LINC

LUND UNIVERSITY FINANCE SOCIETY EST 1991

TRADING & QUANTITATIVE RESEARCH REPORT

Sentiment analysis

Using sentiment analysis to predict movement in equities

In collaboration with:

LYNX

Analysts: Niclas Wölner-Hanssen, Duy Pham-Nguyen & Hannes Brinklert

Introduction & Background

Social media, for instance, Instagram, Facebook, Twitter, and Reddit, is a big part of our everyday life. People post information regarding their life, opinions, and investments. On the 28th of January 2021, the pre-market value of GameStop's stock on the New York Stock Exchange was over 500 US dollars per share, while the share's valuation was 17.25 US dollars at the beginning of January. The reason behind this was a short squeeze[1]. This is when an abundance of short selling causes an undersupply and an overdemand for stock, resulting in a rapid price increase[2]. The short squeeze was caused by people on the internet forum "wallstreetsbets" on Reddit, where users are known to discuss stocks that are off[3]. This event demonstrates social media's impact on the stock market and its relevance when investing.

In collaboration with Lynx Asset Management, a hedge fund based in Stockholm, this paper will investigate whether text from Twitter and Reddit could indicate trading signals which result in successful trades. This will be done by sentiment analysis on the text using the Valence Aware Dictionary and sEntiment Reasoner (VADER), an open-source Python library on GitHub[4]. Ranking the polarity of a text, whether positive (1), negative (-1), or neutral (0), is a fundamental task of sentiment analysis[5]. By grouping posts from social media submitted, for instance, during the same hour, and calculating the mean of the sentiment score, one could determine if people are talking positively about, for example, Amazon's stock.

Background

Machine learning is a branch of artificial intelligence and computer science that focuses on using data and algorithms to imitate the way that humans learn; the goal is to gradually improve its accuracy of human behaviour. Machine learning models have the advantage over humans in that they can iterate over a lot of information in a short period of time, such as analysing a large number of texts and making predictions of the aggregated sentiment on these texts.

Machines are also objective in the sense that they do not rely on one person's unique experience from the past while formulating trading strategies. Therefore, machines have the ability to create trading strategies that are less likely to be overfitted to historical events.

Sentiment analysis

Sentiment analysis is a machine learning tool that analyses texts and gives polarity, from positive to negative. Sentiment analysis models are trained by algorithms to read beyond mere definitions to understand a text just as a human would. In python, there are natural language different libraries with trained machines e.g VADER[6].

Vader

VADER (Valence Aware Dictionary for sEntiment Reasoning) is a model used for text sentiment analysis. It detects polarity and the intensity of emotion within a text. The sentiment score of a text is obtained by summing up the intensity of each word in the text. VADER scores an individual word between -4 and 4, then sums the sentence, after which normalisation is applied to map the value between -1 and 1 as in Figure 1. The VADER sentiment works best for short text, like social media text[7]. For this reason, the sentiment model was chosen as the most eligible one to analyse Reddit comments

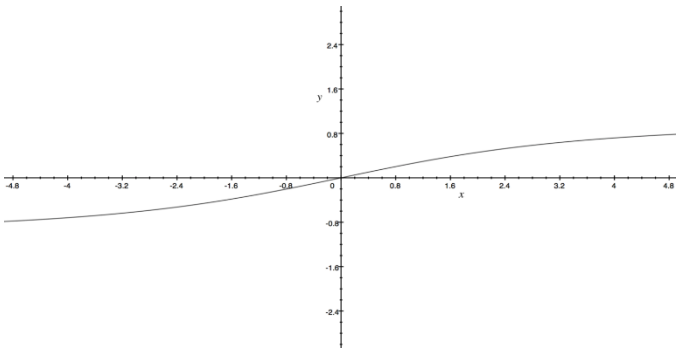


Figure 1: VADER normalization graph

Hypothesis

One of the efficient market hypothesis form states that profiting from predicting price movement is challenging. The main cause behind share price movement is the arrival of new information. A market is "efficient" if all information is already incorporated into prices. So, there is no way to "beat" the market because no undervalued or overvalued securities are available[8]. The EMH states that the market mirrors "all available information in the market". Financial researchers interpret "all available information" in three different forms, weak form, semi-strong form, and strong form by Investopedia. The weak form of EMH asserts that future prices cannot be fully predicted by analysing prices from the past. The semi-strong form suggests that current stock prices react quickly to the release of all new public information, and the strong form of EMH states that the current prices fully reflect all public and private information[9].

In the paper, "Can social microblogging be used to forecast intraday exchange rate?", Papaioannou et. al found that information from microblogging platforms such as Twitter

Background & Data

can enhance the forecasting efficiency of intraday exchange rates. Thus, the paper supports the semi-strong form of the EMH, as the analysis of public discussion on Twitter contradicts the strong form of the EMH [10]. Hence, the hypothesis for this paper is that social media sentiment data is expected to show abnormal returns in the equity market when used as a trading signal.

Scope

The trading strategies based on sentiment analysis conducted in this paper are aimed towards the technology sector within the S&P 500 index between 2015/01/01 to 2021/11/27. The stocks used are in the Technology Select Sector SPDR Fund (XLK). They were chosen because of the growth of the industry, its long-term drivers for revenues and profits, and due to it being the largest sector in the S&P 500 [11]. XLK is an Exchange Traded Fund (ETF) that follows the S&P 500 technology stocks [12]. The strategic approach is straightforward, going long on stocks with positive sentiment and short on negative ones. The choice of a market neutral strategy is to be exposed to company-specific variation only and in turn concentrate the variation of the portfolio to the sentiment signal. The trading period is between 2016/02/19 to 2021/11/02, and trading takes place daily for the entire period for the Bloomberg Twitter data, but only for one year for the Reddit data.

Data

The data collected in this research consists of historical comments from the "daily discussion" thread on the subforum "wallstreetbets" on Reddit, Bloomberg Twitter sentiment data, and tweets mined from specific trader accounts using the Twitter API. The Reddit data is between 2020/11/23 to 2021/11/02, the Bloomberg Twitter sentiment and the Twitter API tweets are from 2015/01/01 to 2021/11/02. The equity prices used are from Bloomberg and are from 2015/01/01 to 2021/11/02. They consist of daily open, high, low, and close prices.

As an important note, sentiment analysis may misinterpret text. There are certain conditions in which it performs poorly. The most common one is sarcasm, where people use positive words to express a negative sentiment. Irony occurs often in user-generated content such as tweets or Reddit comments. An example of faulty sentiment that is stock-related could be: "The stock is down 10%, that's great!" It could be hard for a human being to understand, let alone for a machine.

Reddit comments on "wallstreetbets"

The Reddit data was gathered by using a package called Python Reddit API Wrapper (PRAW) from the "wallstreetbets" subreddit. After receiving the desired data, it consisted of some faulty noise. For example, when checking for stock tickers "IT", "V", or "NOW", it would download a sentence consisting of that word. Another faulty noise was "APHA", this stock doesn't exist in the XLK but since it's similar to "APH" it was also downloaded. For the first problem, those stocks needed to have a dollar sign in front of them. For the second problem, in order to not receive stocks, for instance, "APHA", they were adjusted to have a space before and after the stock ticker. Each comment was then assigned a sentiment score (positive, neutral or negative) using VADER.

According to subreddit statistics for wallstreetbets [13], the number of subscribers to the subreddit was close to 11 million in October 2021. In contrast, this number was 1.5 million only one year earlier and 670 thousand two years earlier. Given that the accumulated buying power grows with the number of subscribers and how the sentiment of this crowd impacts the stock price, the potential signal today is fundamentally different from just two years ago.

Due to the short period of historical data, trading strategies need to be formulated on higher time frequencies such as intraday to have sufficient track record length for the backtest. As a result, transaction costs will have a more significant impact on the results. While using longer holding periods, such as weekly ones, could reduce this impact, this would result in a shorter track record. Furthermore, the sentiment for holding for a week is also weaker than the intraday signal. Therefore, since transaction cost is not considered in this paper, the daily trading strategy was chosen.

Bloomberg Twitter sentiment

Daily sentiment on tweets regarding stocks in XLK was gathered from Bloomberg. The data presents the average, minimum, and maximum sentiment as well as the count of positive, negative, and neutral tweets. The sentiment on the tweets is available 10 minutes before the markets open and it represents data 24 hours back. Bloomberg's sentiment is based on Stocktvist and tweets which contain the company's name and cashtag, for instance, Apple and $AAPL[14]. Stocktvist is a social network aimed towards traders and investors[15].

Data & Method

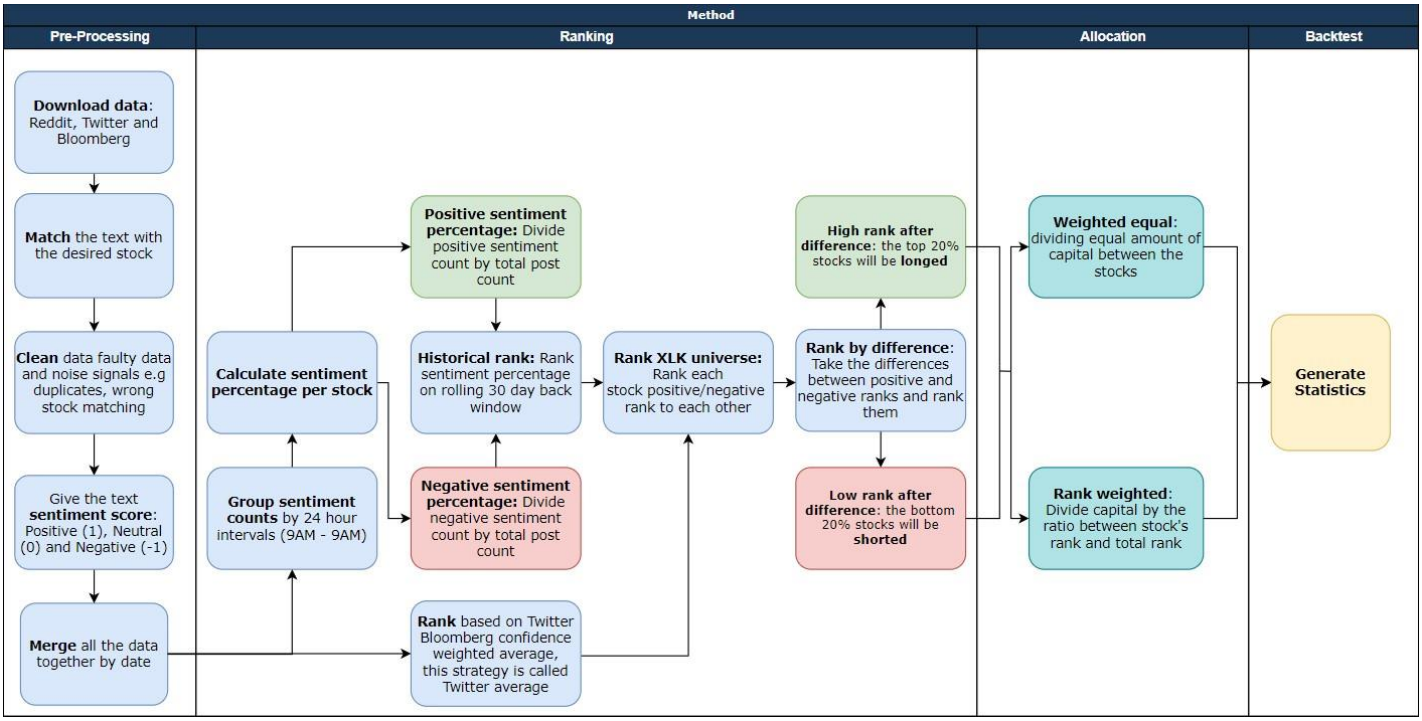


Figure 2: Flowchart of the method

To ensure that the sources are reliable, Stocktvist has official accounts which are accounts that Stocktvist considers professional within the industry[16]. The sentiment is calculated using a machine learning model called support vector machine (SVM).

Tweets mined with Twitter API

To receive historical data from the Twitter API the premium subscription was used. The accounts in Appendix 1 are used to scrape data from Twitter. The query sent to the search-full-archive request on the Twitter API is for example "from:fusionptcapital (AAPL OR MSFT OR NVDA OR V OR ...)". Tweets were gathered between 2015/01/02 and 2021/11/28. The accounts were chosen based on how much they were mentioned on financial news websites, then they were checked based on number of followers and lastly how many tweets they had posted, they were picked intuitively and based on how qualitive they seemed. Afterward, the tweets were processed to be presented in a better manner. Each tweet was linked to one or more tickers, if the ticker or tickers were mentioned in the tweet. This was done to every ticker in XLK except the tickers in Appendix 2. These tickers have to be mentioned with a cashtag after a space afterward, e.g "$V ". This is because they are or a part of common words and if the cashtag was not used, tweets were linked to the above tickers which did not mention the ticker. Each tweet was then assigned a sentiment score using VADER.

Method

As in Figure two, the data was firstly gathered from Reddit, Bloomberg, and Twitter, and then cleaned as described under the "data" section. In this process, the tweets mined with the Twitter API were found to be too noisy and thus after cleaning resulted in too few tweets to implement the constructed trading strategy in, and therefore were excluded from further steps. A Python script was developed to read all the data and execute the trading strategies. The aim of the trading strategies was to formulate market-neutral positions.

Stock sentiment ranking and trading strategy

For the Reddit data, the sentiment is collected on posts from the previous day at 9 AM and then 24 hours forward to match the daily Bloomberg Twitter sentiment data. These are then aggregated to a total count of posts related to each stock. The positive classified number of posts is then divided by the total number to create a positive sentiment percentage of posts for each stock. This percentage is then ranked in a rolling window 30 days back, to receive a historical context of today's positive percentage in relation to 30 days back. These historical rank values are then ranked between the stocks in the XLK to quantify the relation between the stocks. The same procedure is then made for negative sentiment posts to create a negative percentage of the total counts.

Method

To get a more accurate rank for the stocks, the negative sentiment ranks are subtracted from the positive sentiment ranks, and the resulting differences are then ranked. This is so that a stock with high positive ranking and small negative ranking is ranked higher than a stock with low positive ranking and high negative ranking. This subtraction is also utilized to account for situations when a stock has high positive ranking while also having a high negative ranking which creates ambiguity in direction of sentiment. This situation could arise due to a relatively small increase in the number of total tweets and reddit posts. This strategy is called "Reddit-, and Twitter Counts" depending on the data used.

The Bloomberg Twitter confidence-weighted-average sentiment score is also utilized as another strategy [17]. For this strategy no historical comparison for a single stock is made as this relationship is already captured in the confidence-weighted-average metric. The ranks are only made between stocks everyday where the highest score receives the highest rank and the lowest score receives the lowest rank. This strategy is called "Twitter Average".

Allocation strategies

This paper uses two different allocation strategies. Both allocate the same amount of capital on each basket After ranking by difference as in Figure 2 the upper 20% ranked stocks are placed in a basket that goes long and the lower 20% ranked stocks are shorted. What differs the two allocation strategies is how the allocation of capital is done within the baskets. The first one is "equal-weighted", which means that in each basket the capital is divided equally between the stocks

The second one is a "rank-weighted" allocation, which does a new ranking within the long and short baskets (according to the sentiment ranking), and then allocates capital according to the ratio between the rank and the sum of the total ranks. However, in the case of the short basket, some adjustment has to be made since the stock with the lowest rank should be allocated the most capital. This is due to the fact that this stock has the least number of positive sentiment posts or the smallest average sentiment on the posts and this stock's price is predicted to decrease the most. The new calculated ranks are multiplied by -1 and then a new ranking is calculated based on the negative ranks. See Figure 2, which clearly shows the allocation and trading strategies in the form of a flowchart. Lastly, a backtesting code was written to receive statistics of the trading and allocation strategies.

Results

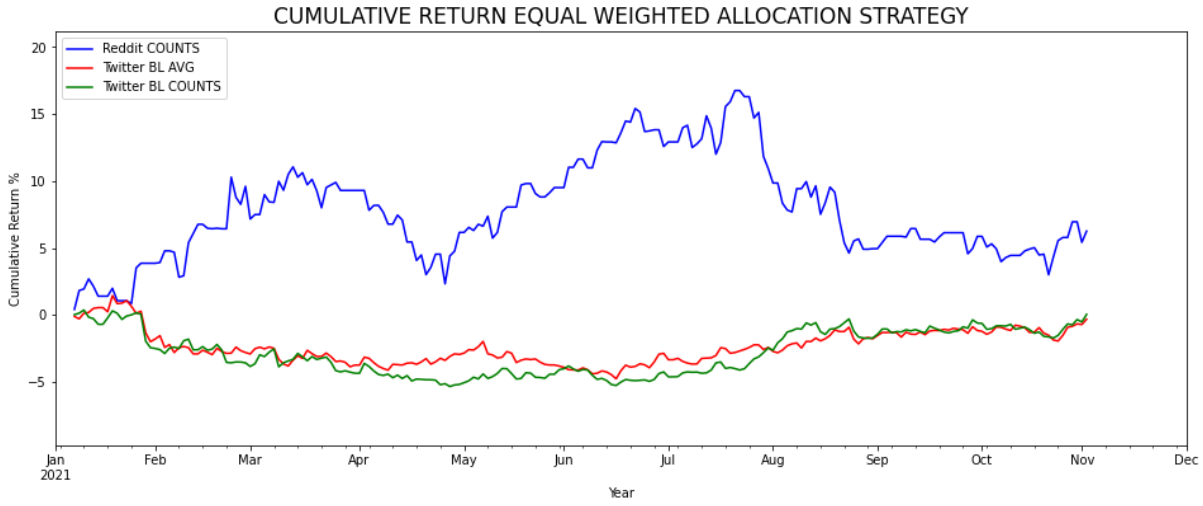


Figure 3: Equal weighted cumulative returns 2021/01/13 – 2021/11/02

Equal weighted	CAGR	Ann. Vol	Cumulative Ret.	MDD	Sharpe Ratio	Sortino Ratio	Calmar Ratio
Reddit COUNTS	7.34%	13.64%	6.25%	-11.78%	0.59	0.85	0.62
Twitter BL AVG	-0.38%	5.1%	-0.33%	-6.12%	-0.05	-0.07	-0.06
Twitter BL COUNTS	0.05%	5.26%	0.04%	-5.69%	0.03	0.04	0.01

Table 1: Metrics for equal weighted cumulative returns 2021/01/13 – 2021/11/02

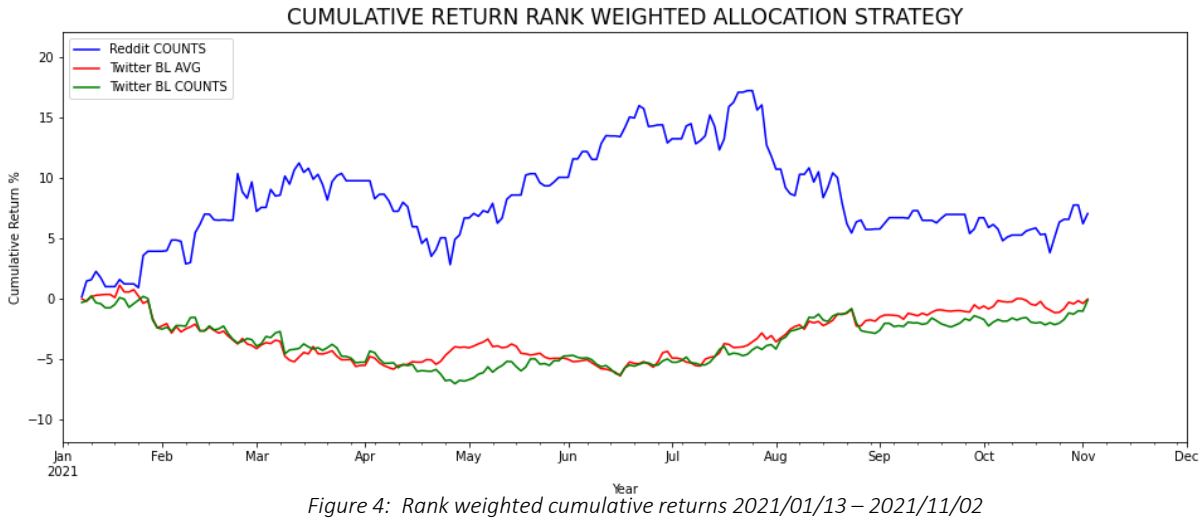


Figure 4: Rank weighted cumulative returns 2021/01/13 – 2021/11/02

Rank weighted	CAGR	Ann. Vol	Cumulative Ret.	MDD	Sharpe Ratio	Sortino Ratio	Calmar Ratio
Reddit COUNTS	8.28%	13.63%	7.05%	-11.43%	0.65	0.94	0.72
Twitter BL AVG	-0.04%	5.59%	-0.03%	-7.42%	0.02	0.03	-0.01
Twitter BL COUNTS	-0.13%	6.09%	-0.11%	-7.25%	0.01	0.01	-0.02

Table 2: Metrics for rank weighted cumulative returns 2021/01/13 – 2021/11/02

Results

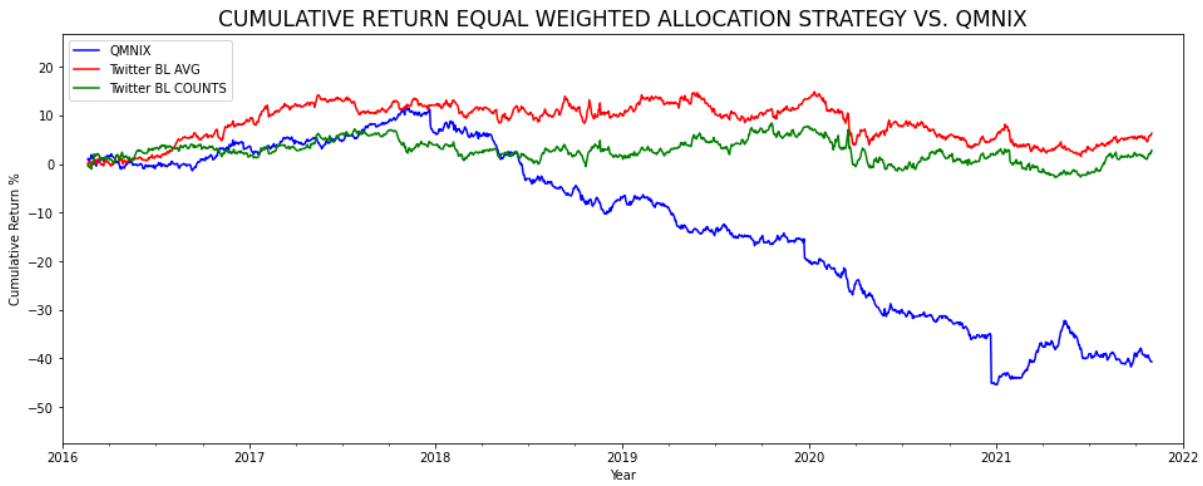


Figure 5: Equal weighted cumulative returns 2016/02/19 – 2021/11/02

Equal weighted	CAGR	Ann. Vol	Cumulative Ret.	MDD	Sharpe Ratio	Sortino Ratio	Calmar Ratio
QMNIX	-8.4%	9.73%	-40.69%	-51.16%	-0.85	-0.98	-0.16
Twitter BL AVG	1.03%	5.44%	6.3%	-11.54%	0.22	0.3	0.09
Twitter BL COUNTS	0.46%	5.61%	2.78%	-10.28%	0.11	0.15	0.04

Table 3: Metrics for equal weighted cumulative returns 2016/02/19 – 2021/11/02

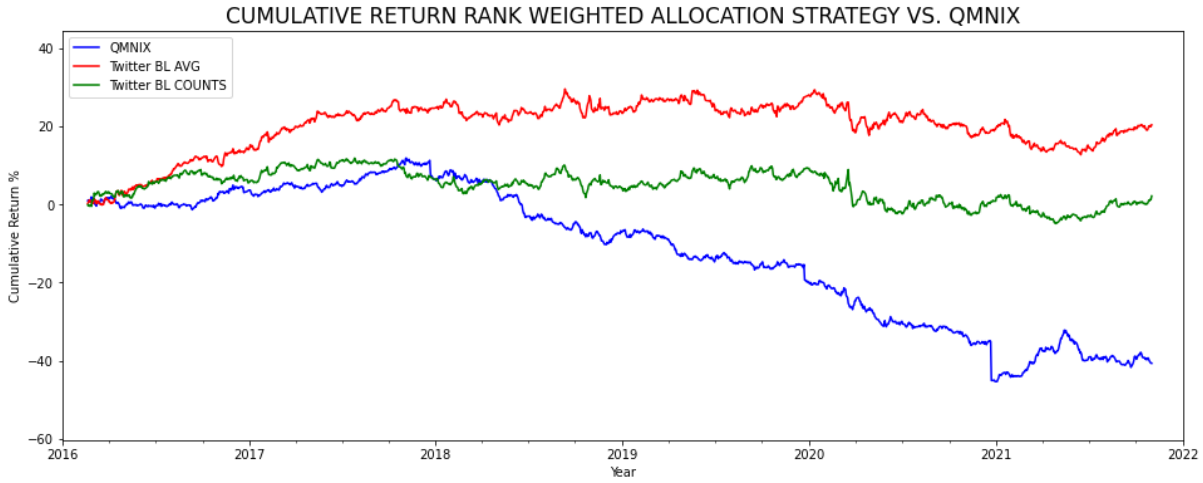


Figure 6: Rank weighted cumulative returns 2016/02/19 – 2021/11/02

Rank weighted	CAGR	Ann. Vol	Cumulative Ret.	MDD	Sharpe Ratio	Sortino Ratio	Calmar Ratio
QMNIX	-8.4%	9.73%	-40.69%	-51.16%	-0.85	-0.98	-0.16
Twitter BL AVG	3.16%	6.08%	20.35%	-13.0%	0.54	0.78	0.24
Twitter BL COUNTS	0.36%	6.58%	2.13%	-14.99%	0.09	0.12	0.02

Table 4: Metrics for rank weighted cumulative returns 2016/02/19 – 2021/11/02

Analysis

Analysis

From Figure 3 and 4 by comparing Reddit against the Twitter Bloomberg portfolio and based on the Compound Annual Growth Rate (CAGR) and cumulative return, the Reddit portfolio is performing much better in the timeframe of one year. In both figures, the Reddit Counts' cumulative return is positive. In the case of the Twitter features in both figures, the cumulative return is below 0 during the majority of the period. When investigating the volatility all the models seem to be relatively low. The low volatility comes from the fact that the trading strategy that is implemented is a market-neutral one, which means that it seeks to avoid some form of market risk by hedging.

Comparing the equal-weighted and rank-weighted allocation strategies by looking at the cumulative return in Tables 1 and 2, the Twitter trading strategies perform almost equally. The Twitter Average's cumulative return is higher in the rank weighted allocation strategy, in contrast to Twitter Counts which perform better with equal weighted strategy. This could also be seen in the Sharpe ratios. The ratio is higher in rank weighted strategy for Twitter Average, and for Twitter Counts the sharpe ratio is higher in the equal weighted strategy. Twitter Average's cumulative return for the long and short positions for both allocation strategies is seen in Appendix 3 and 8, the short and the long position's cumulative return is mirrored around the summarized return, the blue line, the long position results in a positive return and short position in negative return.

This means that the gains from the long position are removed due to the short position's loss, which is why the summarized return is almost a horizontal line. This is the case for Twitter Counts as well, which can be seen in Appendix 5 and 7. However, the Reddit Counts strategy is performing equally in Table 1 and 2. The cumulative returns are 6.25% and 7.05% for equal and rank weighted respectively and the sharpe ratios are 0.59 and 0.65.

A comparison between Reddit Counts' cumulative return for short and long positions can be seen in, Appendix 4 and 9. For both up to April, the short and long positions perform almost equally, and after April the short position's cumulative return declines, while the long position's return increases. All this shows that the Reddit Counts trading strategy performs equally well for both allocation strategies.

This indicates that none of the two allocation strategy is better than the other one since all trading strategies did not perform better in one allocation strategy during the one year period. During the one year period, the average number of stocks for Twitter Average for both allocation strategies in the long and short basket, is 15 and 20.9, Appendix 6 and 10. In Twitter Counts case it is 15.2 for both baskets. However, for Reddit Counts it is 1.4 and 1.04 for both allocation strategies. This could be the reason behind why the Reddit Counts outperform the other trading strategies since it only trades on one stock in each basket.

In Figure 5 and 6, it is seen that both Twitter Counts and Twitter Average with the two different allocation strategies beat AQR Equity Market Neutral Fund(QMNIX) in cumulative return over a six year period. The cumulative return could also be seen in Table 3 and 4, where the trading strategies out performs the QMNIX, and have positive returns. In both allocation strategies the trading strategies have a higher sharpe ratio and CAGR. Comparing the two allocation strategies for Twitter Average, the Twitter Average performs better in the rank weighted strategy than the equal weighted one with regard to CAGR, cumulative return and sharpe ratio. The Twitter Counts performs better in the equal weighted strategy compared to the rank weighted strategy due to the fact that the CAGR is around 28% higher and the cumulative return is 31% higher in the equal weighted strategy. Further on, the cumulative return is 223% higher in the rank weighted strategy compared to the equal weighted strategy for the trading strategy of Twitter Average. One possible reason behind this could be the fact that more capital is allocated to stocks which increases more in price, if it is a long position, compared to the equal weighted one which allocates the same capital for every stock in the basket. Due to this, the rank weighted allocation strategy could be considered better over the six year period.

The performance of the short and long positions over the six year period can be seen in Appendix 11 to 4. In both allocation strategies for both trading strategies the long positions result in a positive cumulative return of at least 400%, while the short positions result in a negative cumulative return of approximately -100%. The large return of the long positions is not odd due to the fact that it has been an up trend market the last few years.

In Table 3, Twitter Average has approximately 127% higher cumulative return compared to Twitter Counts and in alignment with that Twitter Average has approximately 124% higher CAGR than Twitter Counts, for the equal weighted allocation strategy.

Analysis & Discussion

Comparing the same two strategies in the case of the rank weighted allocation strategy in Table 4, Twitter Average has a cumulative return that is approximately 855% higher than Twitter Counts, and for the CAGR that number is approximately 778%. This shows that over a six period the Twitter Average performs better than Twitter Counts independent of the allocation strategy.

Discussion

Vader vs SVM

The method used on Reddit is based on the VADER sentiment as mentioned previously but the Twitter Bloomberg is using a machine learning model called SVM (support vector machine). Given training data, SVM classifies hyperplanes and when it finds the hyperplane it can differentiate the different types of classes, in this case, it would be positive, neutral, and negative sentiment [18]. The VADER sentiment is lexicon-based so it has a pre-assigned value for all the words that exist in the lexicon. Intuitively the SVM should give a more accurate sentiment than VADER given that it can learn and be trained. The sentiment dictionary VADER was the first-hand choice due to it being the most reliable one and of the feedback it had from other users, but it was not flawless. When analysing the Reddit data, it did quite a terrible job assigning the score to the Reddit comments seemed like a hard task for the sentiment model. This could be due to the language the Reddit user has. In the Appendix 16. There are some examples of comments where the sentiment does not seem correct, this might indicate that the VADER model does not understand what the tickers stands for or that the dictionary is not developed enough to do a good job. The things that could be improved is to add own words to the VADER dictionary or since it seems like there is a certain way the reddit users at wallstreetbets write, hence one could adapt the same method as the Bloomberg Twitter sentiment data which is to create a SVM (support vector machine) and after many samples of Reddit comments decide what the comments sentiment should be.

Efficient market hypothesis

Based on the results shown in this report and not considering transaction commission, the hypothesis of this project is supported. The Reddit data showed surprisingly better results than Twitter Bloomberg data within one year. When utilizing the whole 6 year data of Twitter Bloomberg and comparing it to the AQR market neutral fund the Twitter strategies are outperforming in all categories that are shown but here once again the commission fee is not accounted for.

The hypothesis was based on the existence of the semi-strong form of EMH and a conclusion could be made that it is true, with a different strategy, better returns could be achieved, more on this below in the further studies section.

Twitter

The data collected with the premium version of the Twitter API from the accounts was determined to be insufficient, due to a lack of data. The number of tweets that were collected was 131 328 and after the data was processed, 6 445 tweets were left over a six-year period. The chosen accounts did not mention the tickers in XLK as much as hoped. Further on, there were restrictions in the Twitter API. Firstly, there was a limit to the number of requests that could be used. A paid subscription plan of the premium Twitter API was used and the limit of the requests were 1000, with up to 500 tweets per request. Secondly, the enterprise version of the API supports searching for tweets by cashtag, which the premium version does not support. This resulted in tweets that contained words that are tickers in XLK being gathered, while not using the cashtag in the search query.

Strategies

The conclusion made in the analysis is that the Twitter Average performed better than Twitter Counts and the best allocation strategy was rank weighted over a six year period. Twitter Average with rank weighted allocation strategy has a CAGR of 3.14% compared to the market neutral fund QMNIX of -8.14%. The fact that it has been an uptrend market the last couple of years shows the performance of QMNIX is poor. However, the created market neutral strategy with a CAGR that outperforms QMNIX should be considered to be good. One thing to keep in mind is that transaction costs of trades have not been taken into consideration. This results in the cumulative return and CAGR appearing higher than they would be should transaction costs be accounted for. There are approximately 253 trading days in a year and count with a commission of 0.079%[19,20]. With six years of trading and approximately 15 stocks in the long and short baskets each day, it results in a cumulative commission of 3597.66%. This means that the created strategy would perform worse than QMNIX. The Reddit Counts strategy has on average one stock in the short and long baskets, for both allocation strategies. This strategy would not be applicable in real life trading since the risk would be tremendous and is essentially “betting” on one stock in each basket. Due to this, the other strategies with an approximate average of 15 stocks in each basket would be more suited for real life trading.

Discussion & Conclusion

Further studies

In this analysis trading signals were created based on the theory that texts on social media platforms influence market participants to buy and sell securities. A market neutral strategy might not be the best approach in the sense that a negative sentiment score has the same weight as a positive sentiment.

Although the theory seems reasonable, that social media sentiment influences market participants, these market participants might not have the ability to short securities to the same extent as their ability to buy securities. In addition, the disposition effect, that investors tend to sell winners and hold on to losers could be another argument that negative sentiment does not influence negative price movement to the same extent as positive sentiment on positive price movements. Therefore, a better approach might be to focus on positive sentiment only and hedge the market risk by shorting the corresponding sector ETF of each stock instead of shorting stocks with negative sentiment.

This report uses traditional measurements such as sharpe ratios to make inference between strategies. These statistics on financial returns are highly volatile and subject to significant sampling error which makes them insufficient to use as a tool to decide which strategy to allocate capital to out of sample. A better approach here is to make use of machine learning algorithms such as a Random Forest classifier to measure signal predictability between strategies on a training set and then make inference on a test set.

Further studies could also measure signal performance under different market conditions: How does the signal perform during bull and bear markets? How does the signal perform during low and high volatility periods? How does the signal perform conditionally on positive price momentum?

The tweets gathered with Twitter API were not used due to lack of data to the constructed trading strategies. To further analyse and make use of the tweets from the Twitter API one could construct a trading strategy that chooses stocks that are mentioned with a positive sentiment to initiate a long position on, and at the same shorting the rest of the stock in the XLK. This approach is not dependent on a massive amount of tweets.

Conclusion

In conclusion, market neutral strategies were created based on sentiment from social media, with different trading and allocation strategies. The one most applicable trading strategies are the one based on the sentiment on Twitter from Bloomberg, Twitter Counts and Twitter Average. This is due to the amount of data available and the strategies are not initiating positions on only one stock in the baskets, which the Reddit Counts trading strategy did. The Twitter Counts and Twitter Average outperformed QMNIX over a six year period with both allocation strategies, without transaction costs taken into consideration. This shows that a good performing market neutral strategy based on sentiment analysis can be created and a semi strong form of the efficient market hypothesis is supported.

Reference List

[1]: GameStop short squeeze, Wikipedia, https://en.wikipedia.org/wiki/GameStop_short_squeeze, accessed 2021-10-14 22:07

[2]: Short squeeze, Wikipedia, https://en.wikipedia.org/wiki/Short_squeeze, accessed 2021-10-19 22:44

[3]: GameStop short squeeze, Wikipedia, https://en.wikipedia.org/wiki/GameStop_short_squeeze, accessed 2021-10-14 22:07

[4]: vaderSentiment, https://github.com/cjhutto/vaderSentiment

[5]: Sentiment analysis, Wikipedia, https://en.wikipedia.org/wiki/Sentiment_analysis, accessed 2021-10-18 19:22

[6]: NLP, IBM, https://www.ibm.com/cloud/learn/machine-learning accessed 2021- 11-09-17:06

[7]: VADER, https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9 accessed 2021-11-09-17:08

[8]: EHM, Investopedia https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp accessed 2021-11-10 18:25

[9]: Efficiency Market, Investopedia, https://www.investopedia.com/terms/m/marketefficiency.asp accessed 2021-11-10 18:27

[10]: Can social microblogging be used to forecast intraday exchange rates?, Research gate, https://www.researchgate.net/publication/258082669_Can_social_microblogging_be_used_to_forecast_intraday_exchange_rates accessed 2021-11-10 18:28

[11]: Technology sector, https://www.schwab.com/resource-center/insights/content/sector-views accesed 2021-12-14 17:58

[12]: XLK ETF report, ETF, https://www.etf.com/XLK#overview, accessed 2021-11-08 21:16

[13]: wallstreetbets subreddit statistics, https://subredditstats.com/r/wallstreetbets, accessed 2021-11- 19:12

[14]:Trending on Twitter: Social Sentiment Analytics, Bloomberg, https://www.bloomberg.com/company/press/trending-on-twitter-social-sentiment-analytics/, accessed 2021-11-10 20:03

[15]: About, Stocktwits, https://about.stocktwits.com/, accessed 2021-11-10 20:29

[16]: Introducing Official accounts, Stocktwits, Introducing Official Accounts | by Stocktwits, Inc. | The Stocktwits Blog, accessed 2021-11-10 20:28

[17]: Trending on Twitter: Social Sentiment Analytics, Bloomberg, https://www.bloomberg.com/company/press/trending-on-twitter-social-sentiment-analytics/, accessed 2021-11-10 20:03

[18]: Support vector machine , towardsdatascience, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47, accessed 2021-12-05

[19]: Trading day, Wikipedia, https://en.wikipedia.org/wiki/Trading_day, accessed 2021-12-18 21:25

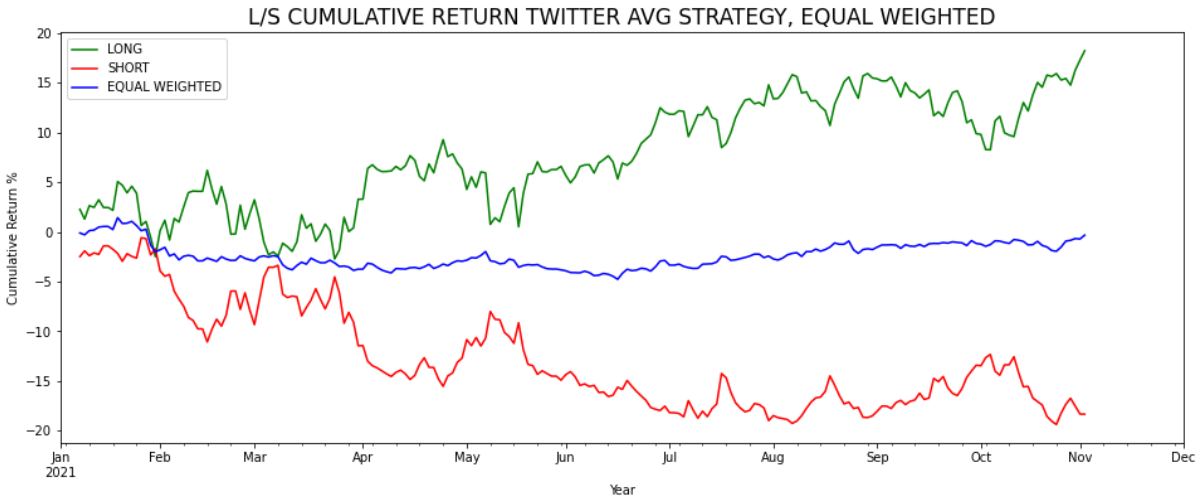
[20]: Prislista, Nordnet, https://www.nordnet.se/se/kundservice/prislista?gclid=Cj0KCQiAqvaNBhDLARIsAH1Pq51yELBxIaEG-BxLFl7l4mW2rLT2ylq20g-YU_M46RiVWBwMn61qxr4aAmBfEALw_wcB, accessed 2021-12-18 21:30

fusionptcapital	hmeisler
SJosephBurns	allstarcharts
charliebilello	GerberKawasaki
timothysykes	DanZanger
howardlindzon	

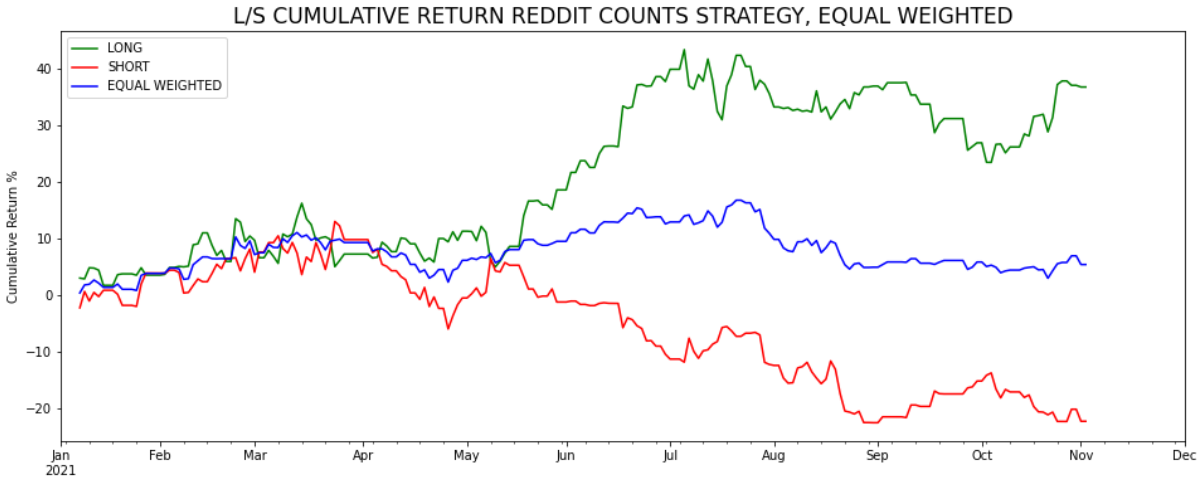
Appendix 1: Selected Twitter account

V	TEL
MA	BR
NOW	IT
MU	WU
FIS	

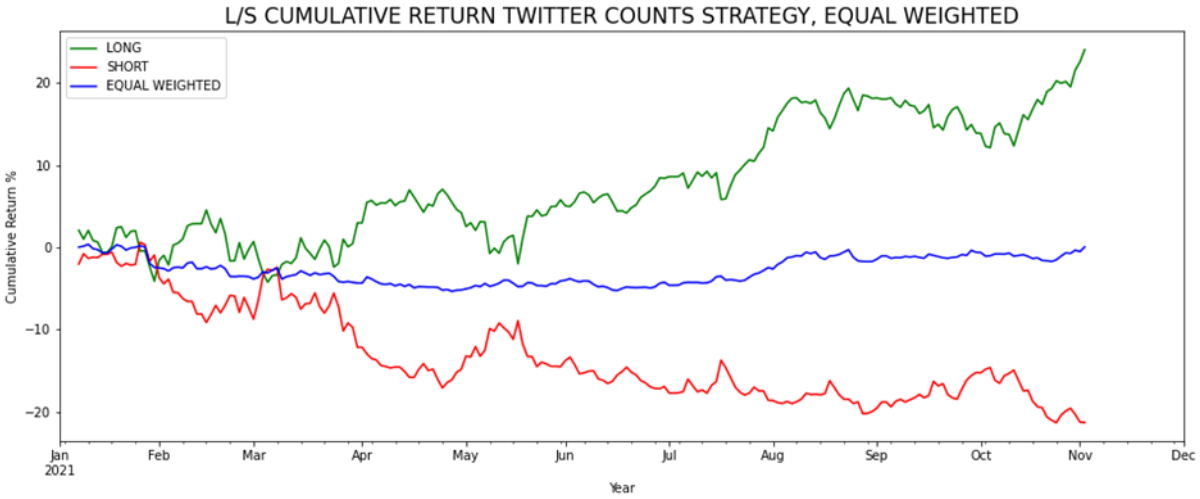
Appendix 2: Equities that were filtered because of noise



Appendix 3: Cumulative returns for long and short positions, Twitter AVG equal weighted (2021)



Appendix 4: Cumulative returns for long and short positions, Reddit counts equal weighted (2021)



Appendix 5: Cumulative returns for long and short positions, Twitter counts equal weighted (2021)

Equal weighted Twitter COUNTS	Mean holdings per day
Long Basket	15.2
Short Basket	15.2

Equal weighted Twitter COUNTS	% of days in long basket
ADBE	29.9
MSFT	29.9
CTSH	29.9
ENPH	29.0
ACN	28.0
CSCO	27.1
QCOM	27.1
AAPL	27.1
HPQ	27.1
ADI	26.6

Equal weighted Twitter COUNTS	% of days in short basket
MU	32.7
AAPL	32.2
V	30.4
ACN	29.9
MSFT	29.4
ENPH	29.4
MSI	29.4
AMAT	29.0
LRCX	28.0
NVDA	27.6

Equal weighted Twitter AVG	Mean holdings per day
Long Basket	15.0
Short Basket	20.9

Equal weighted Twitter AVG	% of days in long basket
TER	46.7
LRCX	43.9
AMAT	43.0
TXN	38.8
ANET	37.9
PAYX	37.9
ADI	36.9
PAYC	35.5
KLAC	35.5
STX	35.0

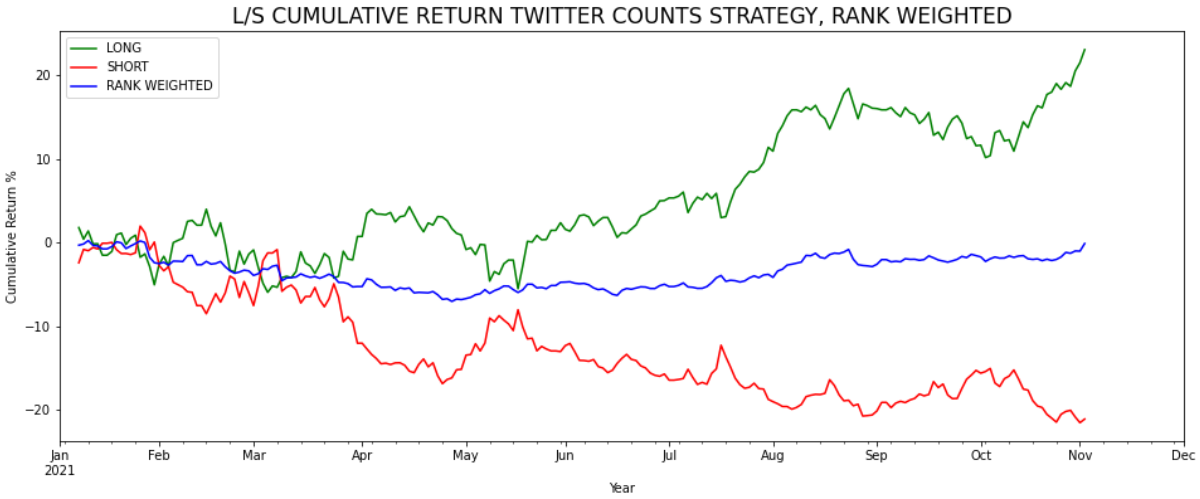
Equal weighted Twitter AVG	% of days in short basket
VRSN	56.1
IPGP	49.1
PTC	46.7
CDW	45.8
WU	41.6
FFIV	41.6
NLOK	40.2
CTXS	39.7
ANSS	39.7
MPWR	37.4

Reddit Counts Equal weighted	Mean holdings per day
Long Basket	1.40
Short Basket	1.04

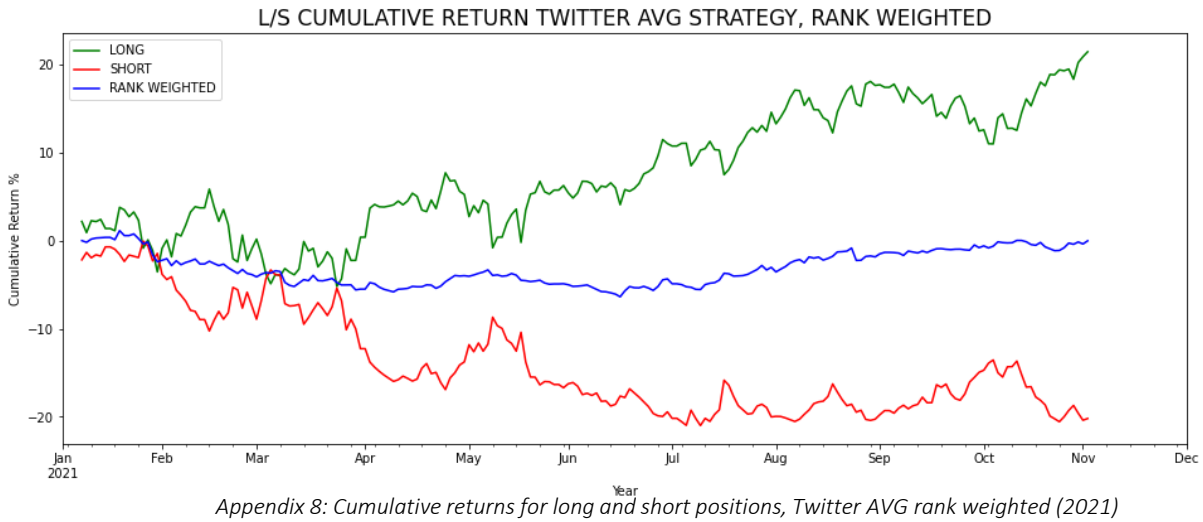
Reddit Counts Equal weighted	% of days in long basket
AAPL	42.7
AMD	38.7
NVDA	27.1
MSFT	14.1
INTC	4.5
PYPL	4.0
MU	3.0
ENPH	2.0
CRM	2.0
IBM	0.5

Reddit Counts Equal weighted	% of days in short basket
AMD	33.7
AAPL	29.6
NVDA	18.6
MSFT	12.6
PYPL	2.5
CRM	2.0
MU	2.0
ENPH	1.0
INTC	1.0
IBM	0.5

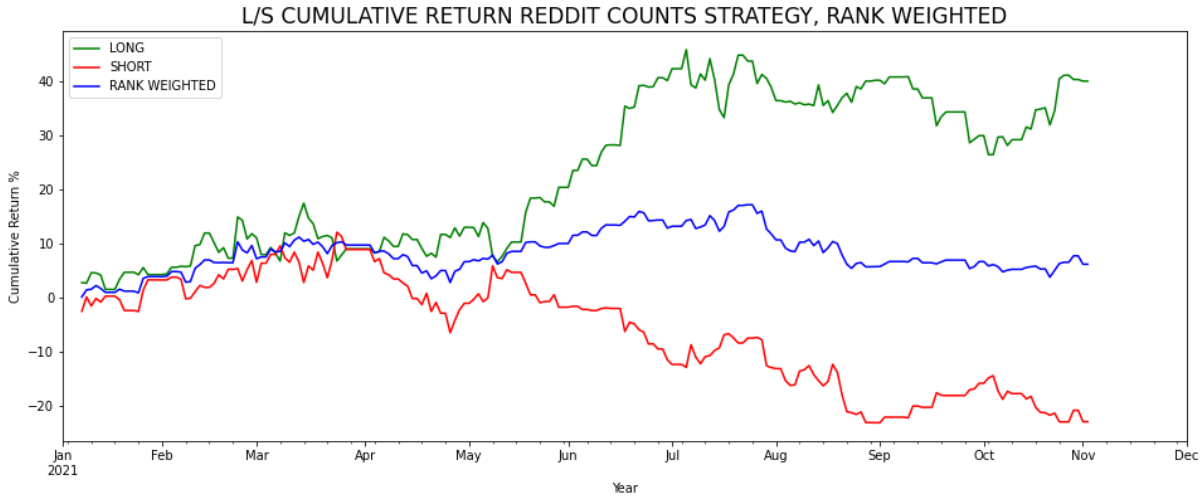
Appendix 6: Metrics for equal weighted allocation strategy of 2021/01/13 – 2021/11/02



Appendix 7: Cumulative returns for long and short positions, Twitter counts rank weighted (2021)



Appendix 8: Cumulative returns for long and short positions, Twitter AVG rank weighted (2021)



Appendix 9: Cumulative returns for long and short positions, Reddit counts rank weighted (2021)

Rank weighted Twitter COUNTS	Mean holdings per day
Long Basket	15.2
Short Basket	15.2

Rank weighted Twitter AVG	Mean holdings per day
Long Basket	15.0
Short Basket	20.9

Reddit Rank weighted	Mean holdings per day
Long Basket	1.40
Short Basket	1.04

Rank weighted Twitter COUNTS	% of days in long basket
ADBE	29.9
MSFT	29.9
CTSH	29.9
ENPH	29.0
ACN	28.0
CSCO	27.1
QCOM	27.1
AAPL	27.1
HPQ	27.1
ADI	26.6

Rank weighted Twitter AVG	% of days in long basket
TER	46.7
LRCX	43.9
AMAT	43.0
TXN	38.8
ANET	37.9
PAYX	37.9
ADI	36.9
PAYC	35.5
KLAC	35.5
STX	35.0

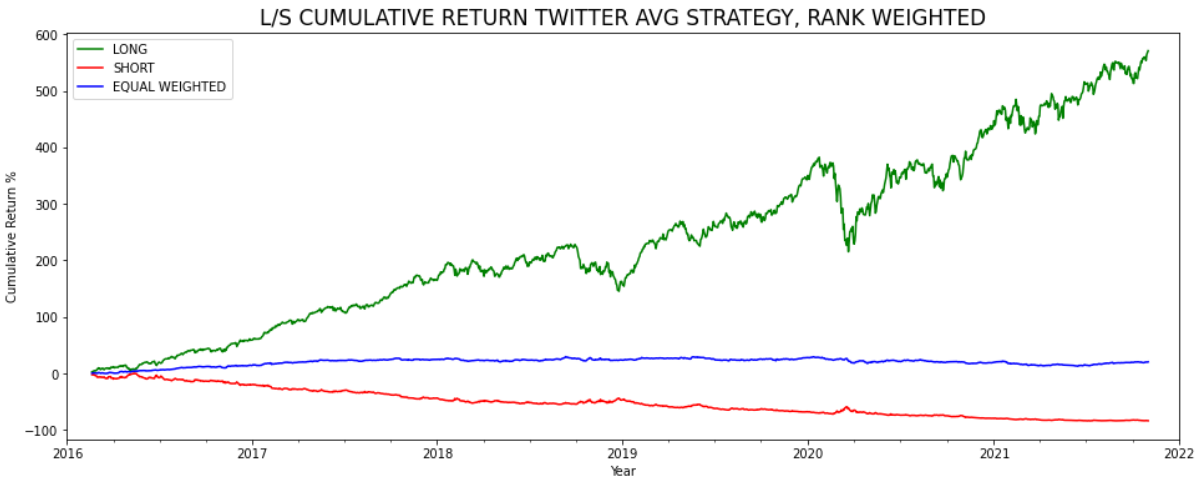
Reddit Rank weighted	% of days in long basket
AAPL	42.7
AMD	38.7
NVDA	27.1
MSFT	14.1
INTC	4.5
PYPL	4.0
MU	3.0
ENPH	2.0
CRM	2.0
IBM	0.5

Rank weighted Twitter COUNTS	% of days in short basket
MU	32.7
AAPL	32.2
V	30.4
ACN	29.9
MSFT	29.4
ENPH	29.4
MSI	29.4
AMAT	29.0
LRCX	28.0
NVDA	27.6

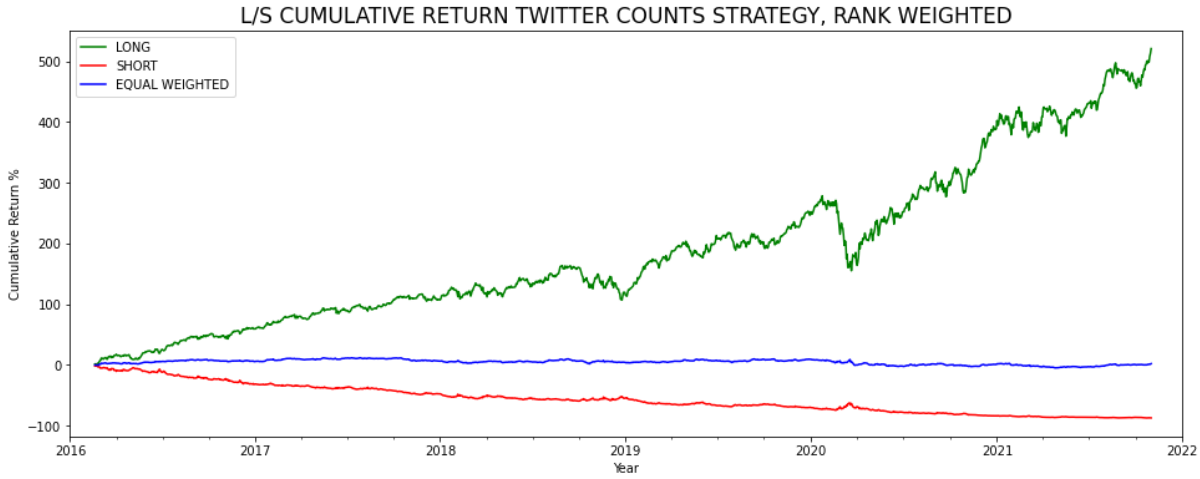
Rank weighted Twitter AVG	% of days in short basket
VRSN	56.1
IPGP	49.1
PTC	46.7
CDW	45.8
WU	41.6
FFIV	41.6
NLOK	40.2
CTXS	39.7
ANSS	39.7
MPWR	37.4

Reddit Rank weighted	% of days in short basket
AMD	33.7
AAPL	29.6
NVDA	18.6
MSFT	12.6
PYPL	2.5
CRM	2.0
MU	2.0
ENPH	1.0
INTC	1.0
IBM	0.5

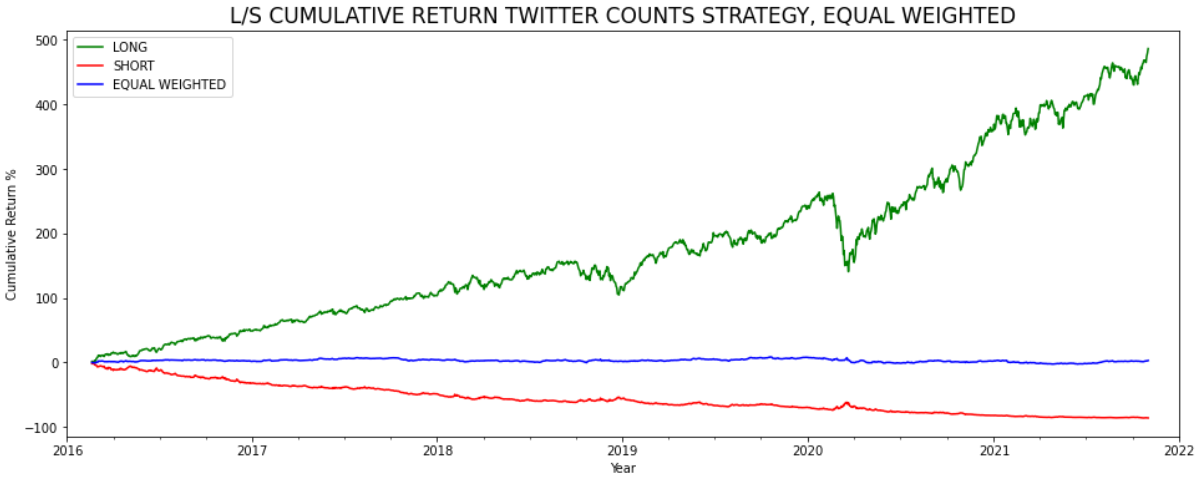
Appendix 10: Metrics for rank weighted allocation strategy of 2021/01/13 – 2021/11/02



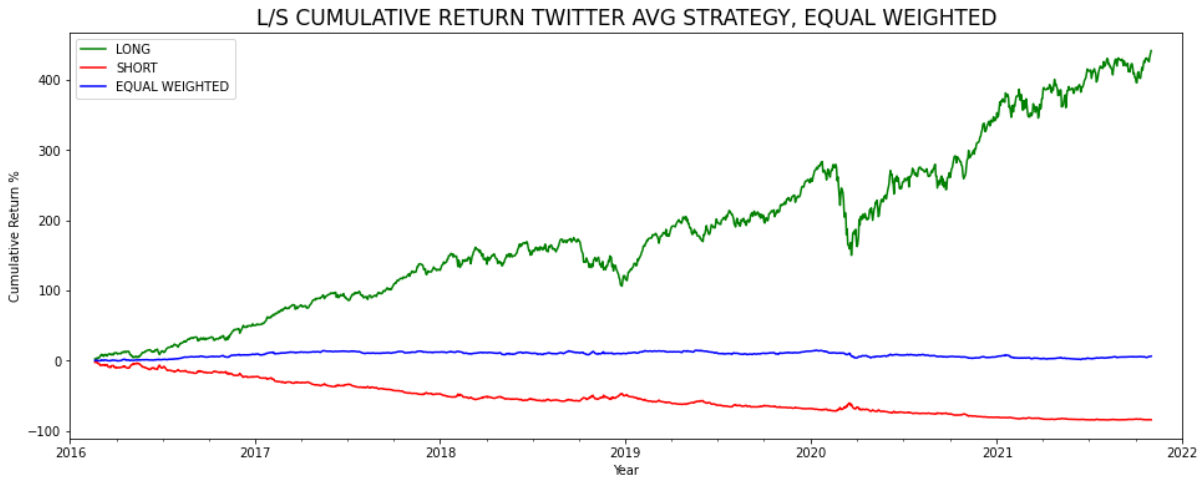
Appendix 11: Cumulative returns for long and short positions, Twitter AVG rank weighted (2016)



Appendix 12: Cumulative returns for long and short positions, Twitter counts rank weighted (2016)



Appendix 13: Cumulative returns for long and short positions, Twitter counts equal weighted (2016)



Appendix 14: Cumulative returns for long short positions Twitter AVG equal weighted (2016)

Rank Weighted Twitter COUNTS	Mean holdings per day
Long Basket	15.2
Short Basket	15.2

Rank Weighted Twitter AVG	Mean holdings per day
Long Basket	15.0
Short Basket	17.6

Rank Weighted Twitter COUNTS	% of days in long basket
AMD	29.1
MU	28.2
LRCX	27.2
WDC	26.5
AVGO	26.5
ORCL	26.3
WU	26.1
V	26.1
CSCO	26.1
AMAT	26.0

Rank Weighted Twitter AVG	% of days in long basket
LRCX	49.4
PAYC	42.2
AMAT	41.3
SWKS	41.2
TXN	37.1
KLAC	35.0
AVGO	33.6
ANET	33.3
TER	33.1
FLT	32.0

Rank Weighted Twitter COUNTS	% of days in short basket
MU	30.8
AMD	29.1
LRCX	27.9
AMAT	27.4
V	27.1
NVDA	27.0
AVGO	26.3
ORCL	26.3
ACN	25.7
PYPL	25.7

Rank Weighted Twitter AVG	% of days in short basket
IPGP	36.5
STX	34.8
VRSN	32.9
CDW	32.5
JKHY	31.9
CTXS	31.8
ZBRA	31.4
MU	31.2
FLT	30.4
ADP	30.2

Appendix 15: Metrics for equal and rank weighted allocation strategy of 2016/02/19 – 2021/11/02

Comment	Ticker	Label	Negative	Neutral	Positive	Compound
Just saying but AMD is popping off	AMD	0	0.00	1.00	0.00	0.00
Go AMD go!	AMD	0	0.00	1.00	0.00	0.00
The amount of times I've seen AAPL declared dead over the years is hilarious. Just buy it and hold it.	AAPL	-1	0.17	0.72	0.11	-0.38
Somebody give me some hope for NVDA calls down 85%	NVDA	1	0.00	0.76	0.24	0.44
thanks AMD facepalm	AMD	1	0.00	0.41	0.59	0.44
Lol AAPL cucking	AAPL	1	0.00	0.42	0.58	0.42

Appendix 16: Example of VADER scoring of Reddit comments

Disclaimer

Disclaimer

These analyses, documents and any other information originating from LINC Research & Analysis (Henceforth "LINC R&A") are created for information purposes only, for general dissipation and are not intended to be advisory. The information in the analysis is based on sources, data and persons which LINC R&A believes to be reliable. LINC R&A can never guarantee the accuracy of the information. The forward-looking information found in this analysis are based on assumptions about the future, and are therefore uncertain by nature and using information found in the analysis should therefore be done with care. Furthermore, LINC R&A can never guarantee that the projections and forward-looking statements will be fulfilled to any extent. This means that any investment decisions based on information from LINC R&A, any employee or person related to LINC R&A are to be regarded to be made independently by the investor. These analyses, documents and any other information derived from LINC R&A is intended to be one of several tools involved in investment decisions regarding all forms of investments regardless of the type of investment involved. Investors are urged to supplement with additional relevant data and information, as well as consulting a financial adviser prior to any investment decision. LINC R&A disclaims all liability for any loss or damage of any kind that may be based on the use of analyzes, documents and any other information derived from LINC R&A.

Conflicts of interest and impartiality

To ensure LINC R&A's independence, LINC R&A has established compliance rules for analysts. In addition, all analysts have signed an agreement in which they are required to report any and all conflicts of interest. These terms have been designed to ensure that COMMISSION DELEGATED REGULATION (EU) 2016/958 of 9 March 2016, supplementing Regulation (EU) No 596/2014 of the European Parliament and of the Council with regard to regulatory technical standards for the technical arrangements for objective presentation of investment recommendations or other information recommending or suggesting an investment strategy and for disclosure of particular interests or indications of conflicts of interest.

Other