Popularity-based Trading Strategy from Reddit Posts

NGAN, Ka Chai

Supervised by: Dr David Rossiter

COMP4971C – Independent Work
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Summer 2022

Abstract

This research investigated navigating the stock market with social media. Since 2020, the stock market has deviated from normal momentums. The infamous GameStop short squeeze is the inspiration for this research, deriving an alternate strategy to traditional technical analysis.

This research comprises 2 main components:

- Named Entity Recognition
- Deriving a trading model based on Reddit data

The first part is using a word2vec model to classify sentences with mentioning of stocks. The model performs well and is able to classify synonyms/ industries for top 10 stocks. It is also performing well with other words where the similarity score is over 0.7.

The final model has 3 separate variables, namely upvotes, number of posts and number of comments. The outcome is measured by the return in an initial investment of $100,000 USD.

Backtest results show that the final model produces output in which the model based on the number of comments rank the best in return% amongst all the models with 71.5% of stocks ranking the highest. However, it still cannot outperform the market in the long term.

However, results show that it was able to recognise the spikes in stock price, but was not able to sell it for profit at the right timing. At the same time, the ‘comments’ model was also better performing in negative return months. Taking GME as an example, it had a better result than the stock itself in 7 out of 9 of the negative return months.

This project showed that there are promising results for navigating the stock market with a reddit-driven trading strategy.

Table of Contents

Introduction

1.1 Objective

Ever since the GME Short Squeeze that happened in Jan 2021[1], it appears that the market for the GME stock has been volatile and more unpredictable with even GME's leadership saying it publicly as well [2]. The public has also proven that they are a huge enough force that the market price changes according to the direction of the public, proving that the strength of many individual investors combined can alter the market

1.2 Hypothesis

This is the hypothesis for the study:

The direction of stock price changes can be observed from the public stock community activity and can be predicted by overall sentiment and activity of individual investors.

1.3 Key Terms

1.3.1 Stock Market

GME Short Squeeze

The GME short squeeze was a coordinated effort triggered by retail investors that purchased the heavily shorted stock GameStop Corporation. Multiple major hedge funds and short sellers have suffered major losses as stock price skyrocketed to 30 times the valuation at the beginning of the month.

Public Equity Float

It is the amount of shares available for public investors to purchase over those internal shares that companies and their subsequent holds to control price & interests

---

[1] Stewart, E. (2021, January 25). *The GameStop Stock Frenzy, explained*. Vox, from https://www.vox.com/the-goods/22249458/gamestop-stock-wallstreetbets-reddit-citron

[2] *GameStop knows its stock is 'extremely volatile' — but leadership says it's completely out of their control*. Business Insider, 2021, https://www.businessinsider.com/gamestop-stock-extremely-volatile-executives-say-cant-be-controlled-2021-3.5.

1.3.2 Machine Learning Techniques

Named Entity Recognition (NER)

NER is the technique to recognize entities in text, such as recognising 'GME' is an organisation within a sentence. It is extremely useful to extract key features in text, in this case, extracting tickers and their synonyms.

Word2Vec

This Natural Language Processing (NLP) technique is an architecture that uses different neural networks and algorithms to predict word association by vectorizing them into a plane. It is used to understand the semantic meaning of words and predict the similarity of words.

BERT

Bi-directional Encoder Representation from Transformers[3], commonly known as BERT, is a pre-trained model on languages using masked language modelling objectives, which masks 15% percent of tokens and BERT is used to predict the context of those tokens. This is one of the best pre-trained models existing currently and a bit of fine-tuning can result in better results in the domain-specific context.

1.4 Initial Difficulties

Data collection is difficult because firstly it is not from official news sources, it was very time-consuming to collect reddit data for two years, multiprocessing was not able to be used as well as it will block the local IP if too many requests are sent in.

Training Named Entity Recognition was a difficult part as there was not much training data existing with labels, while manually labelling data takes a long time and is inaccurate.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Methodology

2.1 Data Collection

Financial News providers such as Wall Street Journal[4] and Newsfilter[5] are popular for analysing the stock earnings reports and stock-specific news. However, in the recent two years, a lot of inexperienced people have diverted their investments from index funds to small market-cap stocks.

People aim for high returns and from the GME short squeeze, we can observe that high returns can come from stocks that are high on public float. In this case, data from discussion website Reddit are used. The stock communities are well-established, consisting of many users. The way of representing sentiment on stocks is different from mainstream news media as well.

2.1.1 Reddit Scraper

Using the Pushshift API to dive into data from Reddit, choosing the subreddit #wallstreetbets only as it narrows our focus to the most talked about stocks in the most popular stock discussion channel.

Data from 01/05/2020 to 31/07/2022 is collected. There was significant data input from this subchannel so the start date was chosen. While the end date was set for completeness

2.2 Data preprocessing

Choosing data from a post, only 6 things are extracted :

- Title
- Context
- Upvote_ratio
- Comments
- Date
- Score

[4] The Wall Street Journal. (n.d.). *Breaking news, business, Financial & Economic News, World News and Video*. The Wall Street Journal, from https://www.wsj.com/

[5] *Newsfilter.io*. Business & Financial News. (n.d.), from https://newsfilter.io/

Fig 2.1.1 *Sample Reddit Post*

For clarity, in this sample post, the '185' is the upvote ratio while the bolded text is the title. Selftext is all the content below. The score predicted is calculated with the upvotes minus the downvote. Other stuff is less significant to the analysis so are processed out and for convenience, the title and selftext are joined together.

To find out what stocks are mentioned in the text, the package TickerExtractor is used to find the stocks mentioned. However, it only detects ticker symbols which misses out on many things therefore when we split the data for analysis, stocks' full name and aliases are added manually to get wider coverage.

Publish times of posts are also altered to differentiate posts before and after trading days, as all the time extracted are in UTC format, they are shifted forward by 10 hours to accurately capture the values before the trading day begins.

2.3 Threshold Setting

Using the gensim word2vec model[6], the threshold for classifying a text is set at 0.7. From checking a few examples, we can see that the result of the model is related to things for most above 0.7. Examples can be found in *3.3*

While for trading stocks, the threshold is also set at 0.7. If the value of the dependent variable is higher than the 70% quantile value from within the data, we will purchase the stock at the opening, whereas if under, the shares will be sold. This is tested with variable values, 65%, 75%, 80%, 85% and 90%, threshold=0.7 is the best performing in backtests, therefore it was chosen as the threshold for stock tradings

[6] Nederhand, R., PS, S., Radim, Payne, B., Kestemont, M., Pinches, M., DG2, Break, Mark, Janson, m, irt24, Simon, McNamara, T., Adam, Aayush, Lachlan, & Parisa. (2013, September 17). *Deep learning with word2vec and gensim*. Pragmatic Machine Learning, from https://rare-technologies.com/deep-learning-with-word2vec-and-gensim/

2.4 Backtesting

To analyse the result of the trading strategy, stock trading is based on the social media data from the previous trading day.

The starting value is $100,000 USD. If the factor has a value higher than the threshold in the previous trading day, shares are bought in and the number of shares bought is the maximum amount the starting value can purchase at the open price of the trading day. When the factor goes down on the trading day, the stocks are sold at the open on the next trading day.

2.5 Evaluation Criteria

The criteria to analyse for the algorithm is the actual return compared to the stocks' return, whether we out performed it or under performed it.

Data Exploration

In this section, 3 stocks based on top 10 is displayed here, others can be found in the Appendix for additional information

3.1 Top 10

To obtain results with a significant pool of data, the top 10 mentioned tickers are extracted to filter out posts that do not mention tickers or mention less popular tickers.

Using the package reticker's TickerExtractor, the following 10 tickers are found with the highest mentions:

Ticker Symbol	Company Name	No. of Mentions
GME	GameStop Corp	92685
AMC	AMC Entertainment Holdings	56437
BB	BlackBerry	17576
WISH	ContextLogic Inc	/
NOK	Nokia	5883
TSLA	Tesla	17415
CLOV	Clover Health	12167
PLTR	Palantir Technologies	4982
SNDL	Sundial Growers	5542
SPY	SPDR S&P 500 ETF Trust	/

Fig 3.1.1 *Top 10 stock symbols with highest Reddit Mentions*

In the following analysis, WISH is not included as most of its mentions are using the literal word 'wish' instead of referencing the stock. SPY is also removed as it has the least mentions and is an ETF of the S&P 500 index. C

3.2 Stock Prices

To see an overview of the stock price to try and find an initial pattern in the change in stock prices for these US stocks.

We can observe a pattern amongst these 3 stocks, There were three main bursts in the volume traded for them, with twice between Jan and Apr 2021, and once during Jun 2021.

The Jan 2021 rise was the infamous GME short squeeze incident, while from the late March rise we could see that the volatility of it has made the price changes volatile. It recorded a 53% increase in price a day the stock price dropped from an under performing quarter recorded from earnings release.

While, we can also see that the price of top mentioned stocks like AMC Entertainment and BlackBerry also shows similar trends as mentioned above. The maximum return from day trading these stocks can be huge, but the losses can also be huge if we navigate the market situation wrong

GME

Fig 3.2.1 *OHLCV of GME from Jan 2020 to Jul 2022*

AMC

Fig 3.2.2 *OHLCV of AMCfrom Jan 2020 to Jul 2022*

BB

Fig 3.2.3 *OHLCV of BBfrom Jan 2020 to Jul 2022*

3.3 Word2Vec

Using all posts from r/wallstreetbets to create the corpus, the gensim word2vec model is used to train on the corpus. On the 100-dimension vector plane, using the top 10 tickers, the mean vector of the top tickers is calculated for using to extract companies from text.

Looping through tokens in each post, if the word has a similarity score of higher than 0.7 relative to the mean vector above, it is then classified and labelled as text positive to the company.

3.3.1 Company Context

Inputting companies in their full name, we can see that the top 10 words that return are all related to the stock, mainly with company names. This

```
Similar words to: facebook
('fb', 0.8533295392990112),
('snapchat', 0.7911134362220764),
('netflix', 0.760712742805481),
('tiktok', 0.7574191093444824),
('amazon', 0.721465528011322),
('apple', 0.721445620059967),
('meta', 0.716083824634552),
('myspace', 0.7153778672218323),
('whatsapp', 0.7152852416038513),
('instagram', 0.7133999466896057)
```

Fig 3.3.1 *Similarity to 'facebook'*

```
Similar words to: aapl,fb
('msft', 0.9059796333312988),
('nflx', 0.8823227882385254),
('amzn', 0.8674663305282593),
('nvda', 0.8508817553520203),
('googl', 0.8300726413726807),
('tsla', 0.8224852085113525),
('appl', 0.8198520541191101),
('goog', 0.815477728843689),
('amd', 0.8106151819229126),
('snap', 0.8057532906532288)]
```

Fig 3.3.2 *Similarity to 'aapl, fb'*

3.3.2 Financial Context

There are a few additional interesting results found from the vector plane. For example, when we search for 'sold', the model returns results for uncommon words but have the same meaning in a financial context. 'Unloaded' and 'liquidated' are some words not usually used but are often used in replacement of 'sold'.

When we query for 'strike', it is able to recognise the meaning of it in financial context. 'Strikes', 'call', 'expire' are all words commonly used when we mention 'strike'.

```
Similar words to: sold

('bought', 0.7876056432723999),
('purchased', 0.7505807280540466),
('unloaded', 0.728362500667572),
('rebought', 0.721016526222229),
('liquidated', 0.6959587931632996),
('cashed', 0.6950024366378784),
('swapped', 0.674940288066864),
```

Fig 3.3.3 *Similarity to 'sold'*

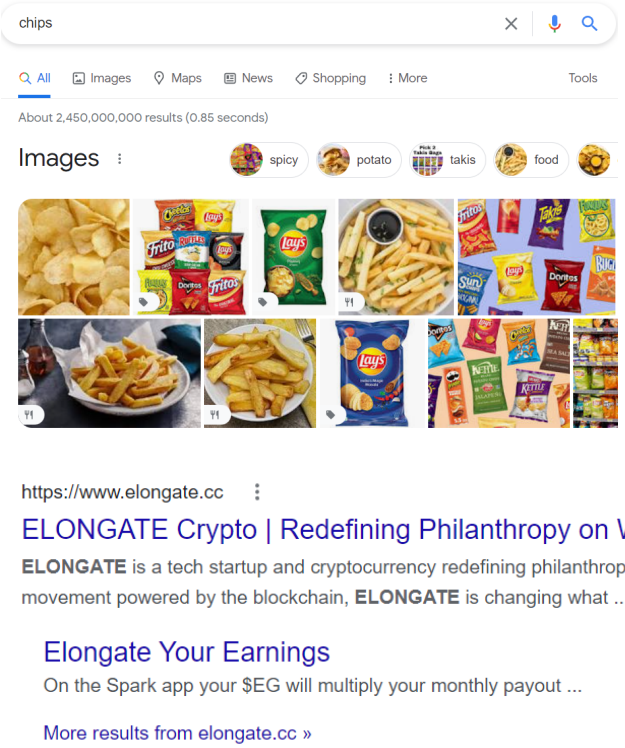
```
Similar words to: strike

('strikes', 0.736444354057312),
('expiry', 0.7151036262512207),
('expiring', 0.6921681761741638),
('expiration', 0.6909973621368408),
('call', 0.662369966506958),
('expirations', 0.6517267227172852),
('option', 0.649758517742157),
```

Fig 3.3.4 *Similarity to 'strike'*

3.3.3 Others

When we search for chips in google search, the results that come up are all potato chips. However, chips are usually not correlated to any of our stock purchases. We would rather know about Central Processing Unit (CPU) chips. When we insert it into the model, it outputs all related words to CPU chips. It can definitely be used to define industries text but in this project it could not provide significant results for identifying stocks.

For 'dogecoin', we can see results of other coins, but as well as 'elongate'. It does not appear as a familiar word if we are not deeply invested in crypto. However, it is positively correlated to 'dogecoin'. After googling, we can see that it is a tech startup powered by blockchain. This model can be used further to find lesser-known relations between various stocks in the market as well.



```
Similar words to: chips

('chip', 0.7570053935050964),
('microchips', 0.7430556416511536),
('cpus', 0.7248637676239014),
('processors', 0.7235096096992493),
('cpu', 0.7191013097763062),
('5nm', 0.7186688184738159),
('gpus', 0.714860200881958),
('macs', 0.7079733610153198),
('tsmc', 0.6785429120063782),
('cpu's', 0.6767925024032593)]
```

Fig 3.3.5 *Similarity to 'chips'*

```
Similar words to: dogecoin

('doge', 0.9379846453666687),
('shiba', 0.8793869018554688),
('shib', 0.8787811994552612),
('xrp', 0.8373001217842102),
('coin', 0.8059687614440918),
('inu', 0.802283525466919),
('litecoin', 0.7937556505203247),
('btt', 0.7922592759132385),
('shibu', 0.7810928821563721),
('elongate', 0.7675046324729919)
```

Fig 3.3.6 *Similarity to 'dogecoin'*

3.4 Sentiment

For the sentiment model, FinBERT is used to classify the sentiment of a text[7]. Both ProsusAI/finBERT and yiyanghkust/FinBERT pretrained models are tested with the outputs, the output for the model is either positive (1.0) or neutral (0.0) or negative (-1.0). There is a huge output imbalance, generating not enough positive & negative labels. Without enough outputs, it is not included in subsequent parts as it cannot show enough results for analysis.

Limitations of this analysis include the financial context. These two finbert models are trained on official, formal financial news such as FinancialPhraseBank. Although it is in a domain-specific context, it does not use the same type of expressions.

For example, the following post is an example of a title from reddit, it is classified as negative with the model. However, it should be classified as negative based on the title, but when we look at the text, it is actually a neutral description of the situation.

'Why's everybody so doom and gloom with BBBY today?'[8]
Therefore, a limitation for the model is that in long paragraphs, it could not get an accurate sentiment score as most text is directly labelled as neutral. This is caused by the existence of a larger portion of neutral words in the text.

---

[7] Sidogi, T., Mbuvha, R., & Marwala, T. (2021, October). Stock Price Prediction Using Sentiment Analysis. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 46-51). IEEE.

[8] *R/wallstreetbets - why's Everybody so doom and gloom with BBBY today?* reddit. (n.d.). Retrieved August 14, 2022, from https://www.reddit.com/r/wallstreetbets/comments/wkgv0c/whys_everybody_so_doom_and_gloom_with_bbby_today/

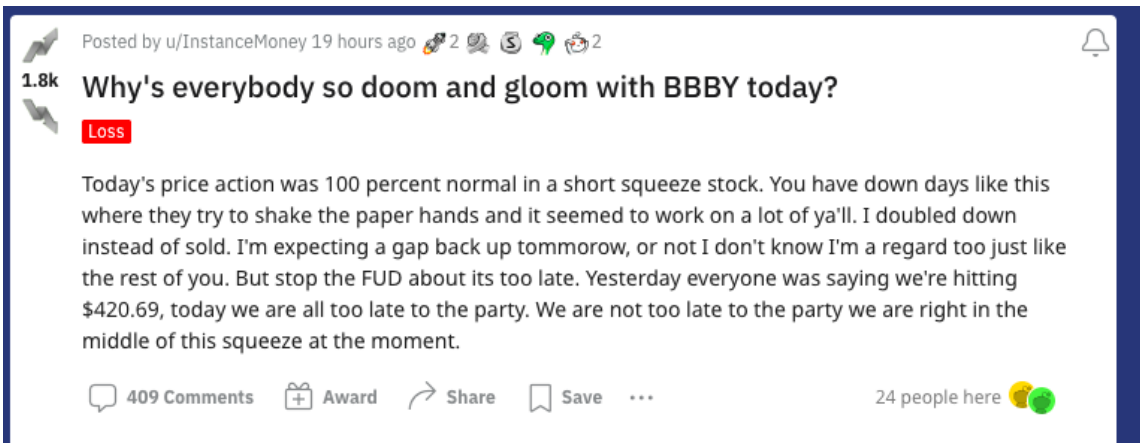


Fig 3.4.1 *Example reddit post with partial varying sentiments*

Strategy & Results

4.1 Overall Performance

Reviewing the results of stocks prices that correlates the most with reddit engagement data, we can observe a similar pattern between the parameters. However, the large volatility coincides with the periods with huge price swings observed above.

In the long run, the stock price has risen for all the #wsb stocks but CLOV, while the model has also achieved a positive return. However, for most of the results, in the long run, if we strong-holded the stock, the return % is be higher than doing day trading.

To interpret this outcome, a possible outcome is that buy and sell orders are only executed on the Opening of that stock trading day, it is an inefficient trading strategy only trading on Open Price, but just to show the effect of the effect of Reddit on the stock price, we can observe that trading based on Reddit engagement values can capture the rises very efficiently, whereas it could not grasp the selling times, leading to a drastic drop after earning a huge profit.

The following figure shows the result of 8 stocks annual return in percentages. The 'Comments' model performed better in April 2020 (start of r/wallstreetbets) to April 2021 whereas from April 2021 to April 2022, there was no specific model that performed especially well.

	2020-04 to 2021-04				2021-04 to 2022-04			
	Stock	Comment	Score	Count	Stock	Comment	Score	Count
AMC	103.86	71.43	42.09	18.89	52.54	-3.62	49.02	-12.69
GME	2929.49	880.69	457.94	451.63	-27.95	-14.95	-43.12	-56.72
BB	106.31	-1.59	21.75	31.47	-35.22	-36.78	-34.46	-33.19
TSLA	353.68	185.69	173.84	148.14	22.74	34.66	7.02	8.49
CLOV	19.14	-36.71	-74.81	-70.46	-83.93	-62.76	-67.03	-65.56
NOK	30.73	37.55	19.12	48.48	7.69	5.37	49.62	50.35
SNDL	51.58	58.18	80.82	76.72	-45.72	-24.54	-84.09	-78.37
PLTR	\	\	\	\	153.05	-45.35	37.39	-19.63

Fig 4.1.1 *Annual Performance of the model on various stocks*

From the results, we can have the hypothesis that the 'comments' model performs the best, with it having the highest annual returns for 71.4% of the stocks from April 2020 to April 2021, while with 62.5% of the stocks from April 2021 to April 2022

The first graph for each figure shows the stock price OHLC chart for stock price, while the 3 below are backtest results trading based on the respective parameters. The green points on the graph represent buys, while red points represent sells. Most of the buy & sells happen within a day, hence we could see the points clustered together.

4.1.1 GME

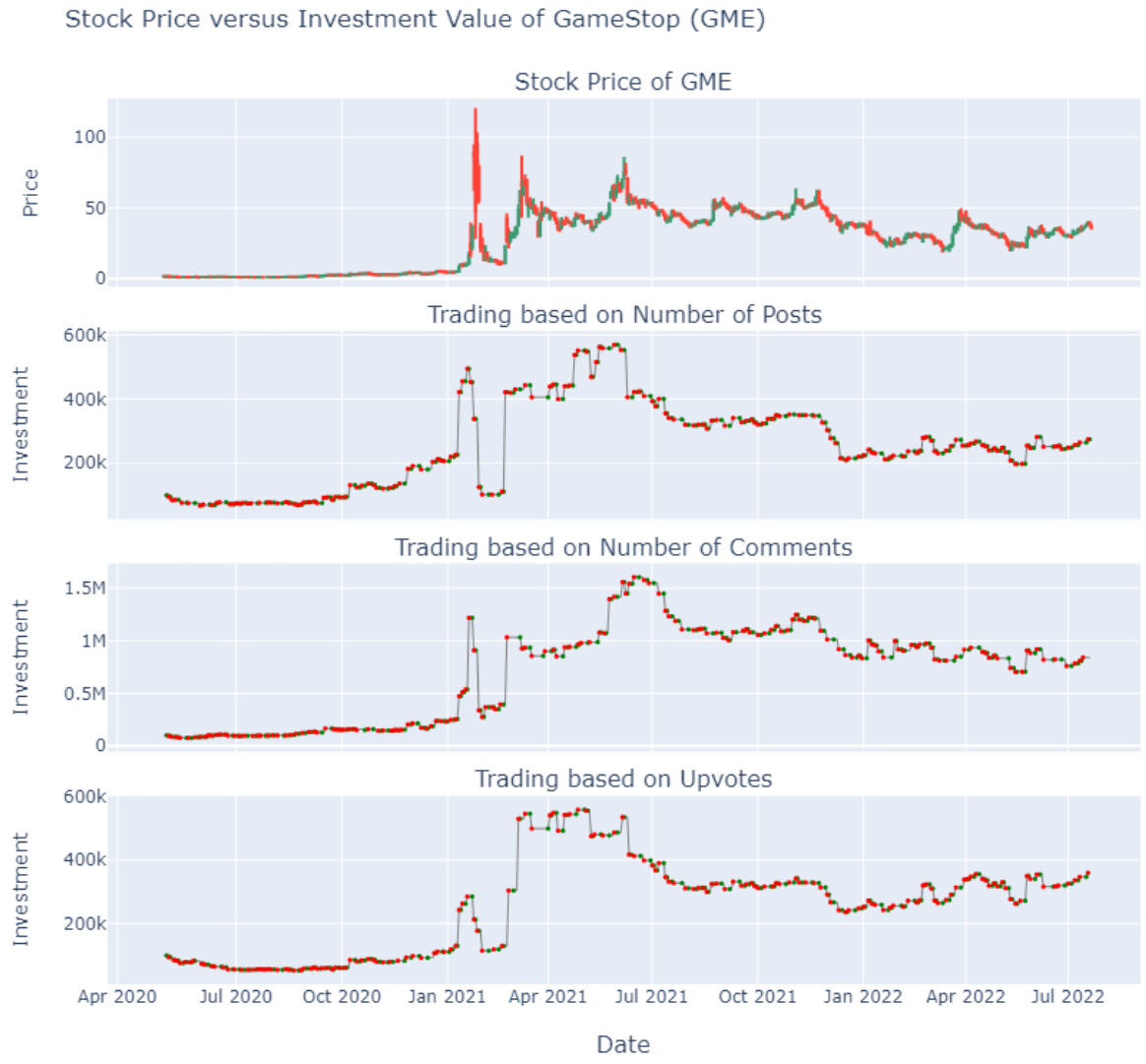


Fig 4.1.2 *Backtest performance of model on GME*

For GME, we can see that trading based on upvotes, can actually earn a near 10X return through the two year period. Although it could not outperform the market, it is still impressive based on the ability to capture profits. And amongst all parameters, Number of comments can capture the highest returns at the threshold of 0.7.

4.1.2 AMC

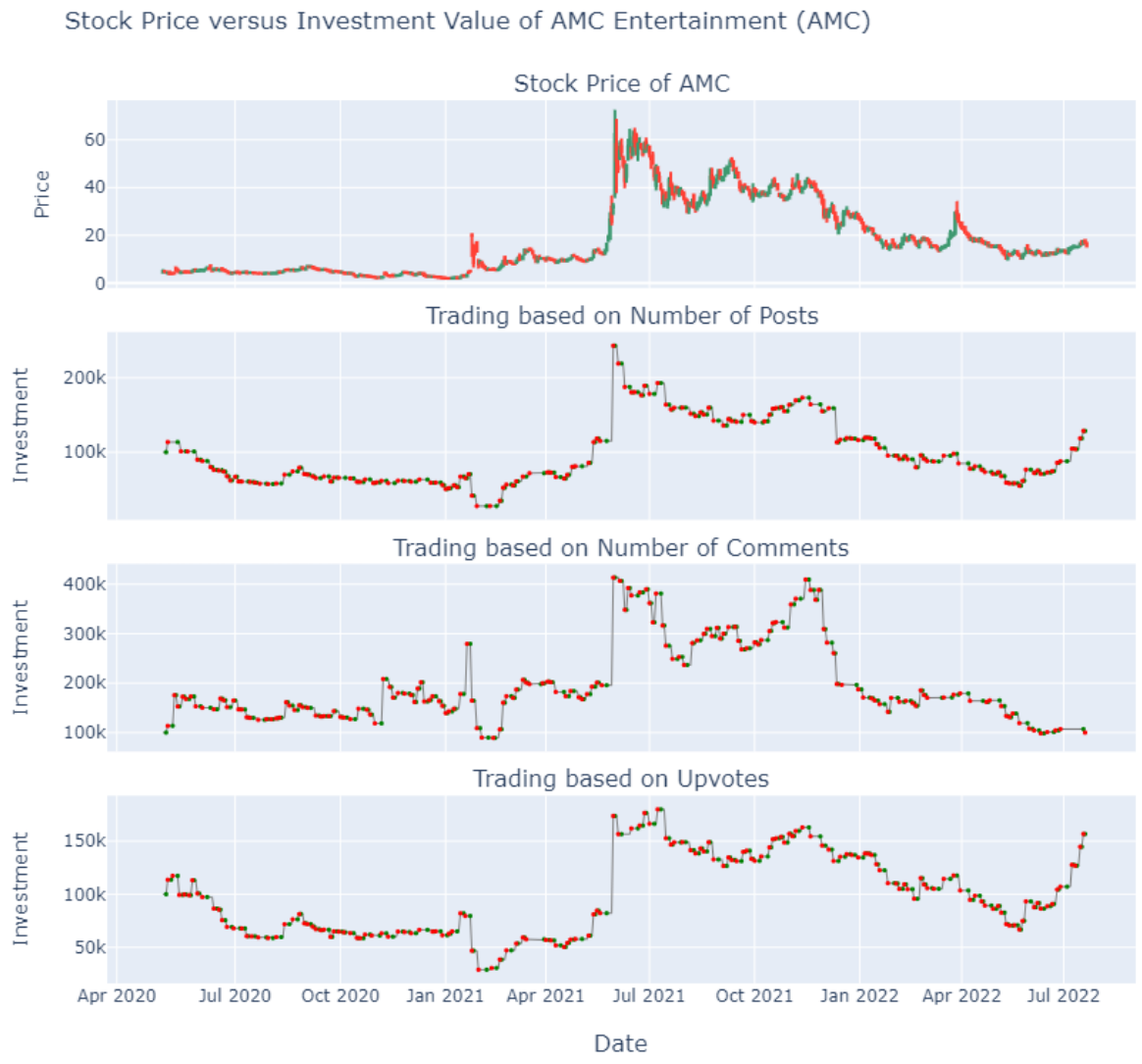


Fig 4.1.3 *Backtest performance of model on AMC*

For AMC, we can see that once again, it captures the rises well but fails to predict and prevent the losses. However, from the graphs, the huge rise was accurately predicted but the resulting rise in investment value is 100% in a day, while the stock price rose to a highest of $72, which is yielding around 400% return. This is probably caused by the disparity of OHLC prices, in which many buy & sell actions happened after the before-the-market or after-the-market trading hours, causing our model to not be able to capture the maximum value.

The model did not perform well for AMC, the result for all 4 parameters could not exceed 100%, compared to the increase in price of AMC, from 5.19 USD to 17.52 USD, is underperforming.

4.1.3 NOK

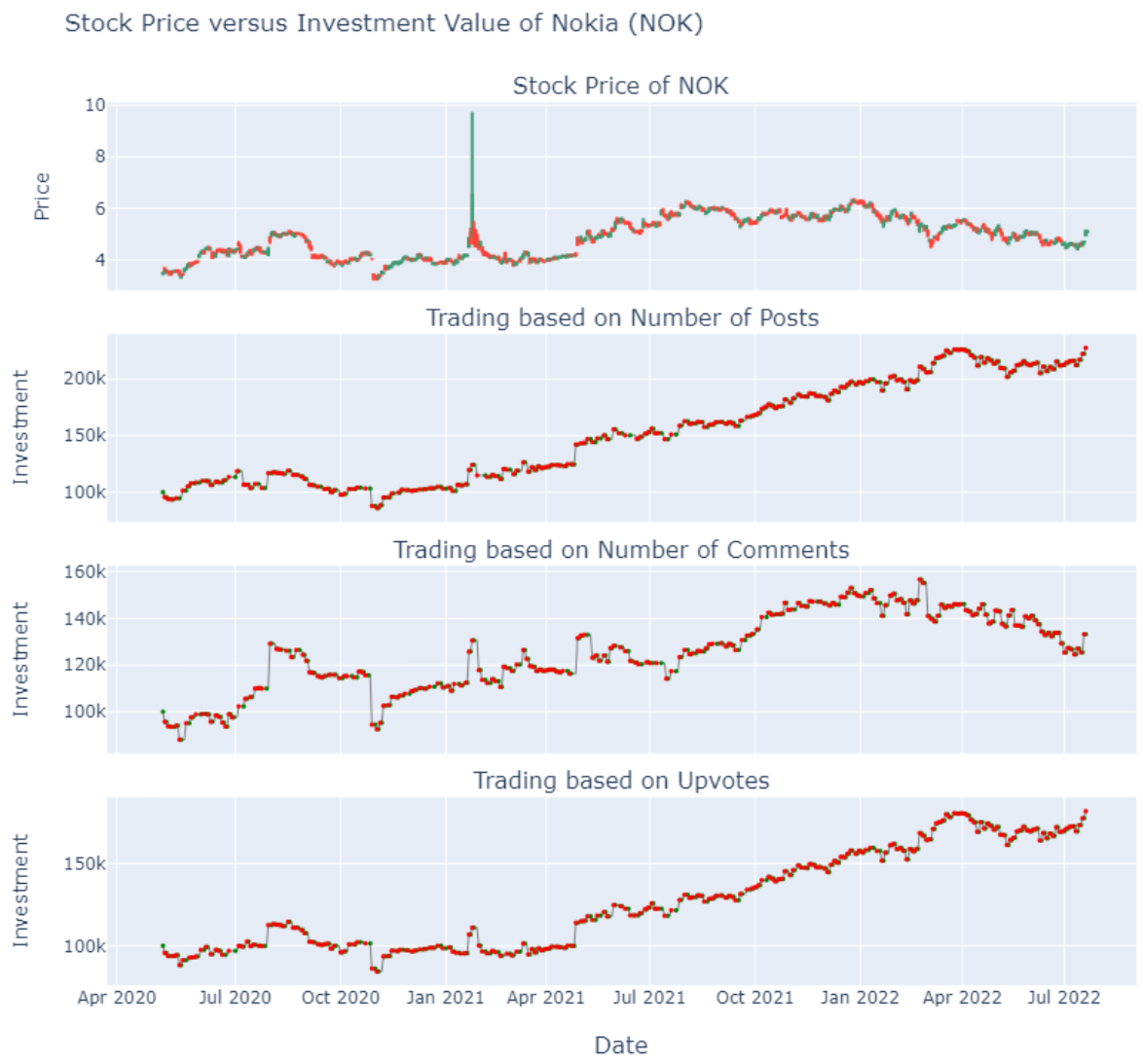


Fig 4.1.4 *Backtest performance of model on NOK*

Out of all the 8 stocks, the best performing stock was Nokia. During the period, it yielded a higher return than the stock if it were traded based on the number of posts per day. The end investment value for the stock is 227k, which was a 127% return for 27 months, whereas the price change of the stock was 3.4 USD to 5.1 USD, which is a 50% return.

However, we can see that there was a huge spike in Nokia's stock price that the parameters did not fully capture, it was due to a similar open prices for that day and the day after, while the stock price once went up to 9.7 USD within the day.

4.2 Comparison of comments, upvote ratio, posts

Stock	Stock Price (%)	Upvotes (%)	Comments (%)	Posts (%)
GME	2440.09	259.60	741.77	174.55
AMC	227.55	56.66	-0.06	28.56
BB	42.10	-26.57	-46.72	-19.17
NOK	46.47	81.73	33.22	127.63
TSLA	409.12	99.06	18.90	82.95
PLTR	10.42	-53.95	-86.44	-68.31
SNDL	-32.26	-74.25	60.59	-65.79

Fig 4.2.1 *Overall return percentages on stocks for each model*

The total returns of the stocks vary a lot for the 7 stocks, High return stocks (TSLA, GME, AMC) have high returns with the models, while those low return stocks (BB, NOK, PLTR) have lower returns than the stock price with some special cases.

CLOV has been left out of here as the decrease in price was drastic and could not get insights from it.

Amongst the three models, the ‘upvotes’ model performs better than the ‘posts’ model in most stocks, while the ‘upvotes’ and ‘comments’ vary in performance but generally, the ‘upvote’ model performs better, while ‘comments’ can yield higher maximum return.

SNDL

SNDL has a negative stock price return the model based on comments exhibits 60% positive return, the graph below presents the result that the comments model predicted the spikes accurately and did noy drop after the rises.

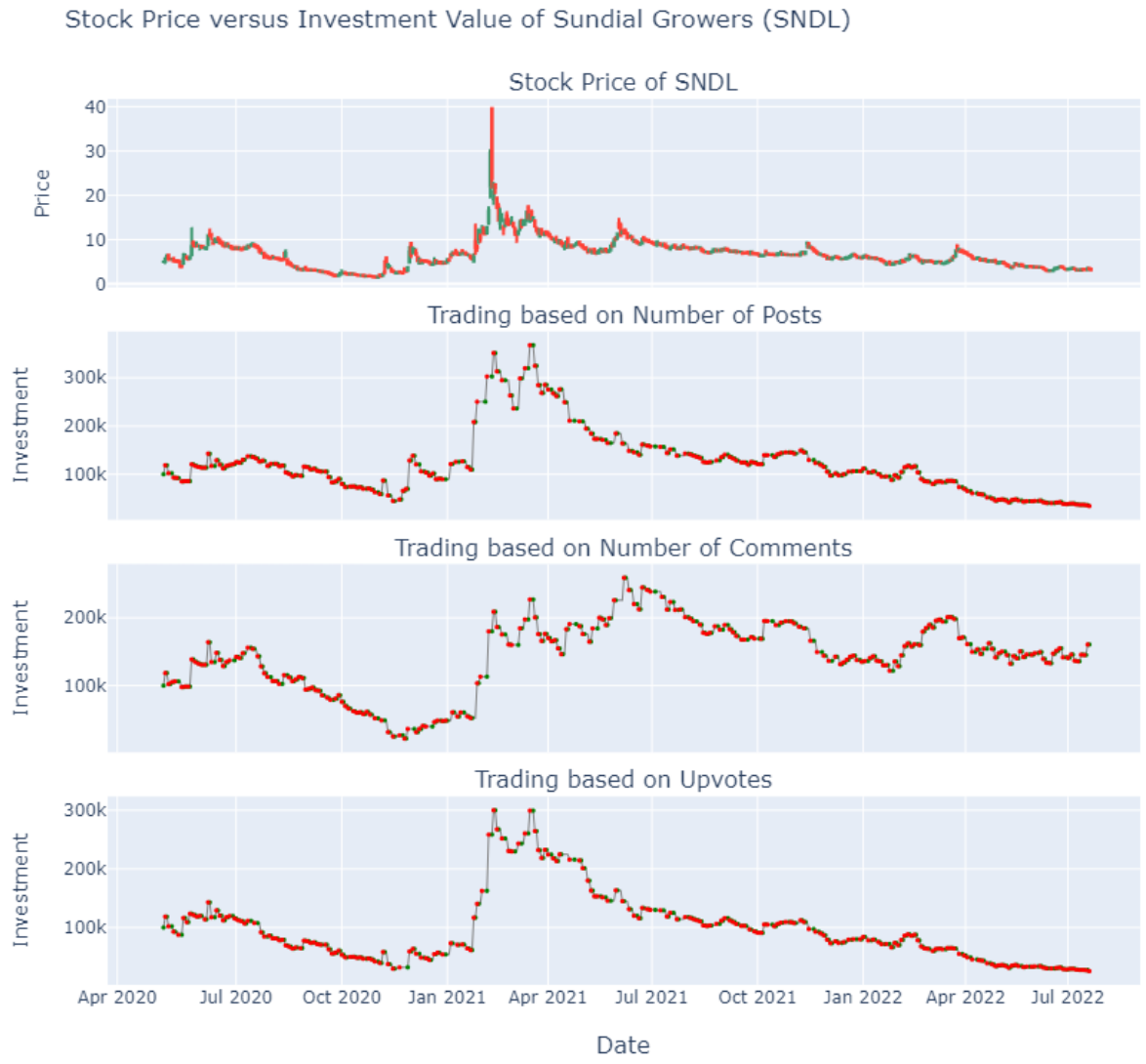


Fig 4.2.2 *Backtest performance of model on SNDL*

4.2.1 Monthly results

The monthly results generally show that the model outperforms the holding of the stock in negative return months but also fails to capture the high return months. For GME, it successfully gets a high return for ‘comments’ model in which the monthly return percentages are the most similar to that of the monthly return of stocks. Oppositely , in months when the monthly return of stock price is relatively low, the return from ‘upvote ratio scores’ is able to go against the market and yield high returns.

GME

GME Monthly Return

	Date	Stock	Comment	Score	Count
1	2020-05-31	-29.14%	-18.30%	-17.14%	-24.40%
2	2020-06-30	6.90%	17.46%	-32.39%	-4.54%
3	2020-07-31	-7.60%	2.22%	0.35%	6.69%
4	2020-08-31	66.58%	24.33%	5.51%	0.96%
5	2020-09-30	52.69%	24.76%	2.78%	20.18%
6	2020-10-31	2.65%	4.12%	31.59%	31.59%
7	2020-11-30	58.17%	28.77%	16.86%	48.34%
8	2020-12-31	13.77%	13.68%	18.94%	13.26%
9	2021-01-31	1625.05%	292.52%	59.19%	63.73%
10	2021-02-28	-68.70%	13.56%	71.20%	24.70%
11	2021-03-31	86.57%	-12.80%	64.08%	-3.84%
12	2021-04-30	-8.55%	8.79%	11.90%	36.03%
13	2021-05-31	27.89%	44.71%	-12.86%	3.25%
14	2021-06-30	-3.54%	9.20%	-18.00%	-28.11%
15	2021-07-31	-24.76%	-28.47%	-21.98%	-21.98%
16	2021-08-31	35.45%	-2.87%	4.79%	4.79%
17	2021-09-30	-19.60%	0.43%	-2.42%	-2.42%
18	2021-10-31	4.58%	1.90%	3.61%	7.78%
19	2021-11-30	6.92%	-0.57%	-4.80%	-7.24%
20	2021-12-31	-24.37%	-23.86%	-20.96%	-32.52%
21	2022-01-31	-26.59%	19.60%	3.46%	1.43%
22	2022-02-28	13.23%	-2.24%	26.02%	26.02%
23	2022-03-31	35.06%	-6.06%	4.64%	-9.65%
24	2022-04-30	-24.92%	-8.97%	-6.20%	-6.20%
25	2022-05-31	-0.26%	5.95%	7.32%	3.84%
26	2022-06-30	-1.96%	-13.97%	-6.06%	-1.48%

Fig 4.2.3 *Backtest monthly performance of model on GME*

AMC

AMC Monthly Return

	Date	Stock	Comment	Score	Count
1	2020-05-31	4.27%	53.09%	0.71%	-10.12%
2	2020-06-30	-16.37%	7.74%	-32.53%	-30.99%
3	2020-07-31	-5.83%	-22.82%	-13.46%	-7.94%
4	2020-08-31	45.54%	18.93%	23.54%	23.53%
5	2020-09-30	-19.90%	-5.33%	-10.84%	-7.09%
6	2020-10-31	-49.89%	-17.28%	-5.61%	-10.83%
7	2020-11-30	80.93%	50.89%	5.42%	4.82%
8	2020-12-31	-50.35%	-13.85%	-4.71%	-8.90%
9	2021-01-31	525.47%	6.81%	-23.69%	-25.63%
10	2021-02-28	-39.59%	5.42%	0.69%	36.04%
11	2021-03-31	27.47%	15.10%	22.16%	27.12%
12	2021-04-30	-1.76%	-14.18%	0.44%	13.00%
13	2021-05-31	160.42%	14.17%	41.92%	41.92%
14	2021-06-30	117.00%	99.15%	114.66%	64.50%
15	2021-07-31	-34.69%	-35.09%	-15.56%	-15.56%
16	2021-08-31	27.31%	23.28%	-10.93%	-10.93%
17	2021-09-30	-19.24%	-13.32%	0.35%	-0.43%
18	2021-10-31	-7.07%	15.55%	11.63%	9.45%
19	2021-11-30	-4.04%	24.34%	-1.93%	-0.05%
20	2021-12-31	-19.86%	-51.77%	-7.72%	-25.09%
21	2022-01-31	-40.98%	-9.19%	-17.94%	-17.94%
22	2022-02-28	17.43%	8.97%	4.28%	0.49%
23	2022-03-31	30.65%	-3.29%	-10.09%	-11.37%
24	2022-04-30	-37.91%	-7.83%	-16.62%	-16.62%
25	2022-05-31	-6.27%	-27.93%	8.10%	8.10%
26	2022-06-30	-5.51%	-10.04%	14.80%	14.80%

Fig 4.2.4 *Backtest monthly performance of model on AMC*

The model performs worse compared to other stocks. It has a good number of well performing months that catches up to the return in stock price, but it couldn't capture the investment value of the huge spike in Jan 2021.

TSLA

TSLA Monthly Return

	Date	Stock	Comment	Score	Count
1	2020-05-31	6.79%	19.65%	19.75%	20.63%
2	2020-06-30	29.32%	13.26%	1.84%	9.39%
3	2020-07-31	32.50%	19.47%	31.64%	22.35%
4	2020-08-31	74.15%	64.85%	69.35%	69.26%
5	2020-09-30	-13.91%	-13.85%	-2.35%	-2.33%
6	2020-10-31	-9.55%	-5.45%	-12.20%	-12.20%
7	2020-11-30	46.27%	34.86%	36.25%	25.68%
8	2020-12-31	24.33%	31.94%	3.75%	3.76%
9	2021-01-31	12.45%	6.96%	7.02%	2.11%
10	2021-02-28	-14.87%	-21.84%	-18.24%	-19.20%
11	2021-03-31	-1.12%	-19.18%	-11.88%	-6.57%
12	2021-04-30	6.21%	9.33%	7.79%	5.35%
13	2021-05-31	-11.87%	-8.14%	-13.98%	-2.46%
14	2021-06-30	8.71%	-0.96%	10.90%	10.98%
15	2021-07-31	1.10%	-6.94%	-3.03%	-3.03%
16	2021-08-31	7.06%	0.57%	7.57%	7.57%
17	2021-09-30	5.40%	-1.38%	-0.00%	-1.76%
18	2021-10-31	43.65%	28.83%	15.66%	16.00%
19	2021-11-30	2.76%	-20.44%	-6.12%	-6.17%
20	2021-12-31	-7.69%	3.36%	-5.03%	-10.32%
21	2022-01-31	-11.36%	-36.36%	-5.73%	-5.71%
22	2022-02-28	-7.08%	9.28%	13.86%	13.85%
23	2022-03-31	23.80%	8.28%	4.63%	0.47%
24	2022-04-30	-19.19%	-2.43%	-7.14%	-7.12%
25	2022-05-31	-12.92%	-21.74%	-20.58%	-20.53%
26	2022-06-30	-11.19%	-13.91%	-9.61%	-9.62%

Fig 4.2.5 *Backtest monthly performance of model on TSLA*

For TSLA, the value of one share is higher than other stocks in the analysis. It can be depicted from the table that the return from prices are relatively stable, but in these months, the model underperforms in most months.

4.2.2 Quarterly results

GME

GME Quarterly Return

	Date	Stock	Comment	Score	Count
1	2020-07-31	-30.02%	-1.90%	-43.79%	-23.00%
2	2020-10-31	161.10%	61.51%	42.71%	59.67%
3	2021-01-31	3004.11%	474.58%	121.27%	175.09%
4	2021-04-30	-46.59%	7.73%	214.31%	63.12%
5	2021-07-31	-7.18%	13.04%	-44.25%	-42.08%
6	2021-10-31	13.90%	-0.60%	5.95%	10.22%
7	2022-01-31	-40.64%	-9.46%	-22.15%	-36.51%
8	2022-04-30	14.82%	-16.40%	23.69%	6.80%

Fig 4.2.6 *Backtest quarterly performance of model on GME*

To show the quarterly results of GME, differing from monthly results, the ‘comments’ model is still the more stable model, with less possibility of a -30-40% quarter.

AMC

AMC Quarterly Return

	Date	Stock	Comment	Score	Count
1	2020-07-31	-17.89%	27.30%	-41.20%	-42.90%
2	2020-10-31	-41.58%	-6.86%	3.97%	2.35%
3	2021-01-31	461.86%	38.84%	-23.34%	-28.98%
4	2021-04-30	-24.36%	4.13%	23.55%	95.43%
5	2021-07-31	269.09%	47.58%	157.23%	97.12%
6	2021-10-31	-4.46%	23.47%	-0.22%	-2.92%
7	2022-01-31	-54.59%	-45.55%	-25.73%	-38.56%
8	2022-04-30	-4.73%	-2.87%	-21.82%	-25.74%

Fig 4.2.7 *Backtest quarterly performance of model on AMC*

AMC results support the observation of the ‘comments’ model outperforming the stock price in bear quarters but fails to capture higher returns.

4.3 Combining models

Tests on using two factors combined consists of trading when the both variables exceed the threshold. However, there was not enough data above both thresholds on the same day to make significant results. Further actions include methods of refining it to become a factor model to trade on.

From all the results, it is also observed, social media actions cannot be the sole indicator to trade on. The number of comments navigates bull markets well while the others could be a good indicator to navigate through a bear market.

Further actions

Limitations

Current stock price data contains columns of OHLCV in daily format, it is hard to measure the effect of social media by hours. However, at the same time, there is not enough social media data for us to measure by hours. Using daily data as a variable is a compromise for both sides.The way to solve this is to increase the amount of social media data. Current data sources of large communities, paid or for free, only include reddit and stocktwits. More ideas of sources include the discussion panel in stock-broker apps like FUTU.

Another limitation is the lack of labelled training data for discussion thread contents' sentiment. There is existing financial sentiment data on news, but the way that comments/posts largely differentiate from news, with sarcasm and abbreviations making it difficult to predict the sentiment.

If there is more data on these two areas, the model can be more complete and increase the accuracy of predict buy & sell.

Build Factor Models

The common factors that are calculated in factor models are market risk, outperformance of small-cap companies to large companies, as well as the value of equity in the companies. It basically compares the factors to understand the market trend and predict returns. Incorporating 'engagement' factor can help increase the accuracy of the model for individual stocks.

Lack of time-wise data

In all the reddit posts collected, there is a limitation of not knowing when the comments or upvotes were given. The underlying assumption groups all of it as the time of posting and within that given day before the market opens.

In addition, in the quick and continuous delivery of discussion threads, the engagement actions should only be recorded for a certain period of time like 30 minutes or even less to allow instantaneous reaction to market price.

Keyword Search Optimization

Currently the model is only able to detect similarity to the word, but provides no insights on whether it is positively or negatively related to it. However, the lack of labelled training data hinders the model from performing such functions.

Currently, content is labelled with tickers based on whether its full name or symbol appears within the text.

Appendix

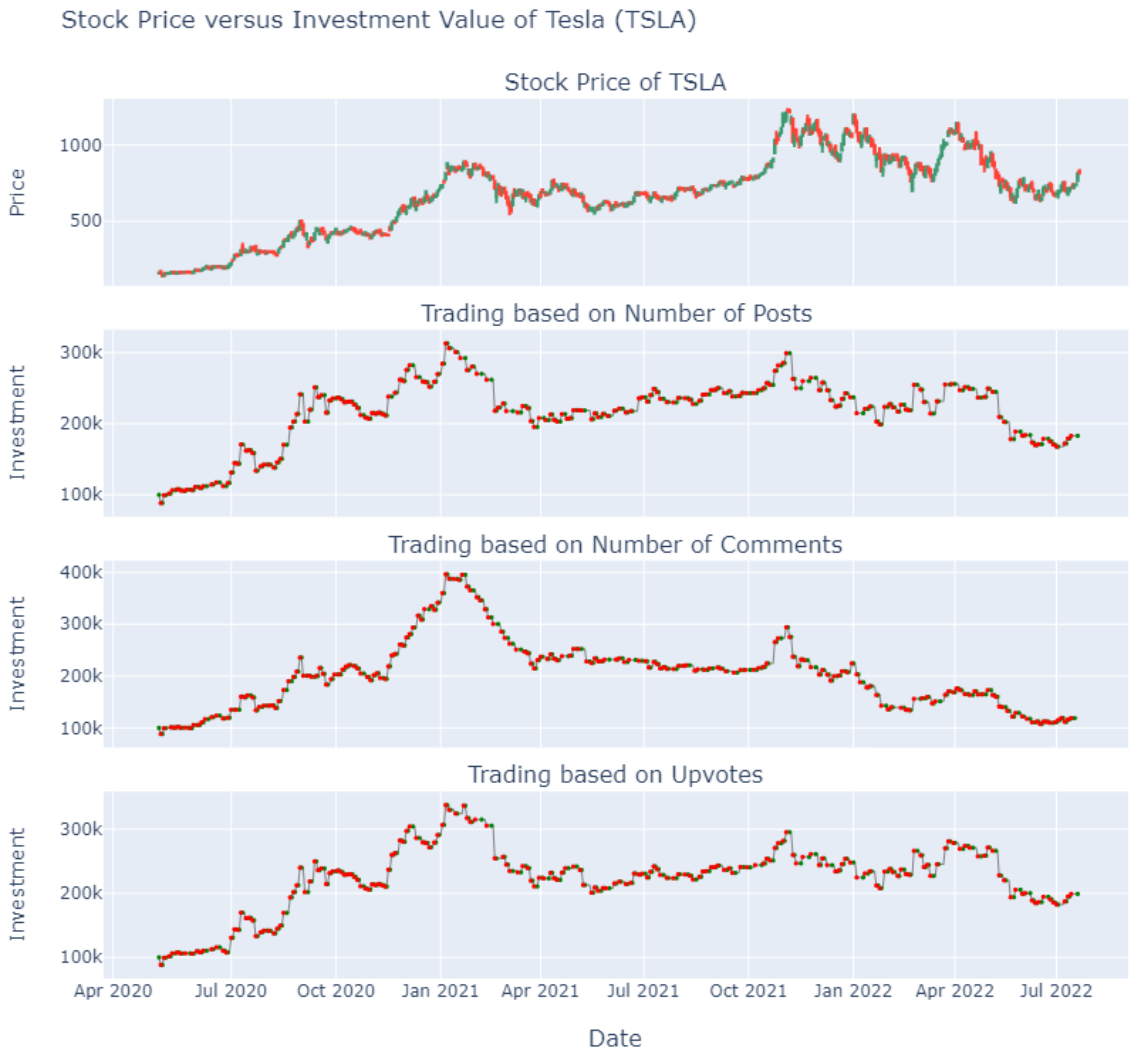


Fig. *Backtest quarterly performance of model on TSLA*

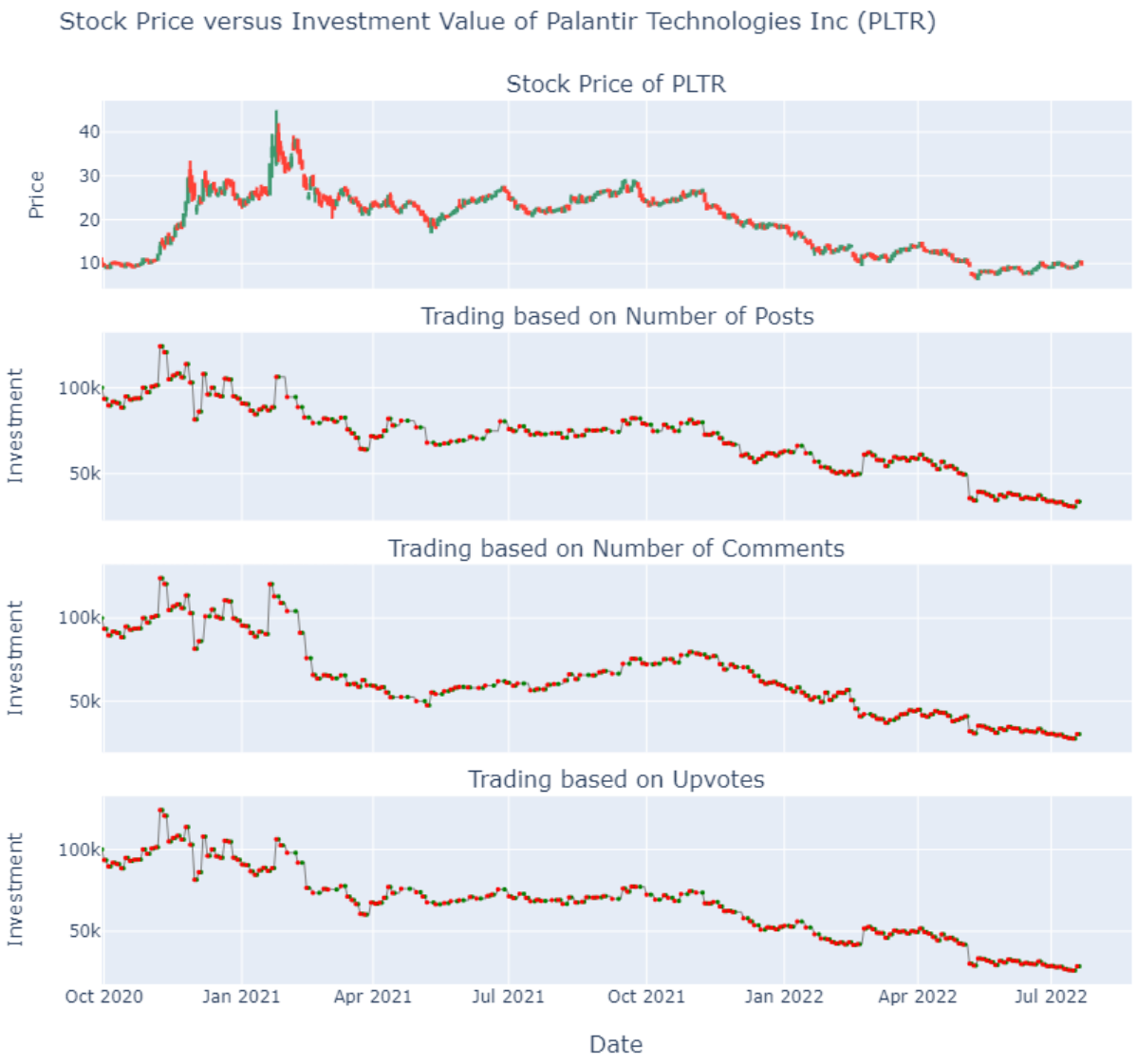


Fig. *Backtest quarterly performance of model on PLTR*

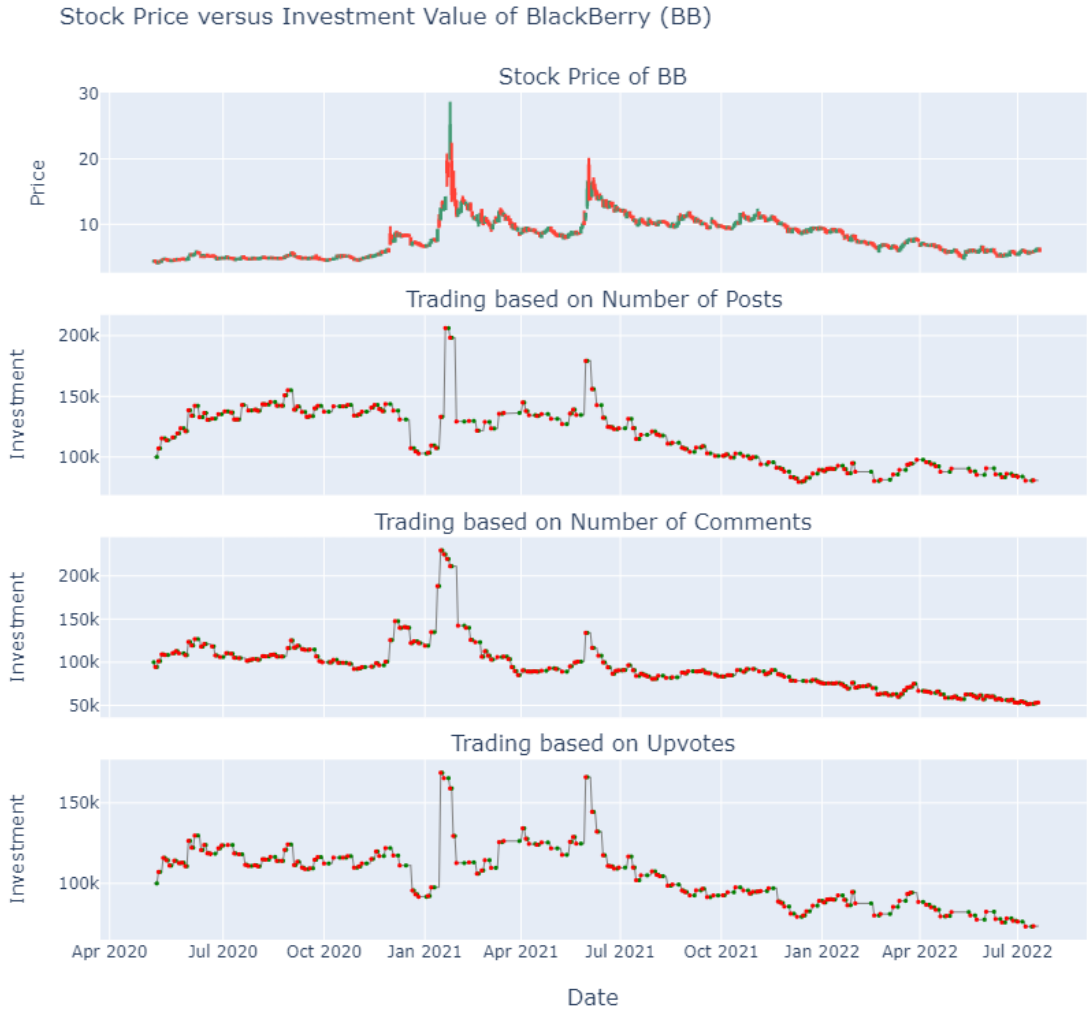


Fig. *Backtest quarterly performance of model on BB*

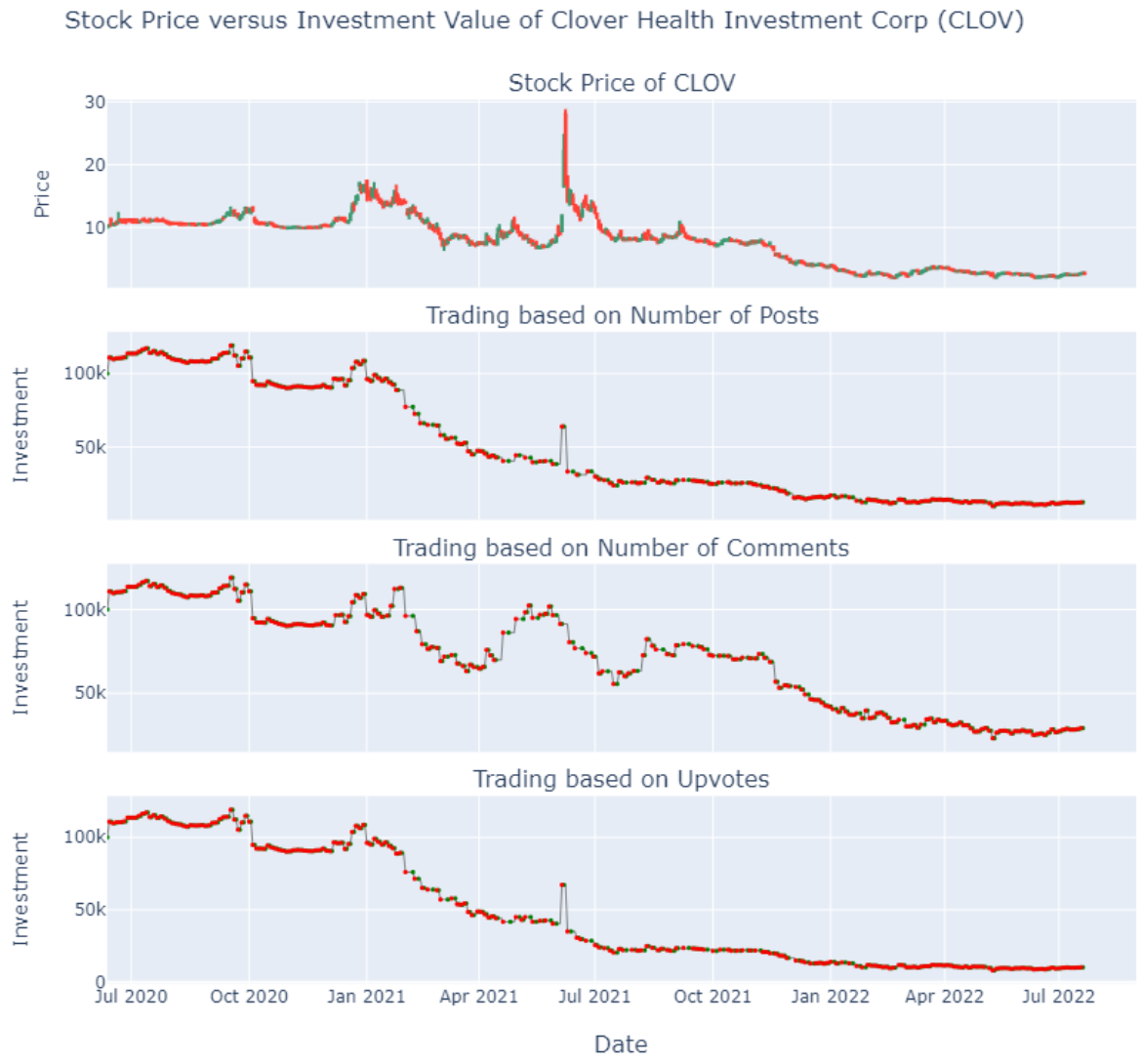


Fig. *Backtest quarterly performance of model on CLOV*