sentiment-portfolios. The dependent variable of all regressions are the excess returns of the portfolio. The independent variables are the Fama-French five factors including the Momentum factor. Do note that the metrics for the short portfolio are presented such that this would be the exposure to the different factors if a long position is placed in the portfolio. Multiply all coefficients for this portfolio by -1 to obtain short exposures.

	Daily Long	Daily Short	Daily Long-Short
	(1)	(2)	(3)
Constant	0.0011*** (4.38)	- 0.0003 (- 1.36)	0.0014*** (3.51)
Mkt-Rf	0.43*** (22.71)	0.73*** (36.79)	- 0.30*** (- 9.63)
SMB	0.08** (2.03)	- 0.11*** (- 2.77)	0.18*** (2.98)
HML	- 0.27*** (- 6.88)	- 0.17*** (- 4.04)	- 0.10 (- 1.59)
RMW	0.058 (1.05)	0.18*** (3.03)	- 0.12 (- 1.29)
CMA	- 0.10 (- 1.44)	- 0.37*** (- 4.83)	0.26** (2.19)
Momentum	0.047* (1.77)	0.055** (1.99)	- 0.0086 (- 0.20)
Adj. R^2	0.43	0.63	0.11

*p<0.1, **p<0.05, ***p<0.01

When it comes to the daily rebalanced sentiment portfolio the results are somewhat different from those we saw for the monthly rebalanced portfolios. Do however note that the daily portfolio is constructed differently and consists of only five stocks. Hence the results are not directly comparable, but yet again some of the findings are interesting.

For the daily portfolios, we see that the intercept is statistically significant at $\alpha = 1\%$ for the long portfolio and the long-short portfolio while it is insignificant for the short portfolio. It appears that for the daily portfolio, the long-short is also able to deliver positive alpha which for the monthly was not the case. The alpha for the long-short portfolio is slightly higher than for the long-only portfolio, indicating that investing in this portfolio will yield a slightly higher abnormal return over the returns of the factor benchmark than the long portfolio. Do however note that the long-short portfolio requires shorting, which makes the long portfolio the better option for the short-constrained investor once again. The alpha of the long portfolio is also not a lot smaller, so a potentially risk-averse investor who does not like the additional risks related to placing short positions can invest in the long-only portfolio and does not have to forfeit a lot of abnormal returns by doing so.

The Mkt-rf factor is statistically significant at $\alpha = 1\%$ for all the portfolios. The long-short is not market neutral for the daily portfolio, which it was for the monthly. This is probably due to how the portfolio is constructed, specifically due to it not always being long and short at the same time, but it can be long/short the entire portfolio at the same time depending on whether the sentiment is above/below the average level. If the sentiment intensity scores are above average for all five stocks, the portfolio will be long all five stocks that particular day.

The SMB - factor is statistically significant for all the portfolios, and the coefficient is positive for the long and long-short portfolio while it is negative for the short portfolio. As the included stocks are solely large companies, we find it peculiar that the coefficient is positive for the long portfolio. However, given that the short portfolio loads more on the underlying risk factors related to larger companies, going short this portfolio indirectly provides exposure to the underlying risks of smaller companies, hence the long-short portfolio has a positive SMB-coefficient.

The HML - factor is statistically significant for both the long and the short portfolios, but is statistically insignificant for the long-short. This indicates that combining the portfolios creates a hedge aganst the underlying risk factor.

The RMW-factor is statistically significant only for the short portfolio, and the coefficient is positive in this case. Indicating that this portfolio has exposure to companies with robust profitability, which is not surprising considering the included companies.

We find it interesting that for the daily rebalanced portfolio, the momentum factor is statistically significant at $\alpha = 10\%$ for the long portfolio and at $\alpha = 5\%$ for the short portfolio, while it was insignificant for all portfolios in the monthly rebalanced case. This indicates that the daily rebalanced portfolio is exposed to this factor more than the monthly rebalanced portfolio, even though these five companies are among those included the most in the monthly rebalanced portfolio. This indicates that a lot of the other stocks we include in the monthly rebalanced portfolio reduces the exposure to this factor.

The $Adj.R^2$ are lower for all three regressions if compared to the monthly rebalanced. For the daily rebalanced long, short, and long-short portfolios, only 43%, 63%, and 11% of the variations in excess returns are explained by the model. The Fama-French five factor model including the momentum factor fits worse for the daily rebalanced portfolio than for the monthly rebalanced.

5.2 Individual stock regressions

We perform regressions of the excess returns and trading volume of the 5 individual stocks on the number-of-mentions factor (NoM_t), lagged number-of-mentions factor (NoM_{t-1}) and the sentiment factor. We control for the Fama-French five factors as well as the returns on the momentum portfolio in the return regressions. These factors are not included in the volume regressions as they are constructed to explain returns and not trading volume. In the following sub-sections we will present the results and discuss our findings.

5.2.1 Amazon (AMZN)

The return-regression for Amazon (AMZN) (Model 1 in table 5) shows that the coefficients on the number-of-mentions factor as well as the lagged number-of-

Table 5: Amazon (AMZN) regression results

The table shows the coefficients, t-statistics and adjusted R^2 of the regressions of daily historical returns and trading volume from 2018 to 2021 for Amazon (AMZN). As described in section 3.3, the NoM-factor is constructed by collecting the number of times the AMZN ticker or company name was mentioned in posts or comments on the two subreddits within a pre-specified 24-hour window. The sentiment factor is a vector of daily sentiment scores estimated from all the posts and comments mentioning the stock.

	Return (1)	Trading volume (2)
Constant	- 0.00 (-0.005)	87,257,490.00*** (19.28)
NoM_t	- 0.00 (- 0.48)	156,086.00*** (5.00)
NoM_{t-1}	- 0.00 (- 1.01)	- 32,277.00 (- 1.04)
Sentiment	0.00 (1.07)	-15,460,980.00 (- 1.65)
Mkt-Rf	0.93*** (29.92)	
SMB	- 0.13** (- 2.14)	
HML	- 0.54*** (- 8.44)	
RMW	0.40*** (4.45)	
CMA	- 0.91*** (- 7.66)	
Momentum	0.04 (0.97)	
Adj. R^2	0.60	0.031

*p<0.1, **p<0.05, ***p<0.01

mentions and sentiment factor are insignificant at the 10% level. This indicates that an increase in discussion of Amazon in the two forums does not have a significant effect on the returns for this particular stock. The adjusted R^2 of this regression is 0.60, meaning that the model explains 60% of the variation in returns for Amazon. The intercept is also statistically insignificant.

When it comes to the trading volume regression (Model 2 in table 5) the results are quite different from those in the return regression. The NoM_t factor is statistically significant at $\alpha = 1\%$ indicating that an increased number of mentions in the subreddits has a significant effect on the trading volume for Amazon stock specifically. This contradicts the results from the return regression to a certain extent as trading volume is highly correlated with price movements. However, this may be due to the high daily trading volume of Amazon stock and that the effect the forums appear to have on trading volume is not large enough to affect prices and returns. The results are nonetheless interesting as it seems medium-sized social media forums have an effect on trading volume and that even larger forums could potentially drive the price through the trading volume. It is also reasonable to assume that the effect will be more prevalent with stocks not traded as often.

The beta-coefficient on the NoM_t-factor is estimated by the model to be 156,086, which means that one comment or post mentioning Amazon in the two subreddits leads to an increase in trading volume by 156,086 shares that same day. The average trading volume for amazon-stock in our sample is 87,257,490, and the increase of 156,086 amounts to approximately 0.18% of the average daily trading volume.

If we lag the NoM - factor the coefficient is not statistically significant even at the 10% level, and it appears the number of mentions at closing time at time t does not affect the trading volume on the following day (t+1), hence the effect appears to only exist on same-day trading volume. We found this to be a bit surprising, as we expected the effect to be weaker when lagged, but not insignificant. We also suspect that the way we collect mentions, specifically that we include mentions published while the markets are open may provide significant results for the NoM_t-variable in the regression even if the causality is the other way around, meaning that trading volume and stock activity may be the driver of discussions in the online forums rather than the opposite. We will be investigating this further in subchapter 5.2.6.

For the trading volume regression, the intercept is statistically significant at the 1% level, indicating that there are several factors explaining the trading volume which are not included in the model. This is to be expected as the objective of the regression is not to fully explain the factors determining the trading volume of amazon stock, but rather to investigate whether the number of times the particular stock is mentioned in posts or comments has an effect on the trading volume. This is backed by the low $Adj.R^2$ of the model, which shows that the model explains only 3.1% of Amazon's trading volume. Specifically, the intercept shows that the daily trading volume of Amazon stock is approximately 87 million shares higher than the model predicts without accounting for the intercept (done by taking the number of mentions today and multiplying with the NoM$_t$-coefficient (156,086) and the number of mentions yesterday multiplied by the NoM$_{t-1}$-coefficient (-32,277). This number will be approx. 87 million lower than actual trading volume for that day).

The sentiment factor is not statistically significant for either of the two regressions.

5.2.2 Apple (AAPL)

The results from the Apple (AAPL) regressions are similar to those of Amazon. All of our added variables are not statistically significant for the return regression. For the volume regression, the NoM$_t$ and the sentiment variable is statistically significant at $\alpha = 1\%$ while NoM$_{t-1}$ is not statistically significant. It is interesting that the sentiment-variable is statistically significant, indicating a relationship between subreddit sentiment intensity and AAPL trading volume. The coefficient is however negative, meaning that a one unit increase in sentiment score negatively affects the same day trading volume. This could indicate that the members of the forums trade more when they are negative

Table 6: Apple (AAPL) regression results

The table shows the coefficients, t-statistics and adjusted R^2 of the regressions of daily historical returns and trading volume from 2018 to 2021 for Apple (AAPL). As described in section 3.3, the NoM-factor is constructed by collecting the number of times the AAPL ticker or company name was mentioned in posts or comments on the two subreddits within a pre-specified 24-hour window. The sentiment factor is a vector of daily sentiment scores estimated from all the posts and comments mentioning the stock.

	Return (1)	Trading volume (2)
Constant	- 0.00 (- 0.14)	126,232,500.00*** (25.54)
NoM_t	- 0.00 (- 0.36)	178,443.70*** (6.26)
NoM_{t-1}	- 0.00 (- 0.41)	- 17,271.97 (-0.61)
Sentiment	0.00 (1.05)	- 45,469,280.00*** (- 3.67)
Mkt-Rf	1.29*** (45.83)	
SMB	- 0.15*** (- 2.77)	
HML	- 0.68*** (- 11.73)	
RMW	- 0.60*** (7.33)	
CMA	0.70*** (6.48)	
Momentum	0.04 (0.89)	
Adj. R^2	0.70	0.065

*p<0.1, **p<0.05, ***p<0.01

and uncertain regarding AAPL-stock, with a decrease in sentiment score (less positive attitude) leads to a higher trading volume.

Also for AAPL the intercept is statistically significant in the trading volume regression, the coefficient is also positive but larger than for the Amazon regression. The magnitude of the intercepts can however not be directly compared as the stocks have very different trading volumes, with the average of AAPL being 123,583,922 which is much higher than for Amazon. Hence a larger intercept does not mean that the model explains less of the respective company's trading volume. In fact, the *Adj.* R^2 indicates that the model explains more of the variation in trading volume for AAPL than for AMZN. For AAPL, the intercept shows that the output estimate of daily trading volume will be approximately 126 million lower than the actual trading volume for that day.

5.2.3 Tesla(TSLA)

Table 7: Tesla (TSLA) regression results

The table shows the coefficients, t-statistics and adjusted R^2 of the regressions of daily historical returns and trading volume from 2018 to 2021 for Tesla (TSLA). As described in section 3.3, the NoM-factor is constructed by collecting the number of times the TSLA ticker or company name was mentioned in posts or comments on the two subreddits within a pre-specified 24-hour window. The sentiment factor is a vector of daily sentiment scores estimated from all the posts and comments mentioning the stock.

	Return (1)	Trading volume (2)
Constant	- 0.00 (- 1.39)	44,101,540.00*** (20.38)
NoM_t	0.000036*** (3.54)	75,160.39*** (8.56)
NoM_{t-1}	- 0.000024** (-2.37)	7,503.81 (1.1)
Sentiment	0.02** (2.28)	- 21,560,170.00*** (- 3.00)
Mkt-Rf	1.28*** (14.86)	
SMB	0.58*** (3.43)	
HML	- 0.29 (- 1.62)	
RMW	- 0.61** (- 2.42)	
CMA	- 0.92*** (- 2.78)	
Momentum	0.30** (2.47)	
Adj. R^2	0.28	0.146

*p<0.1, **p<0.05, ***p<0.01

The regression results for Tesla (TSLA) yielded the results in table 7. The regression results shows that the NoM$_t$ factor is significant at $\alpha = 1\%$ for both the return and volume regressions, and the NoM_{t-1} is statistically significant at $\alpha = 5\%$ for the return regression and not statistically significant for the volume regression. In addition the sentiment-variable is statistically significant at $\alpha = 5\%$ for the return regression and $\alpha = 1\%$ for the volume regression. This indicates that activity in the two subreddits has an effect on the returns of this particular stock, while for the others it appeared the forums had no significant effect. The coefficient on the NoM_t-variable is 0.000036 meaning that 1 more mention of the company on a given day leads to an increase in returns for that day by 0.0036%. Hence, an increase of 1,000 posts or comments mentioning the company on a given day, should increase the return by 3.6% for that day.

We did not find this result surprising as Tesla is the stock out of the 5 included which appeals the most to the average retail investor on Reddit. The company is owned and run by Elon Musk, which is one of the most active business-owners on social media and posts regularly on Twitter. Not only is Musk active on social media, the content he posts is often humorous ”memes” about popular topics. A ”meme” is simply a social media post, often in the form of a picture or a video that spreads a message by using humour. The picture or video is often of something completely different, but is used as an illustration. For example, someone may use a picture of a cat which looks grumpy to humorously present that they lost money on a particular trade. This type of content is by nature very appealing to members of the online community and the content is shared easily between people and can therefore spread fast across the internet.

Whenever Musk posts something related to financial markets it often has a large impact and leads to his companies being discussed all across the internet. We believe the significant NoM-factors and sentiment factor for the return regression of Tesla can be largely attributed to the level of engagement that followers of Elon Musk and Tesla has. It seems that this level of engagement makes discussions surrounding the company convert more frequently into trading.

Seeing as Tesla probably is the largest "meme-stock" of the stocks included in this analysis, the price behavior and price level is not perceived as rational by many institutional investors leading to them often shorting or avoiding the stock entirely. This, combined with the large engagement of retail investors surrounding the company makes it reasonable to believe that a larger portion of TSLA-stock is owned by retail investors, which as discussed previously may trade based on information obtained in online discussion-forums. Hence creating a stronger relationship between forum activity and TSLA - returns.

For the volume regressions, the results are similar to those of the other 4 stocks included in the analysis. The NoM_t variable and the sentiment-variable is statistically significant at $\alpha = 1\%$ while the lagged NoM - variable is insignificant at all levels. Similarly to the AAPL-regression the sentiment-variable is significant and the coefficient is negative. Indicating that an increased positive attitude among the forum members leads to less trading activity. The coefficient on the NoM_t -variable shows that one increase in mentions of TSLA stock leads to an increase in trading volume for that day by 75,160. Compared to the average daily trading volume of Tesla stock over our sample (47,519,553), this increase amounts to 0.16% of the average trading volume. This is marginally lower than the effect of Amazon. The intercept is as for AMZN and AAPL statistically significant with similar interpretations to the two previous.

5.2.4 Meta (FB)

Table 8: Meta (FB) regression results

The table shows the coefficients, t-statistics and adjusted R^2 of the regressions of daily historical returns and trading volume from 2018 to 2021 for Meta (FB). As described in section 3.3, the NoM-factor is constructed by collecting the number of times the FB ticker or company name was mentioned in posts or comments on the two subreddits within a pre-specified 24-hour window. The sentiment factor is a vector of daily sentiment scores estimated from all the posts and comments mentioning the stock.

	Return (1)	Trading volume (2)
Constant	0.00 (0.60)	15,459,750.00*** (17.08)
NoM_t	- 0.000068*** (- 4.33)	151,607.80*** (12.94)
NoM_{t-1}	0.000047*** (2.99)	2,978.86 (0.26)
Sentiment	-0.00 (- 0.13)	2,141,567.00 (1.05)
Mkt-Rf	1.07*** (26.92)	
SMB	- 0.10 (- 1.33)	
HML	- 0.35*** (- 4.29)	
RMW	0.39*** (3.35)	
CMA	- 0.87*** (- 5.73)	
Momentum	0.01 (0.25)	
Adj. R^2	0.50	0.179

*p<0.1, **p<0.05, ***p<0.01

For the return regression of Meta (FB) (Model 1 in table 8), we find that NoM_t and NoM_{t-1} is significant at $\alpha = 1\%$, which is similar to the results we obtained for Tesla. However, the coefficient on NoM_t is in this case negative, indicating that an increase in mentions leads to a decrease in returns for that same day.

Also for this stock, the volume regression yielded results similar to the others, with the NoM-factor being statistically significant, and the lagged NoM-variable being insignificant. As discussed we will investigate this relationship further in subsection 5.2.6. Also for this regression, the intercept is statistically significant at $\alpha = 1\%$ and positive.

5.2.5 Microsoft (MSFT)

Table 9: Microsoft (MSFT) regression results

The table shows the coefficients, t-statistics and adjusted R^2 of the regressions of daily historical returns and trading volume from 2018 to 2021 for Microsoft (MSFT). As described in section 3.3, the NoM-factor is constructed by collecting the number of times the MSFT ticker or company name was mentioned in posts or comments on the two subreddits within a pre-specified 24-hour window. The sentiment factor is a vector of daily sentiment scores estimated from all the posts and comments mentioning the stock.

	Return (1)	Trading volume (2)
Constant	0.00 (0.89)	23,806,920.00*** (19.28)
NoM_t	0.00 (1.28)	87,901.32*** (6.54)
NoM_{t-1}	- 0.000015* (- 1.81)	4,308.64 (0.32)
Sentiment	-0.00 (- 0.01)	49,445,691.00** (2.02)
Mkt-Rf	1.21*** (62.51)	
SMB	- 0.32*** (- 8.31)	
HML	- 0.32*** (- 8.01)	
RMW	0.38*** (6.67)	
CMA	- 0.36*** (- 4.84)	
Momentum	0.12*** (4.53)	
Adj. R^2	0.83	0.084

*p<0.1, **p<0.05, ***p<0.01

For Microsoft (MSFT), the return regression (model 1 in table 9) yielded a statistically significant NoM_{t-1} variable, indicating that for this stock number of mentions at a specific day has an effect on the returns on the following day.

The coefficient is negative, so one more mention of the company in a comment or a post on either of the two subreddits will lead to a 0.0015% decrease in returns the following day.

The volume regression is similar to most of the other stocks in the sense that the NoM_t is statistically significant while the lagged variable is not. The sentiment variable is statistically significant at $\alpha = 5\%$ and the coefficient is positive, meaning that the trading volume of MSFT-stock increases if the discussion surrounding the stock becomes increasingly positive. This relationship is the opposite of what we saw for TSLA and AAPL, where an increase in daily sentiment scores seemed to decrease trading volume for that particular day. Similar to all other stocks, the intercept of the trading volume is statistically significant, with explanation similar to the previous.

It is also worth noting that neither of the 5 stocks has a statistically significant intercept for the return regression, indicating that the model explains a lot of the variation in returns for the different stocks and that the returns of a particular stock can be relatively well replicated by investing in the different factor portfolios at weights equal to the betas. This is strenghtened by the fairly high $Adj.R^2$ for the majority of the return regressions.

5.2.6 Additional regressions

As discussed in the above sub-sections, we have very few statistically significant lagged NoM - variables for both the trading volume and return regressions. This combined with the way that the number of mentions is used in the creation of the NoM-factor, we want to investigate whether it is trading volume and stock activity that drives subreddit-discussions and not subreddit-discussions affecting the stock characteristics. In order to investigate this, we perform an additional set of regressions on all of the individual securities. These regressions will provide further evidence on whether trading volume and returns affect the forum discussions or forum discussions affect trading volume and returns.

The trading volume regressions looks the following and yields the following results:

$$NoM_{c,t} = \beta_1 + \beta_2\ Volume_{c,t-1} + \epsilon_t$$

Where $NoM_{c,t}$ is the number-of-mentions factor (described previously) for company c at time t and $Volume_{c,t-1}$ is the trading volume of company c at time t-1 (lagged).

Table 10: Regression results from regression of NoM-factor on lagged trading volume

The table shows the coefficients, t-statistics and adjusted R^2 of the regressions of the number-of-mentions (NoM) factor regressed on the lagged trading volume of each of the 5 individual stocks.

	AMZN (1)	AAPL (2)	TSLA (3)	FB (4)	MSFT (5)
Constant	52.61*** (13.93)	55.86*** (9.56)	43.48*** (5.34)	14.50*** (6.88)	21.69*** (6.49)
$Volume_{t-1}$	1.4e-07*** (3.65)	2.0e-07*** (4.72)	1.0e-06*** (10.32)	8.5e-07 (10.49)	8.0e-07*** (7.81)
Adj. R^2	0.012	0.021	0.095	0.098	0.056

*p<0.1, **p<0.05, ***p<0.01

The above results indicate that the lagged volume variable has a statistically significant effect on the NoM-variable at $\alpha = 1\%$ for all the stocks analyzed. This means that trading volume from the day before has an effect on the number of times that particular stock is mentioned in either of the two subreddits the following day. The coefficients on all stocks are positive indicating that an increase in trading volume increases the number of times the stock is discussed. For TSLA, which appears to be the most affected stock, a 1 million (shares) increase in trading volume leads to one more mention of the stock the following day. For MSFT, an increase in daily trading volume of 1 million shares only leads to an increase in the number of times the stock is mentioned that particular day by 0.8, which means that a larger change in trading volume is required to increase the number of mentions of MSFT than for TSLA.

The intercept is statistically significant for all regressions and tells us that the model does not capture all variables included in determining the number of times a stock is mentioned in the two subreddits on a given day. As for the trading volume regressions above, explaining what determines the number of mentions was not the objective of the model. For example, for Amazon, the intercept shows that if we input the trading volume yesterday and multiply this by the coefficient, the estimate of number of mentions for AMZN will be approximately 53 mentions fewer than the actual. These 53 mentions will be explained by other variables of influence. The $Adj.\ R^2$'s of the models are also quite low, indicating that the model does not explain a lot of the variation in the number of times the different stocks are mentioned throughout a given day.

The regressions of the NoM-factor on lagged excess returns look the following and yield the following results:

$$NoM_{c,t} = \beta_1 + \beta_2 \; R^e_{c,t-1} + \epsilon_t$$

Where $NoM_{c,t}$ is the number-of-mentions factor (described previously) for company c at time t and $R^e_{c,t-1}$ is the excess return of company c at time t-1 (lagged).

Table 11: Regression results from regression of NoM-factor on lagged excess returns

The table shows the coefficients, t-statistics and adjusted R^2 of the regressions of the number-of-mentions (NoM) factor regressed on the lagged excess returns of each of the 5 individual stocks.

	AMZN (1)	AAPL (2)	TSLA (3)	FB (4)	MSFT (5)
Constant	64.99 *** (40.23)	80.99*** (32.21)	111.36*** (23.55)	33.46*** (28.75)	45.70*** (32.87)
$R^e_{c,t-1}$	97.19 (1.18)	-124.34 (-1.03)	564.55*** (4.91)	- 115.56** (-2.26)	- 67.30 (- 0.91)
Adj. R^2	0.000	0.000	0.023	0.004	0.000

*p<0.1, **p<0.05, ***p<0.01

For the regression of number of mentions on lagged excess return we see that the lagged return variable is statistically significant for TSLA at $\alpha = 1\%$ and for FB at $\alpha = 5\%$, while it is not statistically significant for the other three companies. This indicates that for these three, the returns provided by the stock yesterday does not have a significant effect on the number of times this particular stock is mentioned.

It does however appear to have a significant effect for Tesla (TSLA) and Meta (FB). The signs one the coefficients are opposite each other for the two companies, and for TSLA it appears that an increase in excess returns the day before by 1% increases the number of times this company is mentioned by approximately 564 today. For Meta (FB) on the other hand, the effect seems the to be opposite, and a 1% increase in excess returns at time t - 1 decreases the number of times the stock is mentioned in a post or comment by approximately 116 at time t.

Looking back at the regressions in subchapter 5.2, we saw that for these same companies, the number of mentions factor had a significant effect on returns the next day. The results from these regressions build on these findings and shows that for TSLA and FB it also appears that returns have a significant effect on the number of mentions the following day. For the other companies, the effect of returns appeared to be non-existent.

6 Conclusion

The previous section discussed the results of the various analyses conducted in order to shed light on the research questions and the hypotheses underlying this thesis. In this section, we will conclude the thesis. The main objective has been to investigate the relationship between social media networking effects and the characteristics of the stock market.

Firstly, we aimed to answer the first of our research questions: Can a sentiment analysis of forums/discussions on large social media platforms present opportunities to create profitable trading strategies and portfolios? To answer this, we created portfolios based solely on the sentiment intensity scores of submissions and comments collected from the subreddits r/investing and r/stocks. The results of the analysis suggest that it is possible to generate profitable portfolios based on social media sentiment. By constructing portfolios in the way we did, we were able to create three different portfolios which outperformed two major indices and generated alpha when evaluated on the Fama-French five factor model including the momentum factor. These portfolios were the long portfolios of both the daily and monthly rebalanced portfolios as well as the long-short of the daily rebalanced portfolio. Surprisingly, none of the short portfolios we constructed generated alpha or outperformed the indices. These results indicate that the subreddit members are better at indicating long opportunities than short opportunities.

The second research question we aimed to answer was: Can social media attention be used to predict returns and trading volume? In order to analyze this, we performed regressions on the five individual stocks that were the most talked about throughout the period from 2018 to 2021.

The return regression yielded statistically insignificant NoM-variables and sentiment variables for most stocks. Only the regression of Tesla's returns yielded statistically significant coefficients for both the NoM- and lagged NoM-variables ($\alpha = 1\%$) as well as the Sentiment variable ($\alpha = 5\%$), while some of the other stocks yielded some statistical results as well. This indicates that for Tesla, how frequently the company is discussed, as well as whether the discussion is positive or not has an effect on the returns.

The most interesting results of this analysis did however come from the trading-volume regressions, where all stocks had statistically significant NoM-variables, indicating that an increased number of mentions in the forums does affect trading volume for the same day, however, none had significant lagged NoM-variables. This made us suspicious that it may be trading volume affecting discussions and not the other way around. The sentiment variable was significant for 3 out of 5 stocks, indicating that whether the forums are positive or negative does have an effect on trading volume for these stocks. The coefficient being negative for some stocks and positive for the other makes it difficult to specifically determine whether expressions of negative or positive opinions lead to the most trading. Our results suggest that this may be somewhat firm-specific.

We were also able to find that the trading volume predicts the number of times a stock is mentioned in the two subreddits the following day. Similarly, for Tesla and Meta we were able to find the same relationship for the returns' effect on the number of mentions the following day. However, for the other stocks, the returns appear to not have an effect. The second set of regressions that we ran, specifically regressing NoM_t on lagged trading volume showed that the lagged trading volume variable was statistically significant at $\alpha = 1\%$ for all companies. The results from the regression on lagged returns were however not as conclusive. Similar to the regression using returns as the dependent variable, the relationship seemed to be the strongest for Tesla (followed by Meta).

With Tesla being the stock being one of the most popular stocks across social media (see appendix D for total number of mentions per stock over the sample period), indicates that the relationship between social media and stock returns is the strongest for the most discussed companies. We did not find this surprising.

With trading volume, the relationship seems to be strong for all stocks. However, it is difficult to draw conclusive conclusions from these results as the first set of regressions suggests NoM has a statistically significant effect on same-day trading volume, while having no effect on the trading volume for the following day. In addition, the second set of regressions indicates that trading volume from the day before affects the number of mentions today. In this case, the causality may go both ways. However, due to the way the NoM-factor is constructed we suspect that the significance of the NoM_t may be a result of trading volume affecting discussions and not the other way around. This suspicion is not as large for the return regression, as the NoM_{t-1} variables are significant for the companies with significant NoM_t variables. Further research will be needed in order to draw conclusive conclusions.

For further research, we believe an analysis of the relationship between social media and stock market volatility could yield some interesting results. Does increased social media attention lead to changes in volatility and can this be traded upon? Secondly, an analysis covering whether the direction of the sentiment, positive or negative, is of importance. Or if it is simply the absolute size of the sentiment intensity which affects returns, trading volume, and similar. This could insinuate that emotions in general drive trading behavior and whether the emotions are positive or negative are of little to no influence. To investigate this, a similar approach to the one we have conducted using absolute values for the sentiment scores could be a valid approach. In addition to this, an analysis of whether the results are different using another social media platform for data collection could also be interesting. Is Twitter or Reddit the best at forecasting stock returns?

REFERENCES

Baumgartner, J., S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. 2020. The Pushshift Reddit Dataset. doi:10.48550/ARXIV.2001.08435.

Borg, A., and M. Boldt. 2020. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications* 162:113746. doi:https://doi.org/10.1016/j.eswa.2020.113746.

Buz, T., and G. de Melo. 2021. Should You Take Investment Advice From WallStreetBets? A Data-Driven Approach. doi:10.48550/ARXIV.2105.02728.

Chen, H., P. De, Y. J. Hu, and B.-H. Hwang. 2014. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *The Review of Financial Studies* 27:1367–1403. doi:10.1093/rfs/hhu001.

DeGroot, M. H. 1974. Reaching a Consensus. *Journal of the American Statistical Association* 69:118–121.

DeMarzo, P. M., D. Vayanos, and J. Zwiebel. 2003. Persuasion Bias, Social Influence, and Unidimensional Opinions. *The Quarterly Journal of Economics* 118:909–968.

Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56. doi:https://doi.org/10.1016/0304-405X(93)90023-5.

Fama, E. F., and K. R. French. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1–22. doi:https://doi.org/10.1016/j.jfineco.2014.10.010.

Hutto, C. 2014. VADER-Sentiment-Analysis. *GitHub repository.* https://github.com/cjhutto/vaderSentiment#about-the-scoring

Hutto, C., and E. Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8:216–225.

Islam, M., and M. Zibran. 2017. A Comparison of Dictionary Building Methods for Sentiment Analysis in Software Engineering Text. pp. 478–479. doi:10.1109/ESEM.2017.67.

Klepatch, J. 2021. reddit-sentiment-analysis. *GitHub Repository.* https://github.com/jklepatch/eattheblocks/blob/master/screencast/290-wallstreetbets-sentiment-analysis/data.py

Medhat, W., A. Hassan, and H. Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5:1093–1113. doi:https://doi.org/10.1016/j.asej.2014.04.011.

Pedersen, L. 2021. Game on: Social Networks and Markets. *NYU Stern School of Business Forthcoming* 116:1–22. doi:http://dx.doi.org/10.2139/ssrn.3794616.

Yu, Y., W. Duan, and Q. Cao. 2013. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems* 55:919–926. doi:https://doi.org/10.1016/j.dss.2012.12.028. 1. Social Media Research and Applications 2. Theory and Applications of Social Networks.

Zhang, X., H. Fuehres, and P. A. Gloor. 2011. Predicting Stock Market Indicators Through Twitter – I hope it is not as bad as I fear. *Procedia - Social and Behavioral Sciences* 26:55–62. doi:https://doi.org/10.1016/j.sbspro.2011.10.562. The 2nd Collaborative Innovation Networks Conference - COINs2010.

APPENDIX

A Overview of included stocks in monthly rebalanced sentiment portfolio

Table 12: Overview of stocks in monthly rebalanced portfolio

The table shows the tickers of the different companies that make up the monthly rebalanced portfolio per year. 2015 and 2016 consists of fewer stocks due to lack of data.

2015	2016	2017	2018	2019	2020	2021
AAPL	AAPL	AAPL	AAPL	AMZN	AAPL	AAPL
TSLA	TSLA	TSLA	TSLA	AAPL	TSLA	TSLA
AMZN	AMZN	AMZN	AMZN	GOOG	AMZN	AMZN
META*	META*	META*	META*	TSLA	META*	META*
AMD	AMD	AMD	AMD	AMD	AMD	AMD
BABA	MSFT	MSFT	MSFT	META*	MSFT	MSFT
MSFT	DIS	DIS	NFLX	NFLX	GOOG	DIS
NFLX	NFLX	NFLX	GOOG	DIS	DIS	NFLX
GOOG	GOOG	GOOG	TWTR	NVDA	NFLX	INTC
TWTR	TWTR	TWTR	INTC	MSFT	TWTR	NVDA
INTC	INTC	INTC	DIS	SQ	INTC	BABA
F	F	F	NVDA	BA	NVDA	V
	GPRO	GPRO	ATVI	INTC	ATVI	WMT
	NVDA	NVDA	WMT	V	BABA	UBER
		NTDOY	SQ	TCEHY	V	BA
		CGC	SNAP	CGC	WMT	NIO
		ATVI	SHOP	JD	UBER	PLTR
		GM	V	SNAP	BA	DKNG
		BAC	TCEHY	WMT	LYFT	ZM
		SSNLF	BABA	TWTR	F	SQ

*Ticker for Meta was updated from FB to META during the writing of this thesis

B List of words to update VADER - lexicon

Table 13: List of words used to update VADER - lexicon

The table shows the full list of words used to update the VADER - lexicon accompanied by the intensity score of the word.

’citron’: -4.0	’hidenburg’: -4.0	’moon’: 4.0	’highs’: 2.0	’break’: 2.0
’long’: 2.0	’overvalued’: -3.0	’call’: 4.0	’calls’: 4.0	’put’: -4.0
’puts’: -4.0	’mooning’: 4.0	’tendie’: 2.0	’tendies’: 2.0	’town’: 2.0
’short’: -2.0	’undervalued’: 3.0	’buy’: 4.0	’sell’: -4.0	’gone’: -1.0
’gtfo’: -1.7	’maintain’: 1.0	’bullish’: 3.7	’bearish’: -3.7	’green’: 1.9
’stonk’: 1.9	’bagholder’: -1.7	’money’: 1.2	’print’: 2.2	’rocket’: 2.2
’bull’: 2.9	’bear’: -2.9	’sus’: -3.0	’offering’: -2.3	’drop’: -2.5
’rip’: -4.0	’downgrade’: -3.0	’upgrade’: 3.0	’paper’: -1.7	’pump’: 1.9
’hot’: 1.5	’pumping’: -1.0	’rebound’: 1.5	’crack’: 2.5	

C Number of posts and comments per year in dataset

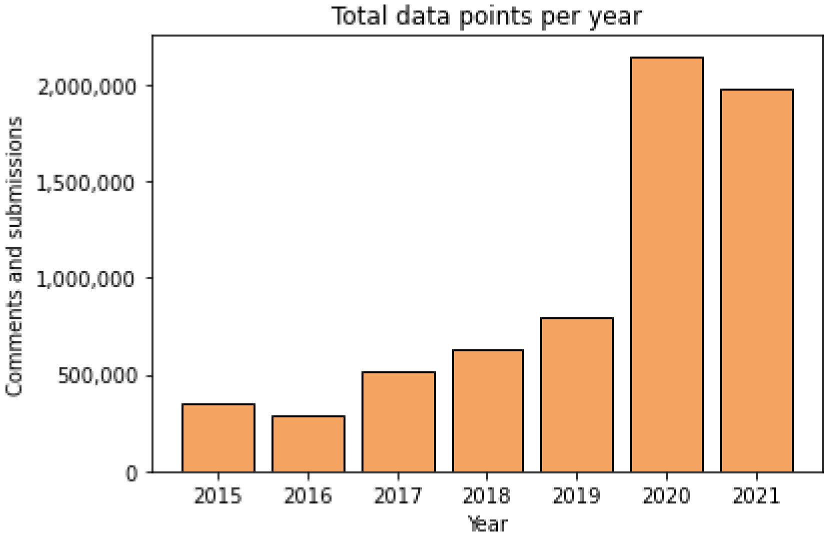

Figure 2: Number of posts and comments per year

D Total number of posts and comments - individual stocks

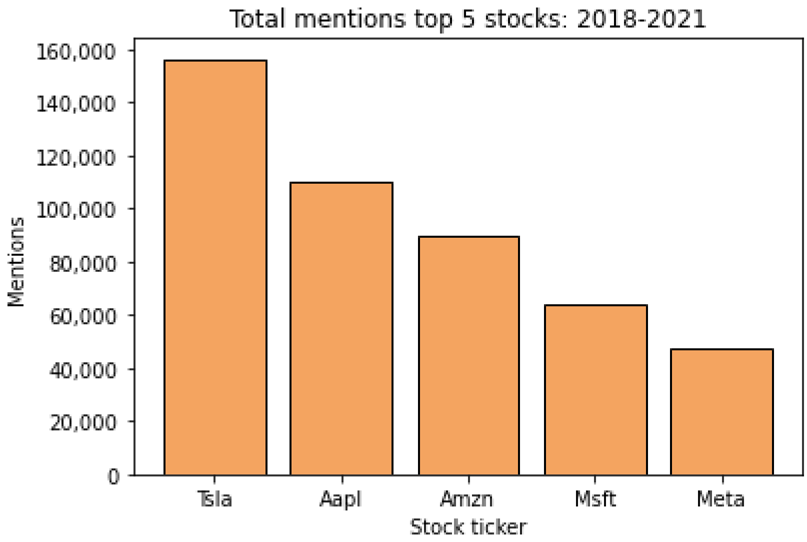

Figure 3: Number of posts and comments for most mentioned stocks

E Cumulative returns: Daily rebalanced— portfolio

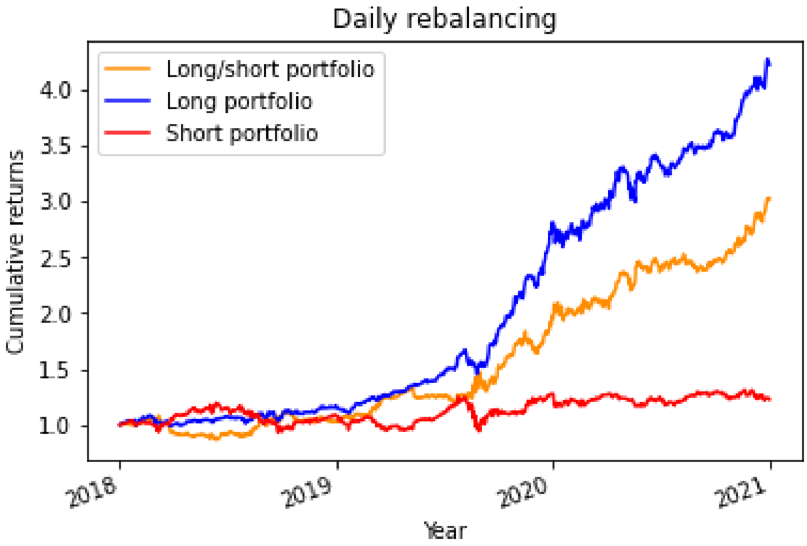

Figure 4: The figure shows the cumulative returns of the daily rebalanced sentiment-portfolios (investing 1 dollar in Jan 2018)

Note that the short portfolio is displayed as if you went long the short portfolio. In order to get the short returns, multiply with -1 (the graph will be inversed)

F Cumulative returns: Monthly rebalanced portfolio

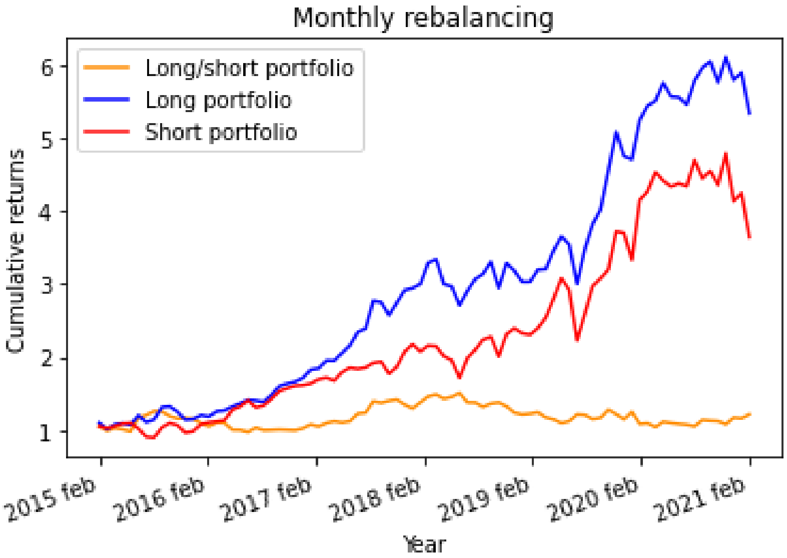

Figure 5: The figure shows the cumulative returns of the monthly rebalanced sentiment portfolios (investing 1 dollar in Feb 2015)

Note that the short portfolio is displayed as if you went long the short portfolio. In order to get the short returns, multiply with -1 (the graph will be inversed)

G Subscriber development - r/investing

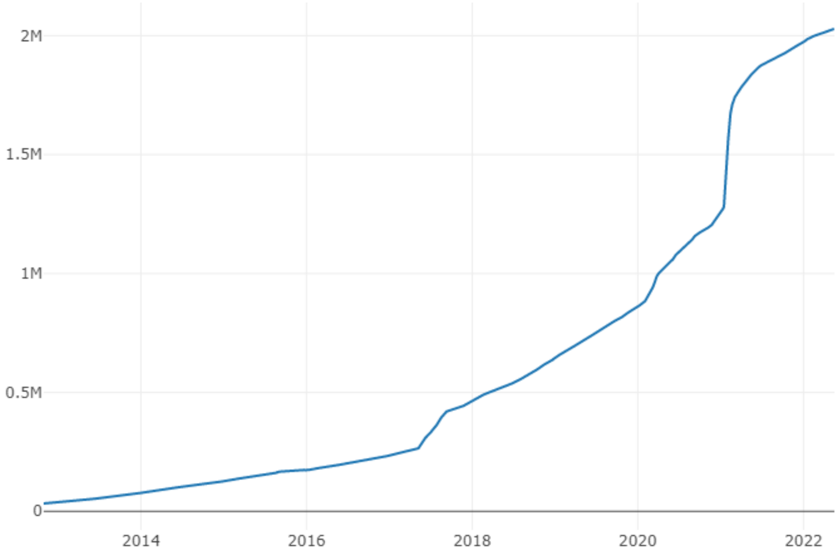

Figure 6: The figure shows the subscriber count in millions from for the subreddit r/investing

H Subscriber development - r/stocks

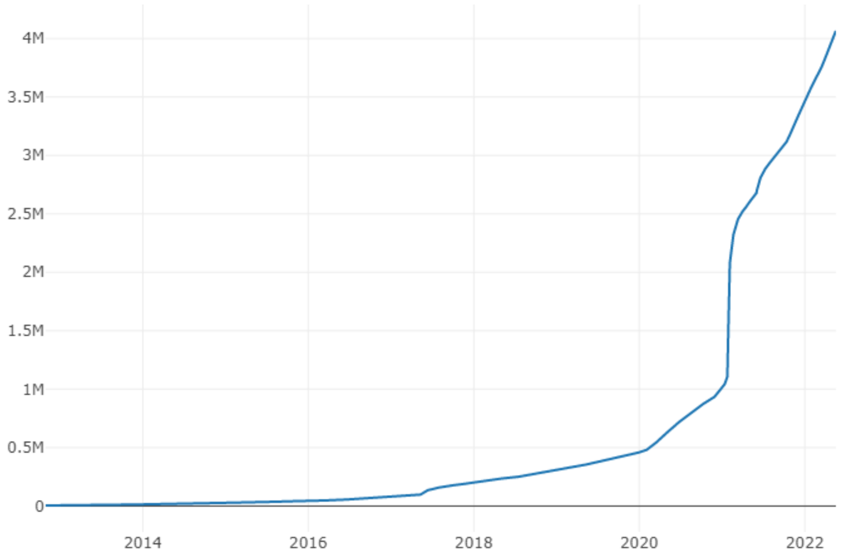

Figure 7: The figure shows the subscriber count in millions for the subreddit r/stocks

I Subscriber development - r/WallStreetBets

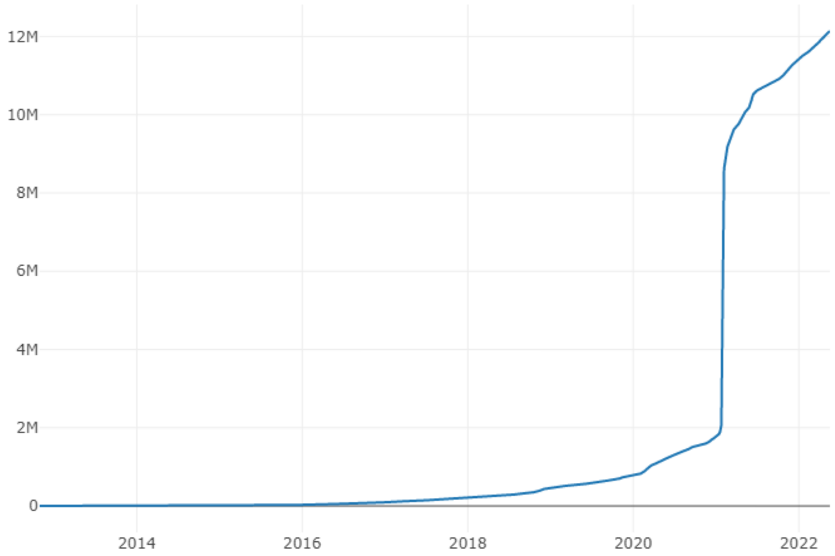

Figure 8: The figure shows the subscriber count in millions for the subreddit r/WallStreetBets