Business Analytics Report

Using machine learning to develop an algorithmic trading strategy based on Reddit's *Wallstreetbets* thread

CID: 01892048

Abstract

Throughout the existence of the stock market, financial institutions, high-frequency traders and also private individuals have sought methodologies to outperform financial markets and generate abnormal financial returns. Countless strategies and algorithms have been developed in the pursuit of that goal and continue to be refined. The emergence of the Reddit forum *Wallstreetbets,* which grew to prominence in 2021, provides a new source for data that may be drawn upon to develop such strategy and generate abnormal returns.

In that context, this report develops and tests a Naïve Bayes, Nearest Neighbour, Logistic Regression and Random Forest model based solely on data from *Wallstreetbets* to develop an algorithmic trading strategy in the pursuit of that goal. By examining data correlations and testing a broad set of features, an algorithm is developed that sends buy, sell and hold signals for a selection of stocks. This algorithm is optimized for a range of buy, sell and hold thresholds and ultimately simulated on a test set from Nokia and Bed, Bath and Beyond. The Random Forest model generates the greatest compounded daily growth rate and yields daily average returns of 1.5% and 2.4%, respectively. The findings are compared to three alternative investment opportunities. These include a passive investment in the respective tickers, a random buy/sell/hold decision, and a market index ETF.

The active trading strategy fails to sustainable outperform a passive investment strategy into the *Wallstreetbets* stocks. However, this report still highlights that an investment strategy incorporating stocks discussed in *Wallstreetbets* data is able to generate superior financial returns compared to the general market and thus create abnormal returns. The findings in this model can be extended to incorporate macroeconomic or financial features, that when intertwined with the data derived from Reddit, may yield in an active strategy that can outperform a passive strategy.

Table of Content

1) Introduction

"Facts only account for 10% of the reactions on the stock market; everything else is psychology" (Kostolany cited in Nann, 2019). The public reddit forum *Wallstreetbets* and the impact that it had on the share price of certain equities, most famously Gamestop ("GME"), are indicative of some truth behind Kostolanys opinion. The online forum grew to prominence as it enabled millions of retail investors to publicly coordinate stock trades, therewith influencing the market price of the underlying equities, contrary to financial fundamentals and macroeconomic climate.

Wallstreetbets has ~10.5 million members as of June 2021, and is primarily used for stock market discussions, predictions, and speculations. GME is a publicly listed electronics retailer for video games that has witnessed a steady decline in its share price since 2014, deteriorating from a share price of $55 USD to $2.8 at its lowest level. In addition to the changing global landscape for retailers and consequences of the Covid pandemic, it has come under pressure from hedge funds that shorted the stock, meaning they borrowed the security and then sold it to the market, with the intend of repurchasing it later for a lower value (Sosnoff, 1966). Short selling places downwards pressure on the price of the underlying equities, particularly when a large amount of stock is short sold. In September 2019, a *Wallstreetbets* member shared information that hedge funds, and particularly Melvin Capital, had a large position short-selling GME, indicating that if large volumes of GME are purchased and the price driven upwards, the hedge fund will lose a large share of the capital invested. This triggered a herd movement in *Wallstreetbets,* with millions of retail investors coordinating trades and purchasing GME stock in masses, pushing the price up to a maximum of $347 and causing Melvin Capital to lose billions of dollars (Schroeder, 2021). Although argued to be a grey area, this coordinated yet public effort to manipulate the market does not constitute "any violations in law" and resulted in coordinated price pumps in further assets (Stacey, 2021).

In that context, the aim of this assignment is to utilize sentiment analysis and machine learning models to identify if the posts and comments in *Wallstreetbets* can be used to generate an algorithmic trading strategy that generates reliable short-term returns (an "active" trading strategy). Reddit data from 10 stocks[1] with 220,621 Reddit comments, is drawn upon and the resulting active trading strategy is compared to a passive position in Reddit stocks, a random buy/sell/hold decision in those and the performance of respective market indices. Comparisons are based upon the ability of generating financial returns to an investor.

[1] Gamestop (GME), AMC Entertainment (AMC), Nokia (NOK), Tesla (TSLA), Blackberry (BB), Microvision (MVIS), Bed Bath & Beyond (BBBY), Clover Health (CLOV), Sundial Growers (SNDL), ContextLogic (WISH)

2) Literature Review

A holistic set of research explores the prediction of financial markets, with the consensus being that movements in financial markets cannot be accurately forecasted in a sustainable fashion (Stibel, 2009). Fama's (1970) efficient-market hypothesis suggests that this inability to predict stems from the fact that financial markets reflect all available public information at any time, with share prices adjusting immediately to any changes in information. In accordance with this theory, outperforming the stock market over a sustained period is deemed challenging and baselines on stock prediction studies tend to be at 50 percent accuracy thresholds (Gupta, 2021). However, the efficient market hypothesis is contested by various scholars, including behavioural psychologist, that allege that stock markets are subject to human bias'. For example, Hirshleifer and Shumway's (2003) study finds correlation in the stock market between sunshine in the city and stock performance. Similarly, herding, meaning mass movements to buy or sell an asset that is counterintuitive to the underlying public information, is also cited as a violation to the efficient-market hypothesis (Graham, 1999). Historically, there are numerous events that have led to herd movements that impacted financial markets, meaning they can be traced back to psychologic decisions versus financial information (Galbreith, 1990), for example the dotcom bubble (Wollscheid, 2012). Inefficiencies as such has facilitated the rise of algorithmic trading and provides potential for active investors to outperform passive investors (Birla, 2012).

The extent of which social media can impact share prices and whether it has predictive power in developing a successful active trading strategy is rather contested amongst scholars. Li et al. (2019) conducted a social influence model to allege that social media posts in the platform Tencent Weibo help explain short term market fluctuations in the Hushen 300 index. Bernatzi (2015) corroborates that social media can feed into an investor's cognitive bias and investment decisions and is empirically validated by Sisk's (2013) study. The latter examined the impact that sentiment on social media has on asset prices. The underlying principal component analysis finds 30 features that justify 25% of the variance, thus alleging that social media sentiment has an impact on asset prices (ibid.). Similarly, Pineiro et al. (2017) apply a logit model and a fuzzy-set qualitative comparative analysis to allege that social media sentiment influences the stock market and verify a correlation between social media posts and VIX prices. However, Valle-Cruz et al. (2021) hypothesis of reverse causality can be cited as a counterargument to their findings. It is suggested that social media reacts to financial market behaviour and can be a source of such correlation for up to 11 days after an event, therewith minimizes predictive power, yet still justifying correlation. Pineiro et al. (2017) have also stressed that particularly nontechnical investors are subject to influence from social media and might adapt their investment strategies accordingly. Given the ability to mobilize masses via social media, it could be deemed a catalyst for herd movements, particularly amongst nontechnical investors that are more prone to social media sentiment, which in turn are a source of violation to the efficient market hypothesis (Lycosa et al., 2021).

Leung et al.'s (2019) frequently cited study does not fall into this category. The study examined 30 million tweets from *StockTwits*, a Twitter forum for traders comparable to *Wallstreetbets* by nature, to verify a predictive ability to forecast share price movements. The scholars applied diverse NLP and baseline regressions to analyse the relationship between social media sentiment and stock returns. Findings indicate a predictive power in social media sentiment at the 1% statistical significance level. However, the scholars indicate that it is not herding that justifies the predictive feature, but rather the users' ability to forecast future earnings. As *Wallstreetsbets* only became renowned in 2021, there is limited research on its predictive ability, yet Gupta's (2021) Recurrent Convolutional Neural Network

that assessed data in the last quarter of 2020 and the first quarter of 2021, resulted in an accuracy of 59.4% for Tesla share price movement predictions and 52.1% for GME predictions, therewith beating the earlier cited 50% baseline benchmark. These findings arguably violate Fama's (1970) efficient market hypothesis at least in the short-term. Diverse explanations may be cited for the findings, including that the public information from the forum is too young to have been accurately priced into markets and will be priced in over the long term (Gupta, 2021). A more revolutionary argument might be that the democratization of equity investments through innovative FinTech's changed market dynamics and provided greater power to retail investors, which will continue to disrupt markets over the long-term (ibid.).

Lycosa et al. (2021) refer to a linear regression model estimated via OLS to find that the next day's price variation can be partially traced back to an increase of activity on *Wallstreetbets* relative to Google searches. Having examined 4 popular equities, the scholars display that the return correlation between those is at 0.614 in January and inexistant to a market ETF (ibid.). This indicates a clear decoupling from the market and hints that superior returns using the forum could indeed be generated. However, for the following months this correlation deteriorated significantly. The scholars also detect that the scope of the returns is of different magnitudes and cannot be argued to be uniform across tickers (ibid.). Buz and Melo's (2021) analysis of *Wallstreetbets* posts suggests a similar conclusion on return potential yet delivers more pessimistic findings for active investors. Having examined data from January 2019 to April 2021 for 20 equities, they find that while a passive investment strategy based on reddit sentiment outperformed the S&P500, the short-term accuracy of buy and sell signals is not found to be statistically significant (ibid.). Comparing the data with random buy decisions within the same timeframe, the researchers argue that a random buy decision is superior (ibid.).

3) Data collection

Data is collected from two sources, Reddit and Yahoo Finance.

3.1) Reddit

Reddit provides the API PRAW that enables developers to access and scrape Reddit posts. The API allows access to diverse sets of information in addition to the underlying comment. This includes username, time of post, thread name, and more. However, PRAW limits the number of submissions that can be extracted, with computational work also being an additional consideration. For example, scraping data with a limit of 100 sub threads for a single ticker is computationally expensive and may take longer than an hour even with a speedy device. Thus, posts and comments with a limit of 250 threads have been scraped using PRAW on the aforementioned 10 Tickers. The selection of tickers originates from media coverage that verifies intensive Reddit coverage of such (van Doorn, 2021; Wigby, 2021).

3.2) Yahoo Finance

Daily share price data has been downloaded locally from Yahoo Finance. This has been stored locally and uploaded. Both opening and closing share prices have been extracted and are used in the analysis.

4) Data processing

Each individual Reddit posts has undergone a sentiment analysis. There has been some data cleaning before and data transformation after the sentiment analysis for Machine Learning purposes.

4.1) Data cleaning

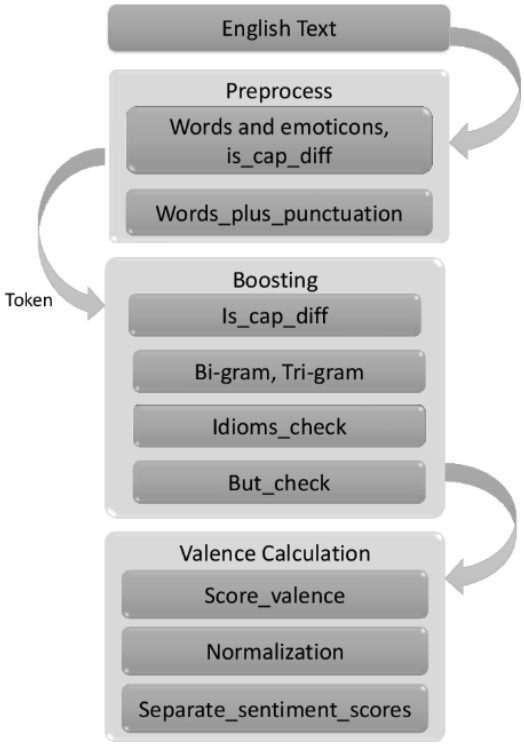
Two data cleaning decisions were made. Firstly, certain posts had missing values. This is to be traced back from posts that contain images, hence, these have been removed from analysis. Secondly, the format of the PRAW API does not provide posts on every historical day but also skips days occasionally. Thus, for certain historical days, the equity share prices are available but not the respective Reddit data. A column reflecting the share price of the previous day(s) has been added to the dataset for future use case, afterwards the share prices on dates with no Reddit data are excluded.

4.2) Sentiment analysis

There are diverse methodologies of conducting sentimental analysis including unsupervised machine learning models or lexical-based approaches (Araujo et al., 2013). Whereas for different needs a different form of appraisal may be deemed appropriate, the nltk library and specifically VADER and the SentimentIntensityAnalyzer tool have been considered suitable. VADER is an efficient pretrained algorithm that enables sentiment analysis for social media applications and proved effective in processing the slang and abbreviations that occur on the reddit forum. It does so by blending a sentiment lexicon approach, grammatic rules and syntactical conventions to convey sentiment polarity and intensity (Ma, 2020). It is argued to be less accurate with longer and structured texts (Mogyoroski, 2021). Given the typical format of a reddit comment, this trade-off has been deemed appropriate for purposes of this analysis.

The tool provides a sentiment score through three steps as visible in figure 1. First, it pre-processes the text and ensures that the posts can be accurately interpreted, which also includes emojis, as is particularly common in Reddit posts. In the second step, the algorithm interprets the underlying text to the extent that it assesses sentiment counterproductive words like *but*. Lastly, a normalized score is calculated for sentiment, including positive, negative, neutral and compounded sentiment. In this context, compounded sentiment is the collective result of the sentiment of the first three measures. The sentiment scores range between 1 and -1, with the first indicating an entirely positive post whereas the latter indicates a completely negative post.

Figure 1: VADER function breakdown



(Amin et al., 2019)

4.3) Data transformation

Following the sentiment analysis, separate datasets are present for each ticker. As the aspiration is to develop an active trading strategy based on machine learning, the algorithm must be able to send buy or sell signals subject to certain features. To attain those, some further transformation has been undergone, a step-by-step explanation of the transformation process is found in appendix 1. Appendix 2 details summary graphs of the transformed data.

5) Correlation discussion

Having randomly allocated NOK and BBBY as the test set, the remaining 8 tickers have all been individually and collectively been inspected for correlations. When combined into a single dataframe, share price is correlated with the respective variables as illustrated below:

Table 1: Correlations between key variables for the combined dataframe

	Daily Sentiment	Proportion assessment	Share price	Post quantity
Daily Sentiment	1.000000	0.922202	0.009278	-0.213791
Proportion assessment	0.922202	1.000000	0.017346	-0.184413
Share price	0.009278	0.017346	1.000000	0.235978
Post quantity	-0.213791	-0.184413	0.235978	1.000000

Proportion assessment refers to the share of positive comments and post quantity to the amount of comments made to the 250 posts on the forum. The value of interest, closing share price, is slightly positively correlated with post quantity (+0.24), yet not correlated at all with daily sentiment or the proportion assessment variable. Two phenomena could explain this finding: Firstly, it is possible that the two variables are truly not correlated to share price by nature. Secondly, in accordance with the findings of Lycosa et al. (2021), who allege differences in Reddit impacts on the underlying equities, it is possible that daily sentiment is only relevant for certain tickers, or perhaps only during the momentum building phase. This could be hinted by the fact that daily sentiment and post quantity display a certain negative correlation (-0.21). Furthermore, exploration of the correlation of the individual equities strengthens this idea, as is most vividly illustrated by the correlation table for CLOV:

Table 2: Correlations between key variables for CLOV

	Daily Sentiment	Proportion assessment	Share price	Post quantity
Daily Sentiment	1.000000	0.945225	0.380929	-0.215566
Proportion assessment	0.945225	1.000000	0.355424	-0.193883
Share price	0.380929	0.355424	1.000000	0.120049
Post quantity	-0.215566	-0.193883	0.120049	1.000000

With a correlation of daily sentiment to closing share price of 0.38, and a correspondingly lower correlation with post quantity (+0.12), the behaviour of CLOV differs noteworthy from the average findings. In accordance with the summary statistics in Appendix 2, most datapoints stem from summer 2021. CLOV witnessed its peak pricing on the 8th of June 2021, whereas other renowned *Wallstreetbets* stocks had their first major peak earlier, for example SNDL on the 10th of February 2021 or GME on the 27th of January. The share prices of both these equities have a negative correlation with daily sentiment as illustrated in appendix 3. Thus, as further discussed in section 8, a point for further study could examine whether daily sentiment is a signal for the early stages of a herd movement on Reddit.

6) Machine learning models

Machine learning models are the basis for the algorithmic trading strategy. Firstly, a diverse set of features has been created. Next, a benchmark strategy has been developed that gives a buy signal (1) if the predicted percentage return time, $t+1$, is above 1%, and a sell signal (2) if the predicted percentage change is below -1%. For predictions in the interim, the position is held (0). The signal is indicated by the y in the logistic regression below.

$$y = \begin{cases} 2 & \frac{1}{1+e^z} < -1\% \\ 1 & \frac{1}{1+e^z} > 1\% \\ 0 & else \end{cases}$$

Where

$$z = \beta_0 + \beta_n x_n + \varepsilon$$

And β_n refers to the coefficient of parameters of the nth parameter x_n and ε to the error term.

Based on that strategy, a Naïve Bayes, a Nearest Neighbour model, a Logistic Regression and a Random Forest model have been trained and tested for accuracy using cross-validation. This allows the model calibration and assessment of features. Section 7 further optimizes this strategy for different buy and sell thresholds and investments amounts.

6.1) Feature engineering

The final set of features is illustrated below:

$$z = \beta_0 + \beta_1 p_t + \beta_2 p_{t-1} + \beta_3 d + \varepsilon$$

Where:

- p_t = post quantity: Number of comments scraped within the set post limit
- p_{t-1} = post quantity in last available day: Number of comments scraped within the set post limit on prior day / last day with available data
- d = difference since last available date: As the structure of the Reddit API omits consecutive days occasionally, but rather has some gaps in between them, this feature provides flexibility to reduce the predictive power of past quantities if they did no occur on the previous day.

The Appendix Jupyter notebook shows further predictive variables that have been tested and excluded. This includes, but is not limited to, daily sentiment and proportion assessment. Their exclusion yielded in better cross-validated accuracies. This is because the models overfit at their inclusion as corroborated by the correlation appraisal.

6.2) Model initialization

Using the benchmark strategy, four cross-validated machine learning models with 9 folds have been initialized and tested. In regard to both mean accuracy and F1 score, the logistic regression outperformed the other machine learning models.

Table 3: Cross-validated machine learning performance on benchmark case

ML Model	Accuracy mean	F1 mean
Naive Bayes	0.53	0.68
Nearest Neighbour	0.48	0.62
Logistic Regression	**0.59**	**0.73**
Random Forest	0.48	0.60

When comparing the models accuracy to Gupta's (2021) recommended forecasting level of 50%, the benchmark model appears promising. However, two counterarguments must be noted. First, machine learning theory implies that this is an overestimation from the true performance as the model is based on only the training and validation data set (Brownlee, 2014). Secondly, exploration of the decisions returned by the logistic regression model indicates that the model in the benchmark case almost always provides a sell signal, and is therefore close in performance to the majority classifier that has a 58% accuracy on the training set if the decision is always to sell. The confusion matrix in appendix 4 is illustrative of that. Thus, although superior in accuracy and F1 score, a logistic regression model is less promising for algorithmic trading in the benchmark case. Furthermore, a range of training and validation splits have been tested to explore the potential of the benchmark case. The metric used to guide the decision on ideal split has been accuracy as the class distribution is similar with financial markets operating with a roughly equal ex-ante probability for a price increase and decrease (Damodaran, 2019).

Table 5: Machine learning performance optimized for ideal split

ML Model	Ideal test split	Accuracy	F1	Precision	Recall
Naive Bayes	0.10	**0.60**	**0.57**	0.66	**0.60**
Nearest Neighbour	0.25	0.46	0.40	0.44	0.46
Logistic Regression	0.11	0.54	0.38	**0.75**	0.54
Random Forests	0.11	0.48	0.45	0.49	0.48

The table represents the results of a single random sample with the optimized split ratio. The naïve bayes model outperforms in the benchmark case when given the flexibility to change the test / validation split ratio and does not solely follow the majority classifier with sell signals.

7) Algorithmic trading

Having initialized the models, these are now utilized to derive an algorithmic trading strategy and compared to alternative investment opportunities.

7.1) Variable optimization

Do develop a trading algorithm, we require certain inputs as to simulate an investment and measure its performance. Specifically, it's inputs include:

- Max budget: The total amount of cash that the investor can devote to the investment strategy. This has been set at $5,000.
- Investment amount: The value or proportion of the max budget that the investor is willing to invest with a buy signal, assuming sufficient cash is held. The underlying algorithm works with a fixed amount.
- Buy threshold: The percentage change predicted, for the algorithm to issue a buy order.
- Sell threshold: The percentage change predicted, for the algorithm to issue a sell order. Any values in the interim are held.

As part of this analysis, performance assessment neglects transaction costs and tax considerations. Using data from the 8 training tickers, thousands of combinations have been tested to retrieve the inputs that generate the greatest daily financial return. Daily financial return is computed using an amended form of the CAGR formula that we label CDGR, compounded daily growth rate (Chan, 2009).

$$CDGR\ (t_0, t_n) = (\frac{V(t_n)}{V(t_0)})^{\frac{1}{t_n - t_0}}$$

Where t_0 indicates trading day 0, t_n the final trading day and V the value of the portfolio at the given time. This optimization task has been computed for all four machine learning models, 4,410 iterations have been simulated for each model. The table below shows the CDGR attained by the optimal combination for each model.

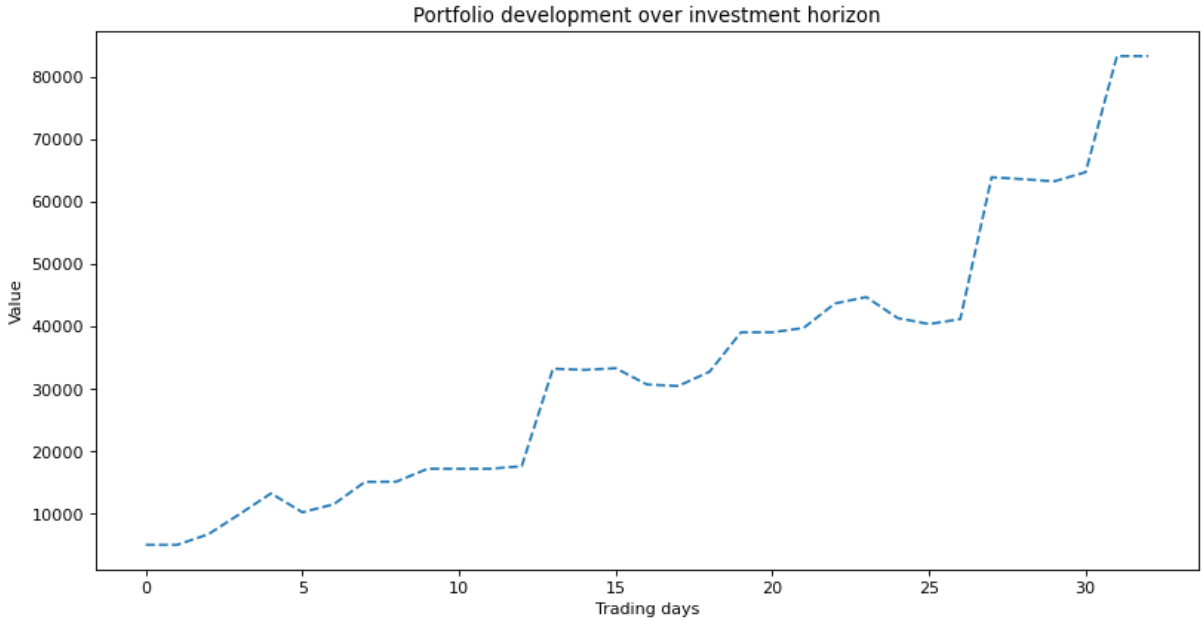
Table 6: Daily return attained on training set

ML Model	Daily return
Naive Bayes	8.8%
Nearest Neighbour	5.5%
Logistic Regression	8.8%
Random Forests	8.9%

It becomes evident that the greatest CDGR is generated by a random forest model, attaining 8.9% of gains on a daily basis using this model. The optimal buy threshold for this model is -10%. This indicates that if the model predicts a percentage change of above -10%, the equity should be bought. This is indicative of the fact that the initialized models systematically underestimate returns, however, the ability to adapt the strategy according to a threshold allows us to incorporate this into our algorithmic trading decisions. The respective optimal sell threshold is set above the buy threshold, and, given the

format of the code, a position is never held but will always be sold if the projected percentage change is lower than the optimal buy threshold. Moreover, the model indicates that an "all-in" initial investment for each trade, meaning $5,000, generates the greatest returns as opposed to splitting the transaction amount into smaller investments.

Figure 2: Portfolio development using a random forest on validation set



The decision set generated by the random forest model is visualised above and generates a multiple on invested capital ("MOIC") of 15.7x over 33 trading days, making a $5,000 worth $78,378 by the end of the investment lifespan. It appears plausible that this is an overestimation of the true performance given that this has been conducted on the validation set, the section below applies the optimized model for the two test sets (Brownlee, 2014).

7.2) Results and discussion

7.2.1) Nokia

Applying the optimal trading strategies to the NOK test set, one can see positive returns. Specifically, over 26 trading days, a return of 30.8% is generated using a random forest model (CDGR of 1.0%). The graphic visualization of the portfolio development is illustrated in appendix 5. Although a CDGR of 1% is less than projected with the training data, such a return is stronger than the performance of many major hedge funds (Baron et al., 2014). However, to truly assess the trading strategy it must be compared to a peer group of alternative investments, specifically three benchmarks are used: a passive NOK investment, a random decision, and the EuroStoxx 50 index in which Nokia trades.

Table 7: NOK benchmark comparison

Benchmark	Superior?	Comment
A passive investment in NOK	No	A passive investment in NOK would have yielded returns of 48.0% (CDGR of 1.5%)
A random buy/sell/hold decision	Yes	A random investment strategy in NOK provides mean returns of 0.8% after 10,000 iterations (CDGR of just above 0%)
Compounding at EuroStoxx 50	Yes	NOK is in the EuroStoxx 50 index, which returned 0.8% over this timeframe (CDGR of just above 0%)

Table 7 illustrates that the returns generated by the random forest model significantly outperforms a random decision and the respective EuroStoxx 50 market index in which NOK trades. Please find in appendix 6 the return distribution of a random decision for NOK. The passive investment, which invests the full investment amount on day 1 and only sells at the end of the investment horizon, has outperformed the active trading strategy.

7.2.2) Bed Bath and Beyond

For BBBY, the model achieved a return of 26.8% over 18 trading days (CDGR of 1.3%). The graphic visualization is available in appendix 8 and the table below compares the results to the three benchmarks.

Table 8: BBBY benchmark comparison

Benchmark	Superior?	Comment
A passive investment in BBBY	No	A passive investment in BBBY would have yielded returns of 53.0% (CDGR of 2.4%)
A random buy/sell/hold decision	No	A random investment strategy in BBBY provides mean returns of 30.4% after 10,000 iterations (CDGR of 1.5%)
Compounding at Nasdaq Composite	Yes	BBBY is in the Nasdaq Composite index, which returned 1.0% over this timeframe (CDGR of just above 0%)

Table 8 illustrates that the strategy does not outperform a passive position in BBBY nor a random active decision. Please find in appendix 9 the return distribution of a random decision for BBBY. The strategy continues to significantly outperform an investment in a market index.

7.2.3) Discussion

For both NOK and BBBY, the active trading strategy failed to outperform a passive position. This is sufficient evidence to out rule that the optimized models is a *sustainable* active trading strategy superior to a passive position for Reddit stocks. This insight falls in line with the findings of Buz and Melo (2021) that report similar findings. Furthermore, the findings partially echo Buz and Melo thesis that an active strategy is not superior to random decisions in *Wallstreetbets* stocks. However, as the active strategy for NOK did outperform the random decision-based trading system, it could still be possible that, on average, an active approach might be superior to random decisions. Limitations on data quantity might also justify this, as is further discussed in section 7.3.1.

An additional insight generated by the random decision benchmark is that of volatility. Although a passive investment in NOK and BBBY would have generated healthy CDGR, 1.5% and 2.4%, the mean CDGR for a respective random decision is unproportionate, close to 0% for NOK and 1.5% for BBBY. This stems from differences in volatility between the equities and leads to a greater chance of engaging in a bad trade for NOK as compared to BBBY. As a result, volatility considerations of the underlying asset, or proxies for such, might be worth examining for future investigations as it could impact the performance of a trading algorithm, this is further discussed in section 8.

Comparing the trading strategy with the respective index benchmark, the random forest model significantly outperformed such for both test sets. This strengthens the claim that Reddit data can be utilized to outperform market indices as hinted by the correlation assessment, even if potentially only in a passive investment strategy. It accordance with corporate finance theory, it must be acknowledged that a greater risk is embarked upon when investing in a single stock as opposed to a market index and thus a larger return is to be expected (Richardsen, 1970). However, CDGR's of 1.5% or 2.4% clearly outperform the expected returns of most single assets with any asset pricing model, including the CAPM model applied in most academic studies (Mullins, 1982).

7.3) Limitations

7.3.1) Data quantity

Whereas the group *Wallstreetbets* has existed since 2012, it has only gained major traction and momentum with the Gamestop break-through. As a result, the majority of comments interesting for predictive purposes is somewhat limited to data in 2021. Although more than 200,000 comments have been scraped and analysed, these are ultimately grouped by date for machine learning purposes and thus restricted in quantity. Furthermore, quantitative limitations also apply for the amount of tickers that can be studied. Given that the herd movement is comprised of many small retail investors, these must target a selected number of stocks as to truly impact prices. As a result, the number of stocks that fall within this category is arguable less than 20 (Ponczek, 2021). This further makes it difficult to judge the success or failure of the active trading strategy as the amount of test data is limited. It cannot be out ruled that the active trading strategy might outperform a passive position for the next trending Reddit stock.

7.3.2) Strategy sustainability

The trading strategy might not be sustainable over the long-term due to two reasons, erosion of herd behaviour and algorithmic adoption. For post quantity to be a successful predictor variable in a trading strategy, it must also correlate with actual purchases of the underlying equity. However, the herd movements that caused the price uplifts for the studied tickers have been fuelled by optimism of financial returns and also aspirations of fighting Wallstreet (Sonnemaker, 2021). With many retail investors also experiences major losses during price declines and complains of bots infiltrating the forum as to promote equities to profit of a pump and dump strategy, the long-term sustainability of basing an investment strategy on Reddit data is threatened (McEnery, 2021). The fact that many leading voices of the *Wallstreetbets* have left the forum in the second quarter of 2021 is indicative of this (ibid.). Secondly, retail traders were able to price squeeze equities out of short-sold positions because Wallstreet has not been equipped for such a movement and did not covert developments in the forum. This vulnerability will have likely been assessed by financial institutions and hedged against for the future, for example, through incorporating algorithms that track *Wallstreetbets* data.

8) Conclusion and future investigations

Through exploring the application of machine learning models in the development of an algorithmic trading strategy using solely Reddit data it appears promising that a profitable strategy can be developed using the provided data. However, the optimized active trading strategy did not outperform a passive position in the test sets and did not clearly surpass the financial returns of a random buy/hold/sell decision for the underlying *Wallstreetbets* stocks. In a future investigation, it may be fruitful to examine Reddit data as predictors in coordination with further non-Reddit specific predictive variables like macroeconomic considerations or financial fundamentals like share volatility. This consideration may also include further creative features, like Lycosa et al.'s (2021) ratio of quantity of *Wallsteetbets* posts over Google searches, that could also assist in the development of a superior active trading strategy.

A key discovery is derived from the performance comparison of the *Wallsteetbets* stocks with the indices in which these trade in. A passive, but also an active, position in both NOK and BBY have significantly outperformed the respective indices in which the equities trade in. This is indicative of the correlation assessment that found post quantity to be correlated with share price. Hence, even if the performance of an active trading strategy remains ambiguous, this paper detected sufficient evidence to allege that investing in the stocks discussed in *Wallstreetbets* displays potential for supernormal returns when compared to traditional investment tools like a market ETF. In that context, future studies might centre on causality versus correlation tasks to validate this hypothesis and thus challenge Valle-Cruz et al. (2021) earlier cited reverse causality argument for correlations between social media and sentiment. Whereas this report has verified a correlation between share price and quantity of posts, the combined dataframe for the 8 tickers in the training set displayed no correlation between share price and sentiment. However, when examining the correlations on an asset basis, the ticker CLOV had a sentiment correlation of +0.39 and only had its first major peak in June 2021, whereas most other studied assets had their first major peak in the first quarter of 2021, and correspondingly no, or even a negative, correlation between sentiment and share price. Therefore, it is possible that sentiment remains a helpful predictive variable for the initial emergence of a herd movement as it could be indicative of momentum building and a signal for an incoming increase in post quantity. Combining this thesis with the examination of the portfolio developments of the NOK and BBBY test

sets in appendix 5 and 7, respectively, it becomes evident that financial returns are driven by sudden share price changes as opposed to slow continuous growth. This indicates that it is recommendable to conduct an event study that is focused on identifying and predicting major pricing events as opposed to continuous price changes. This could in turn be converted into a trading strategy that outperforms the underlying active strategy. The greatest limitation for this study and future investigations remains the quantity of data available, which contest the informative value of such studies. Future studies will continue to face this hurdle as *Wallstreetbets* is a relatively new phenomenon.

9) References

Amin, A., Hossain, I., Akther, A. and Alam, K., 2019. Bengali VADER: A Sentiment Analysis Approach Using Modified VADER. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE).

Araújo, M., Gonçalves, P., Benevenuto, F. and Cha, M., 2013. Comparing and combining sentiment analysis methods. Proceedings of the first ACM conference on Online social networks - COSN '13,.

Baron, M., Brogaard, J. and Kirilenko, A., 2014. Risk and Return in High Frequency Trading. SSRN Electronic Journal,.

Birla, R., 2012. Determinants of the Success of Active vs. Passive Investment Strategy. [online] Core.ac.uk. Available at: <https://core.ac.uk/download/pdf/230538710.pdf> [Accessed 12 August 2021].

Brownlee, J., 2014. Why Aren't My Results As Good As I Thought? You're Probably Overfitting. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/arent-results-good-thought-youre-probably-overfitting/> [Accessed 8 August 2021].

Buz, T. and Melo, G., 2021. Should You Take Investment Advice From WallStreetBets? A Data-Driven Approach.

Chan, E., 2009. Harvard Business School Confidential. John Wiley & Sons (Asia) Pte. Ltd., p.185.

Damodaran, A., 2019. Price patterns, charts and technical analysis: The momentum studies. New York University.

Doorn, P., 2021. We put AMC, GameStop and other meme stocks' numbers to the test — here's which ones came out on top. [online] MarketWatch. Available at: <https://www.marketwatch.com/story/we-put-these-eight-meme-stocks-through-a-rugged-analytical-test-which-are-poised-for-growth-and-which-have-big-downsides-11622810160> [Accessed 1 September 2021].

Fama, E., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), p.383.

Gupta, A., 2021. Wall Street vs r/wallstreetbets: Exploring the Predictive Power of Retail Investors on Equity Prices. [online] Web.stanford.edu. Available at: <https://web.stanford.edu/class/cs224n/reports/final_reports/report010.pdf> [Accessed 14 August 2021].

Graham, J., 1999. Herding among Investment Newsletters: Theory and Evidence. The Journal of Finance, 54(1), pp.237-268.

Hirshleifer, D. and Shumway, T., 2003. Good Day Sunshine: Stock Returns and the Weather. The Journal of Finance, 58(3), pp.1009-1032.

Kahn, J., 2021. The 'stonks' market caught the A.I. algorithms off-guard, too. [online] Fortune. Available at: <https://fortune.com/2021/02/11/stonks-stock-market-gamestop-reddit-wallstreetbets-ai-hedge-funds-losses-gme-amc/> [Accessed 7 August 2021].

Leung, W., Wong, G. and Wong, W., 2021. Social-Media Sentiment, Portfolio Complexity, and Stock Returns. Cardiff University,.

Li, D., Wang, Y., Madden, A., Ding, Y., Tang, J., Sun, G., Zhang, N. and Zhou, E., 2019. Analyzing stock market trends using social media user moods and social influence. Journal of the Association for Information Science and Technology, 70(9), pp.1000-1013.

Lyócsa, Š., Baumöhl, E., Výrost, T.(2021). YOLO trading: Riding with the herd during the GameStop episode. ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg.

Mogyorosi, M., 2021. Sentiment Analysis: First Steps With Python's NLTK Library. [online] Realpython.com. Available at: <https://realpython.com/python-nltk-sentiment-analysis/> [Accessed 8 August 2021].

Ma, Y., 2020. NLP: How does NLTK.Vader Calculate Sentiment?. [online] Medium. Available at: <https://medium.com/ro-codes/nlp-how-does-nltk-vader-calculate-sentiment-6c32d0f5046b> [Accessed 8 August 2021].

McEnery, T., 2021. WallStreetBets is dying, long live the WallStreetBets movement. [online] MarketWatch. Available at: <https://www.marketwatch.com/story/wallstreetbets-is-dying-long-live-the-wallstreetbets-movement-11624714750> [Accessed 5 August 2021].

Nann, S., 2019. Predictive Analytics on Emotional Data Mined from Digital Social Networks with a Focus on Financial Markets. Inaugural Dissertation for the Attainment of the Doctorate from the Faculty of Management, Economics and Social Sciences at the University of Cologne,.

Pineiro-Chousa, J., Vizcaíno-González, M. and Pérez-Pico, A., 2016. Influence of Social Media over the Stock Market. Psychology & Marketing, 34(1), pp.101-108.

Ponczek, S., 2021. There are still 16 meme stocks with at least 100% gains in 2021 - BNN Bloomberg. [online] BNN Bloomber. Available at: <https://www.bnnbloomberg.ca/there-are-still-16-meme-stocks-with-at-least-100-gains-in-2021-1.1559061> [Accessed 14 July 2021].

Richardson, L., 1970. Do High Risks Lead to High Returns?. Financial Analysts Journal, 26(2), pp.88-99.

Schroeder, S., 2021. Hedge fund Melvin Capital is down $4.5 billion after epic squeeze by Reddit traders, report says. [online] Mashable. Available at: <https://mashable.com/article/melvin-capital-loss> [Accessed 16 August 2021].

Sonnemaker, T., 2021. Reddit day traders wanted to beat Wall Street to prove the system is rigged. Instead, they did it by losing.. [online] Business Insider. Available at: <https://www.businessinsider.com/reddit-wallstreetbets-traders-failed-war-wall-street-proves-system-rigged-2021-2> [Accessed 2 August 2021].

Sisk, J. 2013. Methods and systems for predicting market behavior based on news and sentiment analysis. Thomson Reuters Global Resources.

Sosnoff, M., 1966. Hedge Fund Management: A New Respectability for Short Selling. Financial Analysts Journal, 22(4), pp.105-108.

Stacey, K., 2021. GameStop mania: why Reddit traders are unlikely to face prosecution. [online] Ft.com. Available at: <https://www.ft.com/content/8caa3c75-944a-468e-8a68-9deeec8b67d8> [Accessed 21 August 2021].

Stibel, J., 2009. Why We Can’t Predict Financial Markets. [online] Harvard Business Review. Available at: <https://hbr.org/2009/01/why-we-cant-predict-financial> [Accessed 3 August 2021].

Verheggen, R., 2017. The rise of Algorithmic Trading and its effects on Return Dispersion and Market Predictability. [online] Arno.uvt.nl. Available at: <http://arno.uvt.nl/show.cgi?fid=145161> [Accessed 8 August 2021].

Wigby, M., 2021. What is a meme stock? Could Blackberry be the next GameStop or AMC?. [online] HITC. Available at: <https://www.hitc.com/en-gb/2021/06/03/what-is-a-meme-stock-could-blackberry-be-the-next-gamestop-or-amc/> [Accessed 13 July 2021].

Wollscheid, C., 2012. GRIN - Rise and Burst of the Dotcom Bubble. [online] Grin.com. Available at: <https://www.grin.com/document/197166> [Accessed 13 August 2021].

10) Appendix

Appendix 1: Data transformation procedure

The cleaned dataset scraped from Reddit entails the following features for each comment for each ticker:

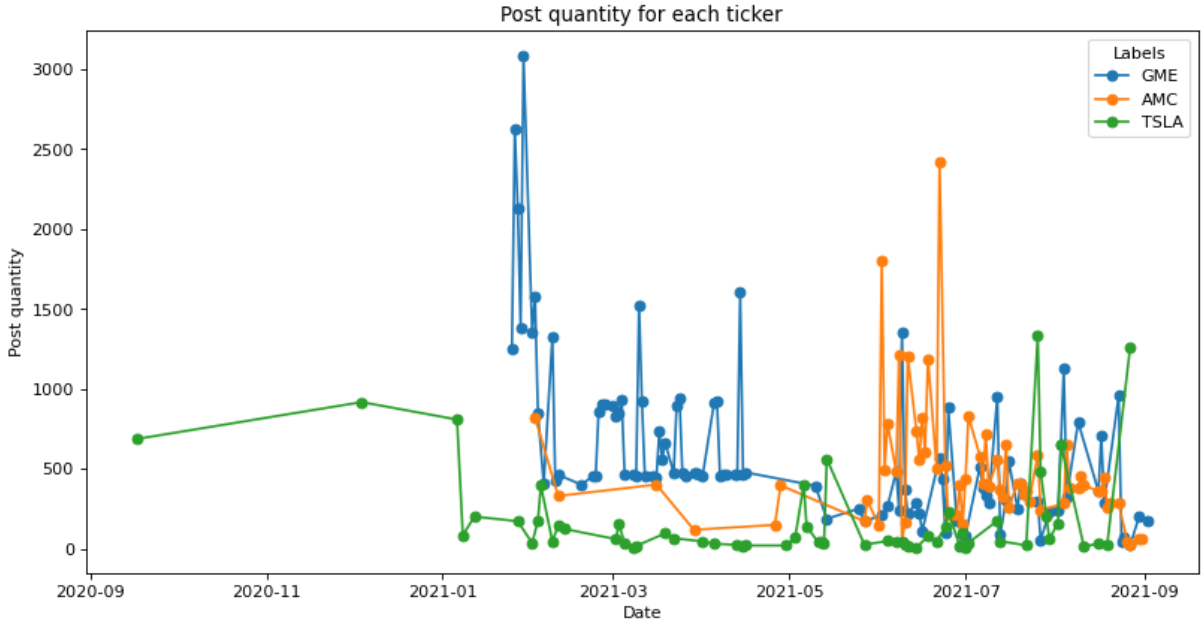
- Thread owner of each post
- Thread upvote ratio
- Cross posts
- Date
- Sentiment (positive, negative, neutral, compounded)

As to be able to utilize the data for machine learning purposes, following transformations have been made:

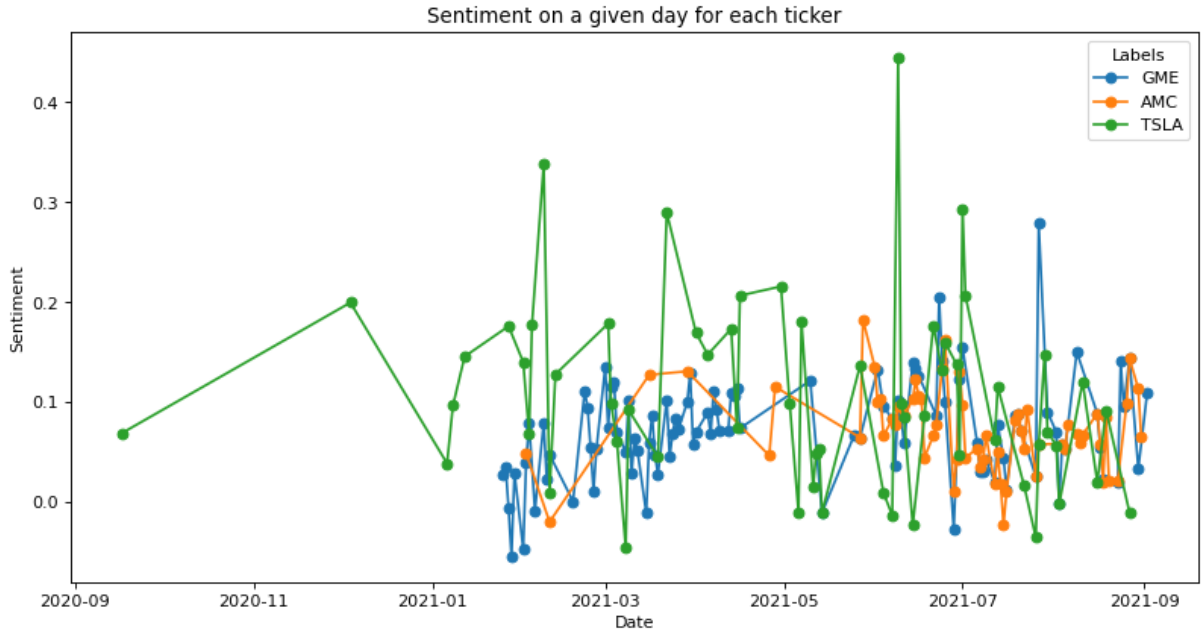
- Group by date and ticker, aggregate sentiment statistics with count and mean function. Attain value for:
    - Amount of comments per day for the 250 scraped posts
    - Average daily sentiment
    - Proportion of posts with positive sentiment
- Join each dataset with relevant ticker share price information incl. lagged share prices (already normalized prior), joined on date
- Combine 8 separate datasets into a large dataset and assign column to distinguish between tickers
    - As Lycosa et al. (2021) argue that the scale of which *Wallstreetbets* impacts shareprice is different for each ticker, the data of 8 training tickers have been combined into a large dataset to minimize the ticker-specific fluctuations
- Remove datapoints for which the last available share price is more than 3 days to prevent overfitting

Appendix 2: Summary statistics

The graph below shows the amount of posts scraped for each date for the sentiment analysis. The machine learning section (Section 6) further trims the data for occurrences beyond only the 1. January 2021.



The graph below shows the average normalized sentiment for each day for a given ticker.



Appendix 3: Correlations

SNDL

	Daily Sentiment	Proportion assessment	Share price	Post quantity
Daily Sentiment	1.000000	0.942342	-0.313034	-0.407988
Proportion assessment	0.942342	1.000000	-0.267507	-0.358756
Share price	-0.313034	-0.267507	1.000000	0.363082
Post quantity	-0.407988	-0.358756	0.363082	1.000000

GME

	Daily Sentiment	Proportion assessment	Share price	Post quantity
Daily Sentiment	1.000000	0.962012	0.056131	-0.313001
Proportion assessment	0.962012	1.000000	0.078730	-0.271459
Share price	0.056131	0.078730	1.000000	0.045975
Post quantity	-0.313001	-0.271459	0.045975	1.000000

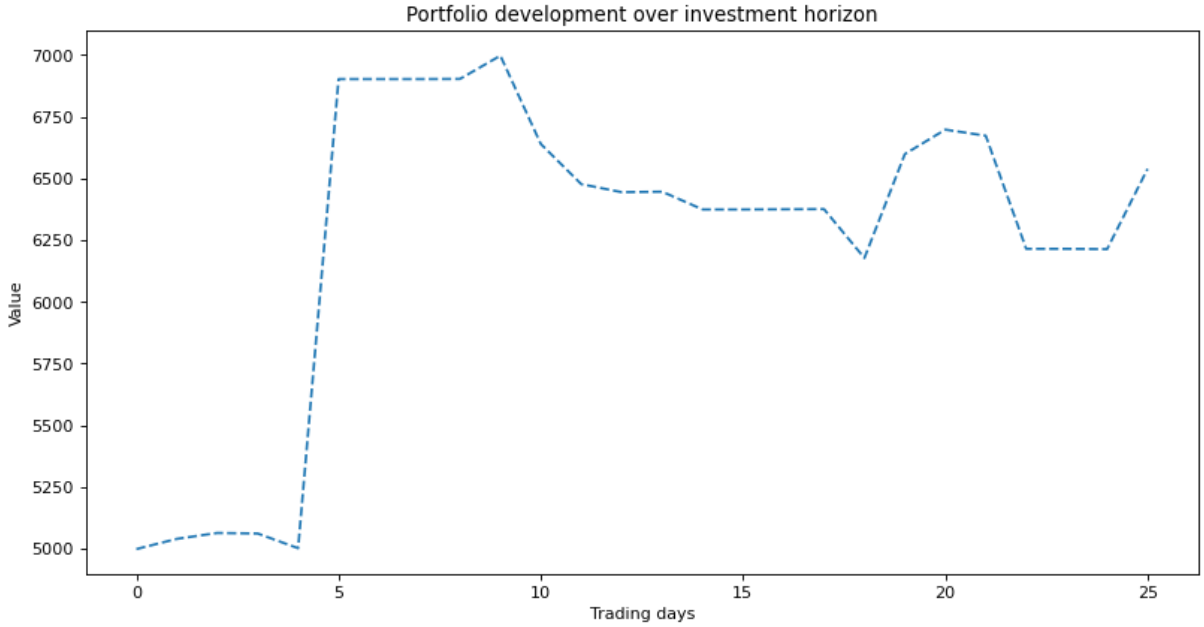
Appendix 4: Confusion matrix logistic regression

The table below shows the confusion matrix for a sample logistic regression with a validation split of 23% in the benchmark case. It becomes evident that the accuracy is driven by issuing mainly sell signs and thus almost mimicking the majority classifier in terms of performance.

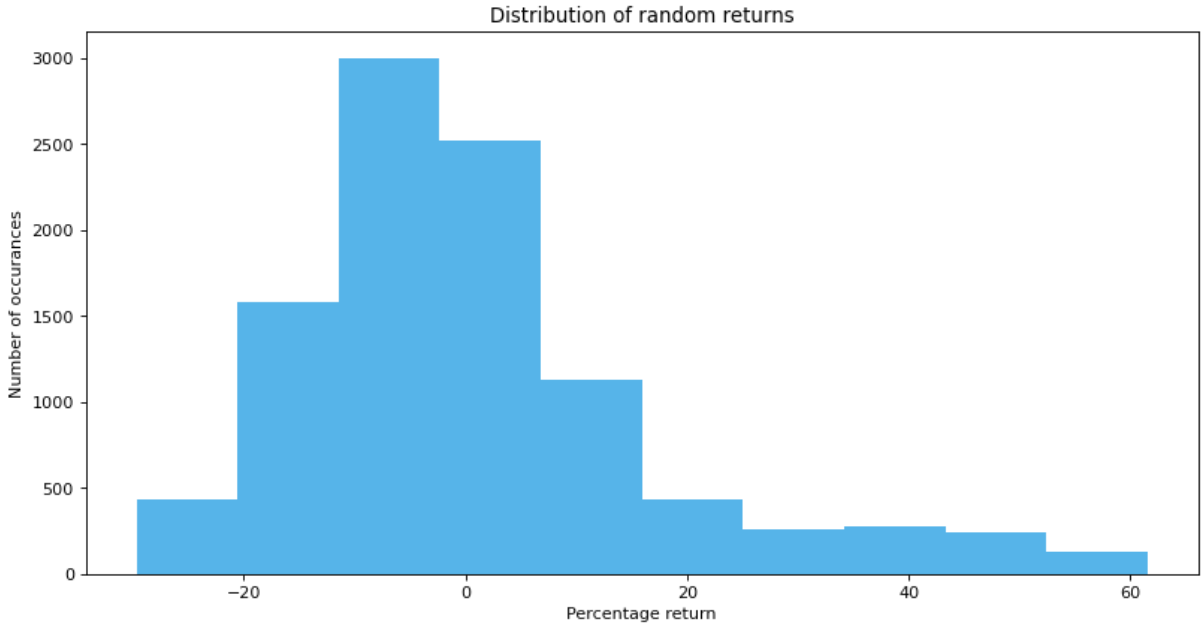
	0	1	2
0	0	0	8
1	0	1	22
2	0	0	27

Appendix 5: Portfolio developments in NOK test set

The trading simulation below uses a random forest model as the baseline model

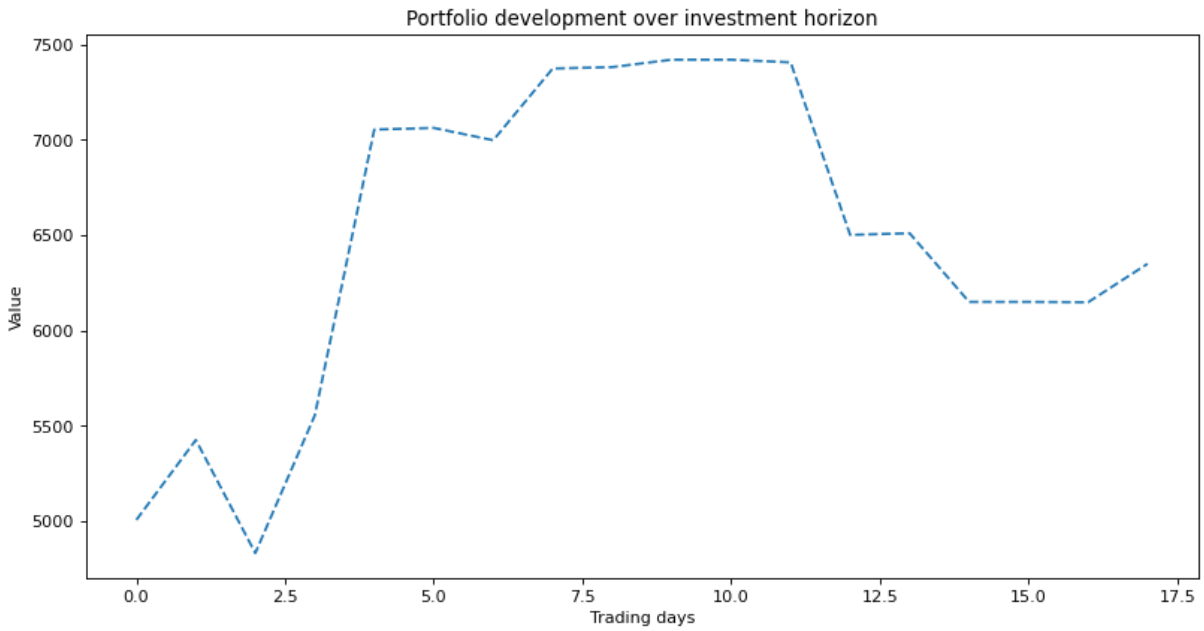


Appendix 6: Distribution of a random trading decision with the NOK test set

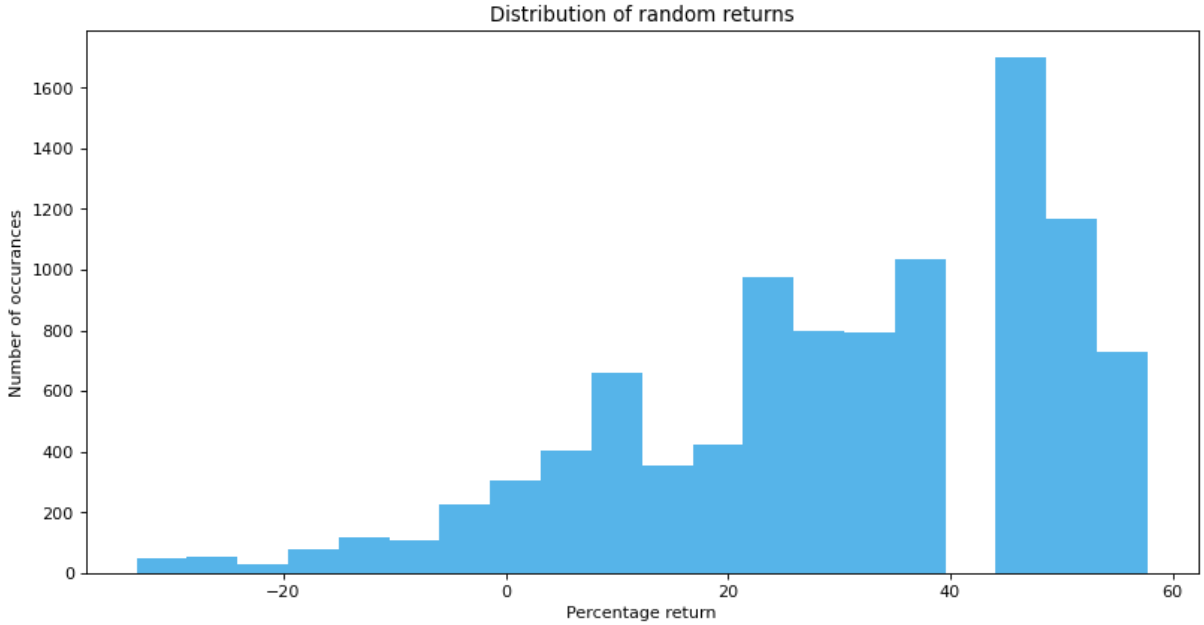


Appendix 7: Portfolio developments in BBBY test set

The trading simulation below uses a random forest model as the baseline model



Appendix 8: Distribution of a random trading decision with the BBBY test set



7.2.3) Ambigious comments

Whereas utilization of VADER has provided helpful for the sentiment analysis, there are certain sentimental features that it lacked to capture. Firstly, by nature of comments on reddit many tent to be sarcastic and meant to provocate others, even if the core message is positive. This may also include insult, which are predominantly interpreted as negative sentiment but might truly be positive. The opposite is less likely true, where positive words are used to describe a negative core message. As a result, it may be argued that there is a bias in the daily sentiment feature, therewith reducing the value from what it truly should be.

7.2.4) Equity impact on predictor sensitivity

Daily sentiment data has been collected for several equities and combined in one dataframe. In other words, the assumption has been made that predictor variables have the same impact on the independent variable *percentage change.* This has been done to minimze the limitation of data quantity, where a large dataset is beneficial for the underlying machine learning model. The assumption may not hold as it is possible that the true predictive impact of a certain feature might be different for two equities.