Predicting the Stock Price through r/WallStreetBets Posts

PySights
Simon Gee, Hak Song Lim, Aeri Liu, Sophia Song, Rosa Won

IEOR E4523 Data Analytics
April 22, 2021

Abstract
The WallStreetBets (WSB) subreddit has become notorious this year for its unique brand of financial commentary and potential effect on the stock market. Pysight believes that we can use the WallStreetBets dataset on Kaggle to investigate whether the forum's mood on a given day truly affects the stock market. We begin by running positive/negative sentiment analysis on post titles and bodies, which gives us a picture of WSB sentiment on a given day. We then fit several machine learning models – using features such as WSB sentiment, short interest, and other financial metrics – to try and predict changes in the following day's stock price for GameStop (GME) and several other volatile companies including Tesla (TSLA), AMC Entertainment (AMC), Palantir Technologies (PLTR), and Nokia (NOK).

Background
While the WallStreetBets subreddit has been around for many years, it shot to mainstream prominence in January 2021 as the stock price of GameStop, an American gaming retail company that had been widely sold short, rocketed up from $20 to an intraday high of ~$500. Shorting GameStop had become such a popular trade on Wall Street that, by mid-January 2021, the short interest was more than 140% of GameStop's outstanding shares. The ensuing rapid price rise would force these short sellers, primarily hedge funds, to exit their short position at an enormous loss. Short selling is an investment strategy that speculates on the decline in a stock or other security's price. An investor borrows a stock, sells it on, and then buys it back later - hopefully at a cheaper price - to return it to the lender.

According to media reports at the time, this "short squeeze" was driven by WSB users who would buy shares in GameStop and other heavily shorted companies and never selling (termed "holding the line"), reasoning that if there were no sellers, the price could only go up, thus simultaneously causing short sellers to take huge losses and making money for themselves in the process. Users who had bought shares would then post on the WSB subreddit, which may have then created a positive feedback loop of buying and holding which may have exacerbated the short squeeze. It is this aspect which we will primarily focus on in this project.

For this analysis, we selected five companies (stock tickers in brackets): GameStop (GME), AMC (AMC), Nokia (NOK), Tesla (TSLA), and Palantir (PLTR). GME, AMC, and NOK were three of the "pure" meme stocks that were particularly popular on WSB, with GME chief among them. TSLA and PLTR were chosen as stocks that also experienced explosive price rises, but were not as associated with WSB.

Exploratory Data Analysis
By observing all the posts from WSB, we noticed that out of 44,000 posts, approximately half of the posts were posted on Fridays. This is likely skewed by the fact that the stock price of GME went up significantly on Friday, January 29th. The number of posts from just that one day was 16,000, more than 16 times the number of posts on any other day.

[See Appendix for Figure 1]

Pre-Processing

Titles vs Bodies

We observed that many posts did not have a body associated with them as Reddit posts are often links or images. Although 52% of posts had empty bodies (Figures 2&3), bodies where they exist tend to be much longer than titles, and to discard them would be to ignore a valuable source of sentiment data. We therefore decided to incorporate both the title and the body texts into our analysis. To achieve this, we combined each corresponding title and body into a single string, which we will refer to as a “post” going forward.

Frequency Count & Wordcloud

In order to count the frequency of the words in the posts, we concatenated all the titles and the bodies into a string. Then, we tokenized the words and removed the stop words to get the most frequent words. The extracted keywords in the word clouds in Figure 4 shows that the actions words, such as ‘buy’, ‘like’, and ‘share’ are most frequently used in the posts.

Sentiment Analysis

Positive/Negative

In order to find out whether the sentiment in the posts affect the stock prices, we performed sentiment analysis by date on posts about each stock and generated positive and negative scores. To determine whether a post was about a given stock, we filtered for posts containing the stock’s ticker, its name, or mentions of its CEO if high-profile enough (e.g. Elon Musk for Tesla). As the plots in Figures 5-9 illustrate, we did not observe any immediately obvious correlation between daily stock price and positive/negative sentiment.

We therefore decided to include the net sentiment (pos - neg) and absolute sentiment (pos + neg) as features as well, reasoning that if the mood of WSB was overwhelmingly one way or the other in the case of net, or highly polarised in the case of absolute, the stock price could be affected. While we also considered using emotion scores, we concluded that the WSB environment was driven more by raw positive or negative sentiment than by complex emotions, so distinguishing between, say, negative emotions (e.g. fear and anger) would be meaningless.

Machine Learning

Feature Selection

We attempt to use the daily positive and negative sentiment scores for each stock as features in a machine learning model to predict the next day’s percentage change in stock price. Along with the raw positive and negative sentiment scores, we added net and absolute sentiment as features for reasons discussed above. We also include the daily number of posts as a feature since more users discussing a particular stock could be driving higher volatility.

In addition, we also incorporate basic financial data available through the Yahoo Finance API. This includes fields such as daily opening price, closing price, volume traded. Using these fields we calculated each stock’s close-to-close percentage change from the previous day. We also calculated each stock’s 10 day realised volatility, a measure of the average magnitude of the

daily movement of a stock in percentage terms over the previous 10 trading days, which we reasoned could be predictive of at least the magnitude of the next day's move.

Since one of the most important themes discussed in the reddit posts were the short sales of certain companies, we also included short interest volumes as additional variables. We wanted to find out if a high volume of shorts could have a negative effect on the increment of the future prices stocks. Intuitively it should serve as a strong variable predicting the decrease in stock values, so we wanted to prove this hypothesis. The features are: short volume, short exempt volume and other related variables such as percentage and changes on these volumes.

Models Without Short Volume

Random Forest

We first decided to fit a model with the sentiment analysis scores but without short volumes as a control. After some experimentation, we found that a Random Forest was the optimal model to use for this purpose. We decided to take two separate approaches here: first we tried to classify the direction of the next day's percentage move (up or down) using a Random Forest Classifier; and second we tried to predict the actual percentage move (e.g. -4%) using a Random Forest Regressor. The results are shown in Figures 10 and 11. The classifier clearly outperformed the regressor in terms of out of sample F1 score, with the TSLA, AMC and PLTR classifiers performing best of all. Significantly, the sentiment score features appeared near the top of the feature importance rankings for GME and AMC, perhaps indicating that WSB sentiment is predictive of price movement for these stocks.

Models With Short Volume

1) Regression Models

We have conducted regression analysis using linear regression, ridge regression, and lasso regression for the 5 datasets. We tried to predict the next day percentage change of the closing price using the features described above. However, as seen from the result of the linear regression, all three methods performed very badly at predicting the next day percentage change of the close price. See Figures 12 and 13.

All three methods' out of sample R2 scores were similar, by looking through the coefficients of regression we see that lasso regression couldn't take the effect of sentiment analysis into account to predict the percentage change. Although ridge regression considered the effect of sentiment analysis compared to lasso regression, its effect was very weak. In contrast, in linear regression, the effect of sentiment analysis was huge compared to other two methods. Results were slightly better when predicting next day's trading volume.

2) Decision Tree

Even though we had limited sentiment data, the model was able to predict the stock changes with a fair prediction rate as seen in Figure 14. Sentiment indicators actually have a strong influence when we are trying to predict the future stock price changes, which might prove our hypothesis that sentiment factors can actually predict the future stock changes. In addition, we observe that more sentiment data leads to an improved prediction model in the case of GME in

Figure 14 which was the one with the best prediction for both training and test dataset. This result could sound obvious, however considering the fact that more sentiment data doesn't add up more variable columns and since the sentiment indicators are normalized by the total amount of dataset, the results on improving prediction rate was unexpected. We also found that short volumes were a good indicator to predict the stock prices, some are being more affected and others less. In our project TSLA was one of the companies most affected by this variable. A larger time span would provide greater confidence, but this provides us with an initial positive outcome to investigate further.

Conclusions

Our team combined sentiment analysis and 3 machine learning techniques to utilize WSB subreddit posts data in predicting stock price changes. As we continued modeling, we confronted the limitations and found possible further improvements for this project.

1) Size of the dataset too small
Since we used The WallStreetBets (WSB) subreddit dataset from 2021-01-28 to 2021-03-31, the data we could use for predicting was roughly only two months. Moreover, we divided the dataset into 80% of the training set and 20% of the test set so only 5-7 data points were included in the test set. If there were enough data to train the model and test it, we could've drawn better results with higher accuracy.

2) Difficulty cleaning the text data
WSB subreddit dataset are the text data which are written by users with a lot of abbreviations, slang, vulgarities, and emojis. We did our best to clean the data before going onto sentiment analysis, but we couldn't catch every single word that shouldn't be included in the analysis. Also, some phrases (abbreviations or swear words for example) that are critical to determine the tone of the text weren't included in the analysis. If we can find a better way to clean the data, we would be able to get more accurate results of sentiment analysis.

3) No comments in dataset
Our WSB dataset didn't include the comments. However, comments play a critical role in reddit posts and are as important as the title and body of the posts (sometimes even considered more critical). If we did sentiment analysis including comments, we might be able to draw more accurate results for the positiveness and negativeness of the posts, which might result in better prediction for stock price, trading volumes, or percentage changes.

Even with the above limitations, the decision tree model showed that the sentiment factors can predict the changes in the stock prices. We believe that if the limitations could be improved in the future, we could further optimize our models to yield better predictions in the stock price changes.

Appendix

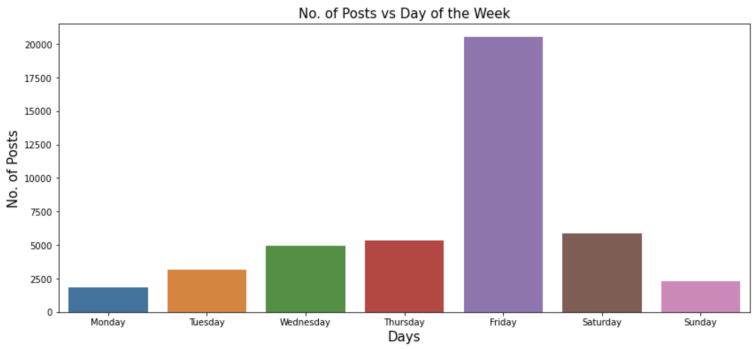


Figure 1: Bar Chart of Number of Posts vs Day of the Week

title	body
It's not about the money, it's about sending a...	NaN
Math Professor Scott Steiner says the numbers ...	NaN
Exit the system	The CEO of NASDAQ pushed to halt trading “to g...
NEW SEC FILING FOR GME! CAN SOMEONE LESS RETAR...	NaN
Not to distract from GME, just thought our AMC...	NaN
...	...
A M C YOLO Update — Feb 25, 2021	NaN
Hold the line you diamond-handed apes!	NaN
Did I do it right by not selling when I was up...	NaN
Rocket Companies ($RKT), who owns Rocket Mortg...	Rocket Companies (ticker RKT) is Rocket Mortga...
That didn't go exactly as I had planned	NaN

Figure 2: Title Texts vs Body Texts

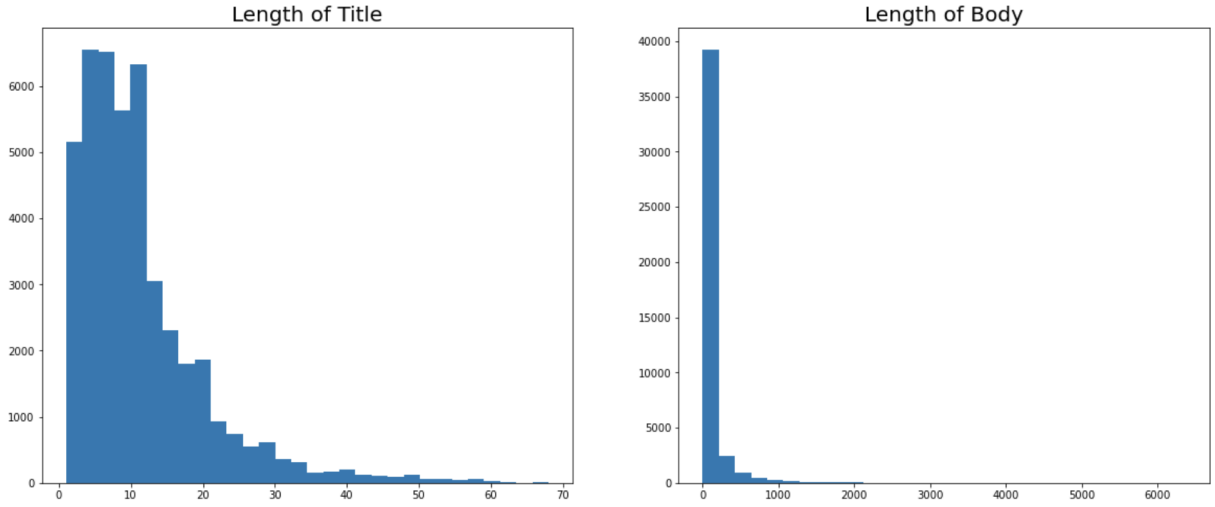


Figure 3: Length of Title & Length of Body

Figure 4: WordClouds for Posts on GME & AMC & NOK & TSLA & PLTR

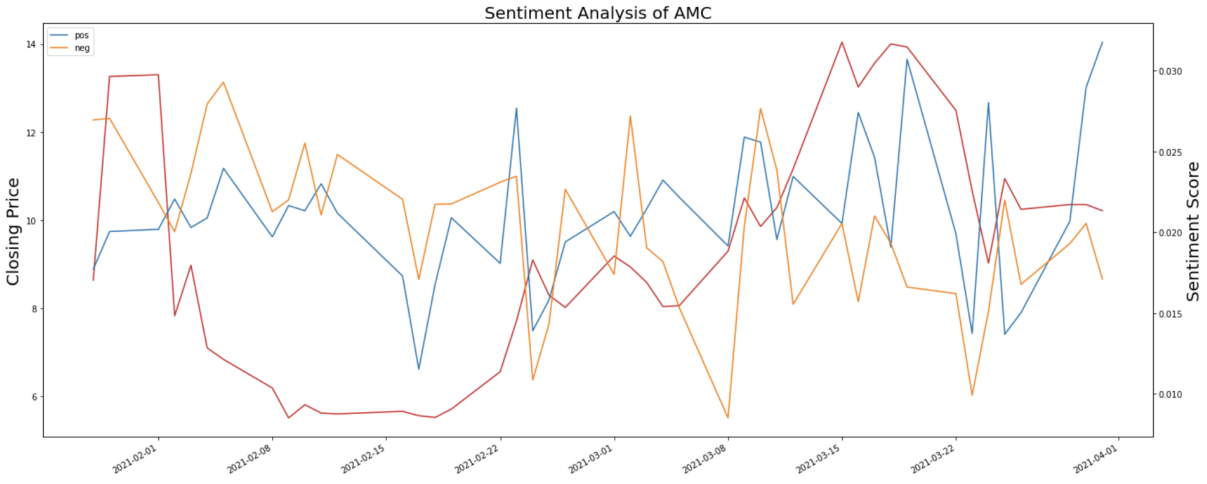


Figure 5: AMC daily positive (blue) and negative (orange) sentiment overlaid with daily closing price (red)

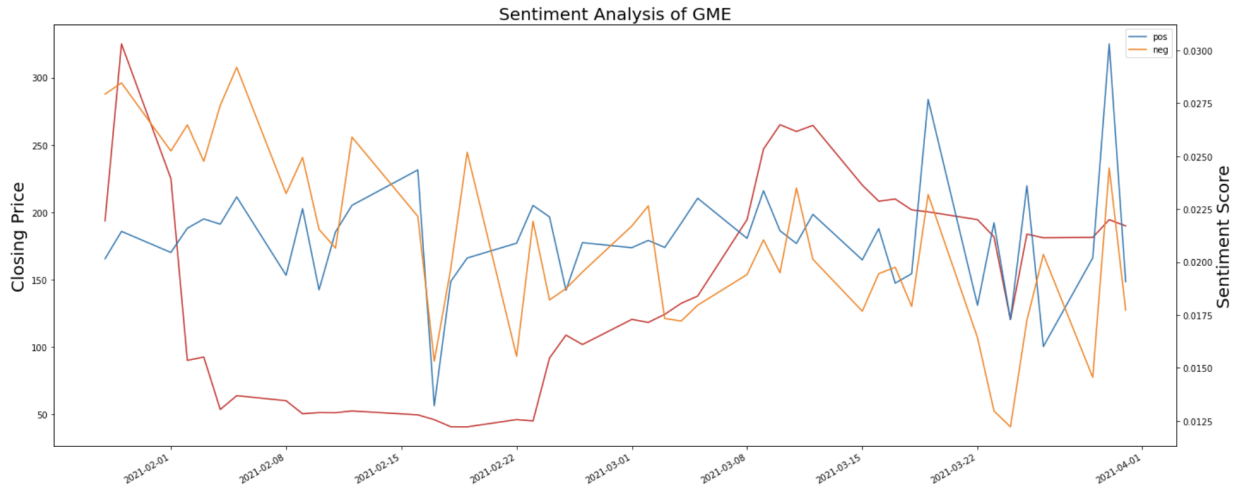


Figure 6: GME daily positive (blue) and negative (orange) sentiment overlaid with daily closing price (red)

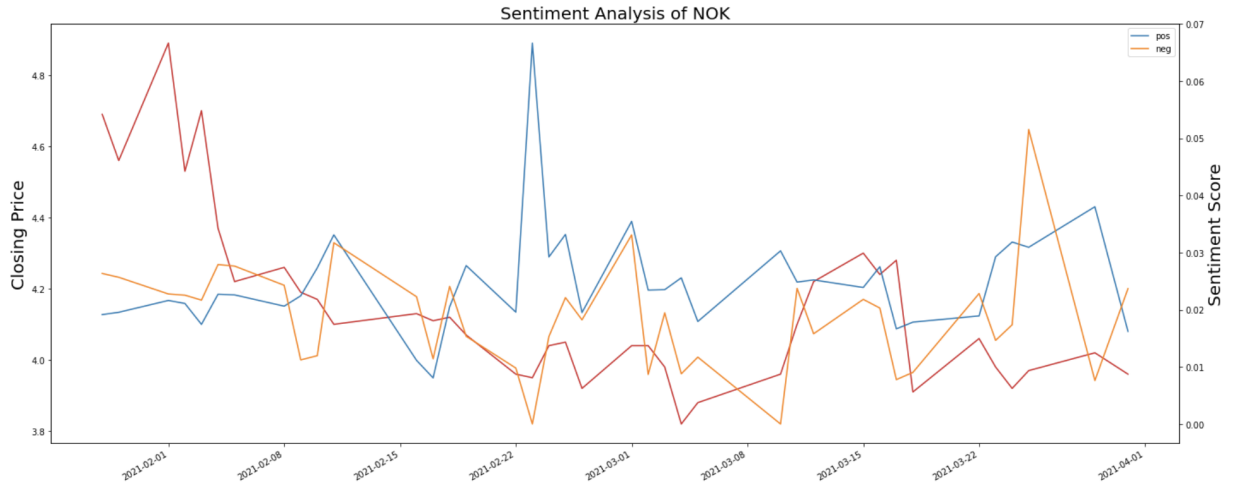


Figure 7: NOK daily positive (blue) and negative (orange) sentiment overlaid with daily closing price (red)

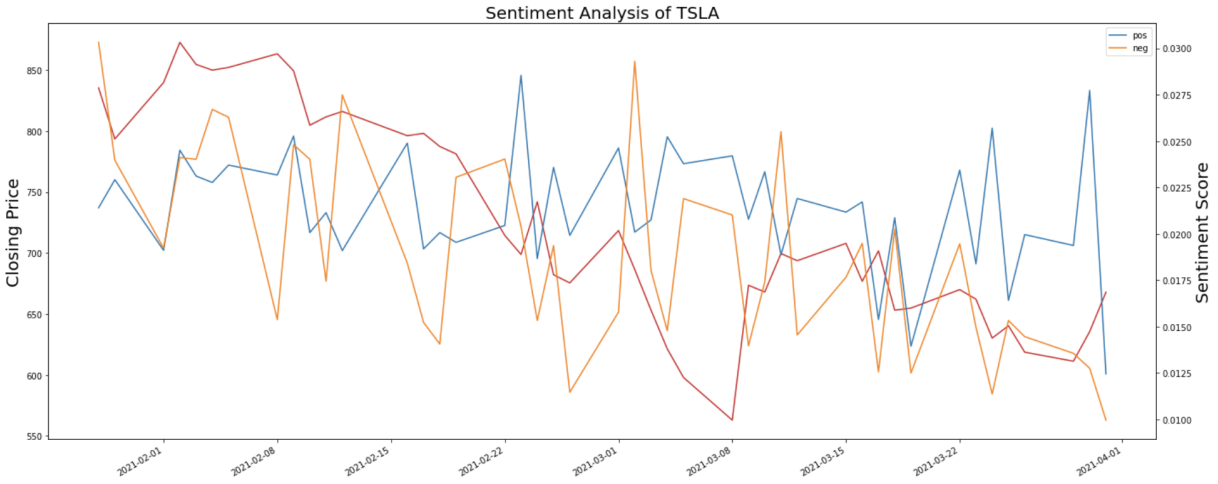


Figure 8: TSLA daily positive (blue) and negative (orange) sentiment overlaid with daily closing price (red)

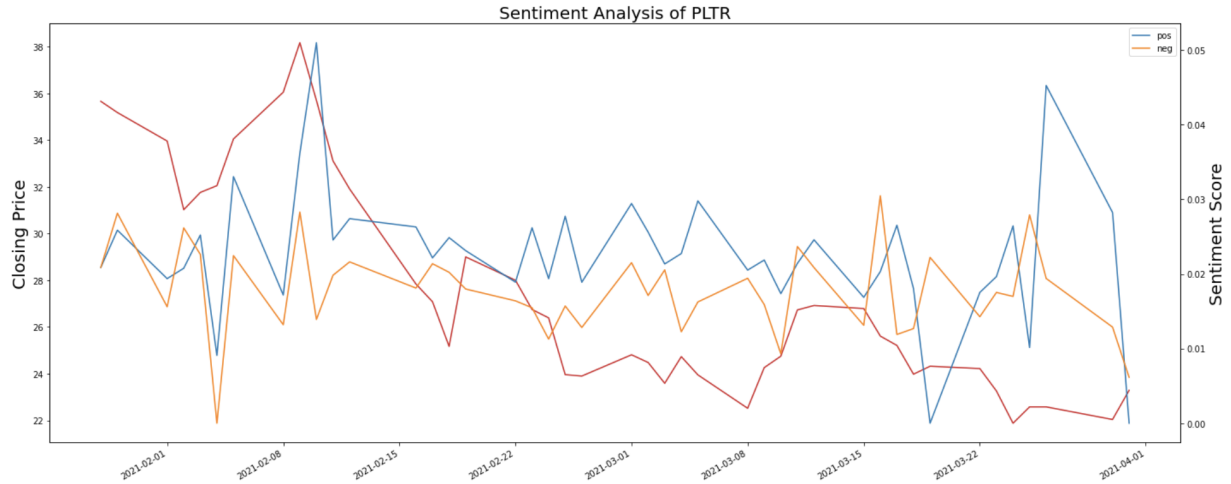


Figure 9: PLTR daily positive (blue) and negative (orange) sentiment overlaid with daily closing price (red)

GME

	down_next	prediction
2021-03-18	1	1
2021-03-19	1	0
2021-03-22	1	1
2021-03-23	1	0
2021-03-24	0	1
2021-03-25	1	0
2021-03-26	0	1
2021-03-29	0	1
2021-03-30	1	0

GME	
Train	0.9926
Validation	0.6933
Test	0.3636

NOK

	down_next	prediction
2021-03-16	0	1
2021-03-17	1	0
2021-03-18	0	1
2021-03-22	1	1
2021-03-23	1	1
2021-03-24	0	0
2021-03-25	1	0
2021-03-29	1	0

NOK	
Train	8296
Validation	0.6743
Test	0.4444

PLTR

	down_next	prediction
2021-03-17	1	1
2021-03-18	0	1
2021-03-19	1	0
2021-03-22	1	1
2021-03-23	1	1
2021-03-24	0	1
2021-03-25	0	0
2021-03-26	0	1
2021-03-30	0	1

PLTR	
Train	0.8623
Validation	0.7687
Test	0.5454

AMC

	down_next	prediction
2021-03-18	1	1
2021-03-19	1	0
2021-03-22	1	0
2021-03-23	1	1
2021-03-24	0	0
2021-03-25	1	1
2021-03-26	0	1
2021-03-29	0	1
2021-03-30	1	0

AMC	
Train	0.7651
Validation	0.3543
Test	0.5454

TSLA

	down_next	prediction
2021-03-18	0	1
2021-03-19	0	1
2021-03-22	1	1
2021-03-23	1	1
2021-03-24	0	0
2021-03-25	1	1
2021-03-26	1	1
2021-03-29	0	1
2021-03-30	0	1

TSLA	
Train	0.8249
Validation	0.5709
Test	0.6666

GME

	Relative Importance
net	0.139913
pct_change_c2c	0.126846
abs	0.116919
count	0.114564
pos	0.111352
pct_change_intraday	0.097615
Adj Close	0.085316
Volume	0.078358
Open	0.065773
neg	0.033559
10d_realised_vol	0.029786

NOK

	Relative Importance
Open	0.3
10d_realised_vol	0.2
net	0.2
pct_change_intraday	0.1
pos	0.1
neg	0.1
Adj Close	0.0
Volume	0.0
pct_change_c2c	0.0
count	0.0
abs	0.0

PLTR

	Relative Importance
Open	0.17
pos	0.17
10d_realised_vol	0.12
Adj Close	0.08
Volume	0.08
pct_change_intraday	0.08
neg	0.07
net	0.07
abs	0.07
pct_change_c2c	0.06
count	0.03

AMC

	Relative Importance
Open	0.2
10d_realised_vol	0.2
net	0.2
abs	0.2
pct_change_intraday	0.1
neg	0.1
Adj Close	0.0
Volume	0.0
pct_change_c2c	0.0
count	0.0
pos	0.0

TSLA

	Relative Importance
Volume	0.16
pct_change_intraday	0.16
pos	0.16
count	0.12
10d_realised_vol	0.12
Open	0.08
neg	0.08
pct_change_c2c	0.04
net	0.04
abs	0.04
Adj Close	0.00

Figure 10: Results of using a Random Forest Classifier to predict direction of the next day's stock price movement (top), model feature importances with sentiment-based features (bottom)

GME

	pct_change_c2c_next	prediction
2021-03-18	-0.733579	0.893665
2021-03-19	-2.886103	2.972726
2021-03-22	-6.550468	7.164587
2021-03-23	-33.788173	25.555849
2021-03-24	52.692376	2.963072
2021-03-25	-1.496599	2.972726
2021-03-26	0.165748	0.893665
2021-03-29	7.258689	5.964194
2021-03-30	-2.386094	2.972726

GME	
Train	0.4912
Validation	-1.9751
Test	-0.5623

NOK

	pct_change_c2c_next	prediction
2021-03-16	0.943407	-0.296927
2021-03-17	-8.644862	-0.124667
2021-03-18	0.995024	0.623622
2021-03-22	-1.970441	0.105145
2021-03-23	-1.507536	0.105145
2021-03-24	1.275509	0.105145
2021-03-25	-0.985221	0.105145
2021-03-29	-0.751879	0.105145

NOK	
Train	0.4408
Validation	-0.4672
Test	-0.1771

PLTR

	pct_change_c2c_next	prediction
2021-03-17	-4.879015	0.500446
2021-03-18	1.417849	0.948797
2021-03-19	-0.411186	2.203025
2021-03-22	-3.963663	1.092634
2021-03-23	-5.932937	1.199598
2021-03-24	3.199272	1.087306
2021-03-25	0.000000	2.700619
2021-03-26	1.426606	1.387416
2021-03-30	5.671506	1.403749

PLTR	
Train	0.3693
Validation	-0.4105
Test	-0.1868

AMC

	pct_change_c2c_next	prediction
2021-03-18	-0.499998	1.080578
2021-03-19	-10.337405	1.726359
2021-03-22	-14.651721	2.405640
2021-03-23	-15.384610	2.405640
2021-03-24	21.286020	3.140719
2021-03-25	-6.398536	5.400928
2021-03-26	1.074225	2.405640
2021-03-29	0.000000	2.405640
2021-03-30	-1.352660	3.301268

AMC	
Train	0.4470
Validation	-0.4047
Test	-0.2803

TSLA

	pct_change_c2c_next	prediction
2021-03-18	0.261808	-1.274837
2021-03-19	2.310383	-1.434665
2021-03-22	-1.170153	-1.035285
2021-03-23	-4.816050	-1.616481
2021-03-24	1.605660	-0.737551
2021-03-25	-3.385436	1.435597
2021-03-26	-1.199277	-1.145465
2021-03-29	3.980111	-0.174304
2021-03-30	5.083225	5.861647

TSLA	
Train	0.4633
Validation	-0.2169
Test	0.1516

GME

	Relative Importance
Volume	0.32
10d_realised_vol	0.30
Open	0.18
net	0.10
neg	0.04
Adj Close	0.02
pct_change_intraday	0.02
pos	0.02
pct_change_c2c	0.00
count	0.00
abs	0.00

NOK

	Relative Importance
Adj Close	0.50
pct_change_c2c	0.30
Open	0.10
pct_change_intraday	0.05
count	0.05
Volume	0.00
10d_realised_vol	0.00
pos	0.00
neg	0.00
net	0.00
abs	0.00

PLTR

	Relative Importance
Open	0.35
Volume	0.21
pct_change_c2c	0.11
pos	0.11
Adj Close	0.09
abs	0.04
pct_change_intraday	0.03
neg	0.03
net	0.02
10d_realised_vol	0.01
count	0.00

AMC

	Relative Importance
Open	0.31
pct_change_intraday	0.22
net	0.19
pct_change_c2c	0.14
10d_realised_vol	0.06
Volume	0.05
Adj Close	0.03
count	0.00
pos	0.00
neg	0.00
abs	0.00

TSLA

	Relative Importance
Open	0.34
Adj Close	0.30
count	0.11
pct_change_intraday	0.09
Volume	0.05
pos	0.04
net	0.04
abs	0.02
neg	0.01
pct_change_c2c	0.00
10d_realised_vol	0.00

Figure 11: Results of using a Random Forest Regressor to predict the next day's stock price movement (top), model feature importances with sentiment-based features (bottom)

Regression Summary

	Coefficient
Intercept	-13.0203
Open	-0.9710
Adj Close	1.0124
Volume	0.0000
pct_change_c2c	-1.0064
pct_change_intraday	0.1572
10d_realised_vol	-0.0185
count	0.0174
pos	663.0208
neg	-618.2092
net	1281.2300
abs	44.8116
Short Volume	-0.0000
Short Exempt Volume	-0.0000

	pct_change_c2c_next	y_pred
15	13.328406	-10.028898
9	-0.195317	0.402323
0	67.871896	-12.809426
8	1.769031	-5.479930
17	103.935950	-1.690757

Coefficient R-squared: -1.1194028292931102

Linear_GME

	Coefficient
Intercept	6.4738
Open	-3.1854
Adj Close	4.0222
Volume	0.0000
pct_change_c2c	-0.6800
pct_change_intraday	0.1481
10d_realised_vol	-0.0385
count	0.0212
pos	0.1506
neg	-0.0270
net	0.1777
abs	0.1236
Short Volume	-0.0000
Short Exempt Volume	0.0000

	pct_change_c2c_next	y_pred
27	-6.190473	7.901501
15	14.912288	6.654095
23	-6.410259	9.897316
17	18.051953	-12.311198
8	5.454549	-7.748006
9	-3.275863	0.429616
29	8.560313	1.499786

Coefficient R-squared: -1.8057767651585195

Ridge_AMC

Regression Summary

	Coefficient
Intercept	12.2383
Open	-0.0000
Adj Close	-0.0000
Volume	0.0000
pct_change_c2c	-0.5620
pct_change_intraday	0.2307
10d_realised_vol	-0.0401
count	0.0192
pos	0.0000
neg	0.0000
net	0.0000
abs	0.0000
Short Volume	-0.0000
Short Exempt Volume	0.0000

	pct_change_c2c_next	y_pred
27	-6.190473	3.099450
15	14.912288	7.352618
23	-6.410259	8.242005
17	18.051953	-13.171007
8	5.454549	-7.857763
9	-3.275863	1.241379

Coefficient R-squared: -1.623404038837812

Lasso_AMC

Figure 12: Best results of regression analysis of stock price changes

Regression Summary

	Coefficient
Intercept	2.324374e+08
Open	2.195313e+08
Adj Close	-2.209618e+08
Volume	2.545772e-01
pct_change_c2c	-2.796653e+06
pct_change_intraday	2.718242e+07
10d_realised_vol	1.121612e+05
count	1.162424e+05
pos	-3.795626e+09
neg	9.183351e+08
net	-4.713961e+09
abs	-2.877291e+09
Short Volume	-8.606832e-02
Short Exempt Volume	-8.410083e-01

	Volume_next	y_pred
15	173409000.0	9.807395e+07
9	55482400.0	1.368011e+08
0	602193300.0	4.729043e+08
8	152810800.0	1.094091e+08
17	376881800.0	1.190025e+08

Coefficient R-squared: 0.4896648884825041

Linear_AMC

	Coefficient
Intercept	1.690169e+08
Open	-5.433346e+06
Adj Close	2.625362e+06
Volume	1.049032e+00
pct_change_c2c	-2.205361e+06
pct_change_intraday	1.474291e+06
10d_realised_vol	-4.298922e+05
count	1.211355e+06
pos	-8.182140e+05
neg	-6.024482e+05
net	-2.157658e+05
abs	-1.420662e+06
Short Volume	-1.827757e+00
Short Exempt Volume	-6.445593e+01

	Volume_next	y_pred
8	45177200.0	5.229734e+07
13	313175100.0	1.532847e+08
9	51863200.0	6.043696e+07
21	73539700.0	1.093115e+08
0	42030900.0	1.043437e+07
11	180294300.0	7.822355e+07

Coefficient R-squared: 0.35376290079434625

Ridge_PLTR

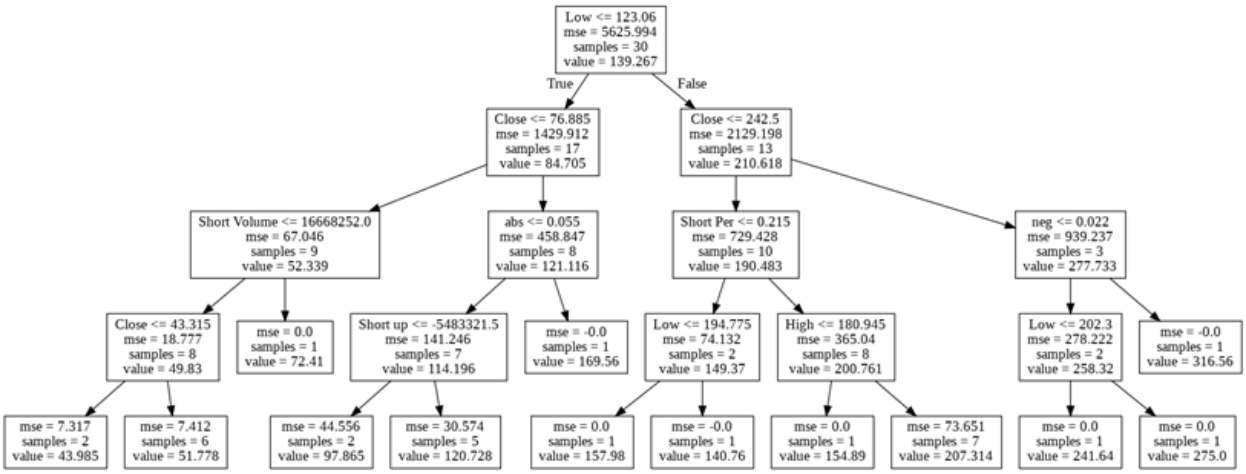
	Coefficient
Intercept	9.253586e+07
Open	-1.731374e+05
Adj Close	9.024958e+04
Volume	5.500000e-02
pct_change_c2c	-4.266229e+05
pct_change_intraday	-1.441264e+06
10d_realised_vol	1.652478e+05
count	-2.135923e+03
pos	-5.583411e+08
neg	1.378702e+08
net	-6.293629e+07
abs	-1.157941e+07
Short Volume	1.137000e-01
Short Exempt Volume	-2.680000e+00

	Volume_next	y_pred
8	36216100.0	1.743534e+07
16	66606900.0	3.886657e+07
0	34990800.0	1.476470e+07
24	89396500.0	4.702480e+07
11	19802300.0	2.538459e+07
9	21622800.0	2.876204e+07

Coefficient R-squared: 0.10207841244491755

Lasso_TSLA

Figure 13: Best results of regression analysis of volume traded

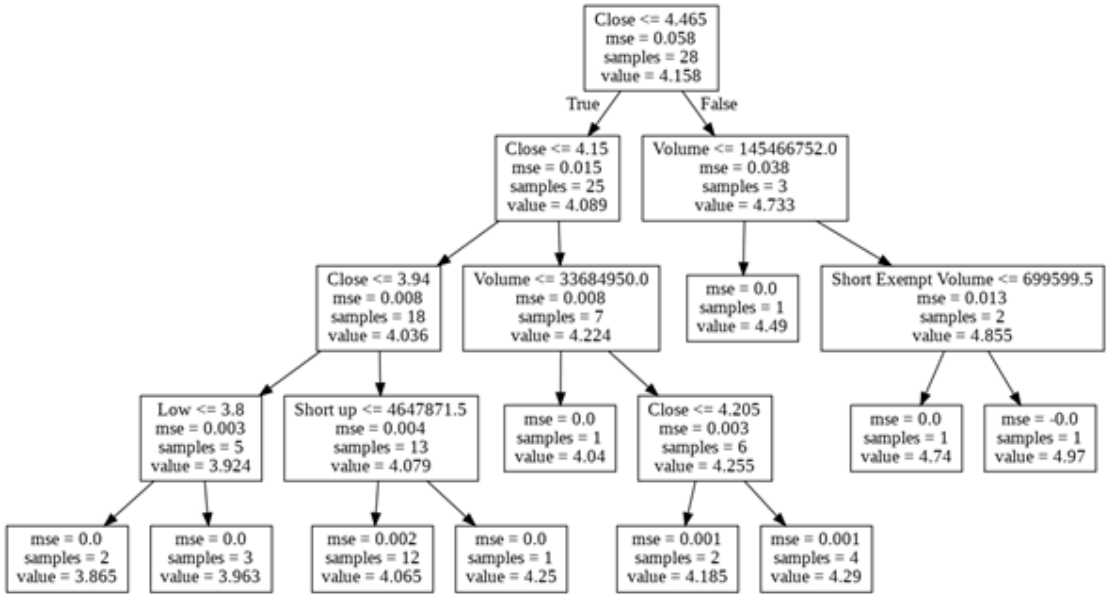


GME	
Train	0.9951
Validation	0.9699

Stock prediction: 20210326	
Prediction	Next Open
217.8399	197.5000

Up or Down?	
Prediction	Actual
Up	Up

Decision tree: GME

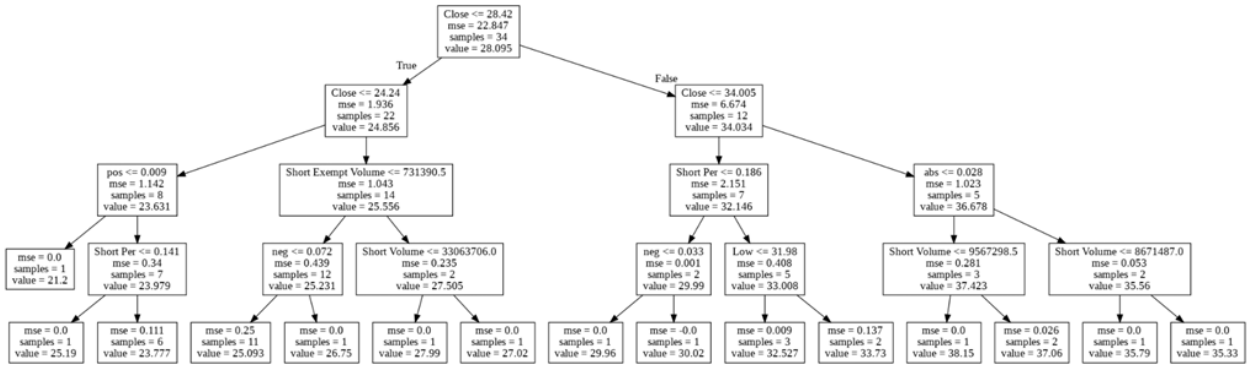


NOK	
Train	0.9854
Validation	0.5724

Stock prediction: 20210324	
Prediction	Next Open
3.94	3.92

Up or Down?	
Prediction	Actual
Down	Down

Decision tree: NOK

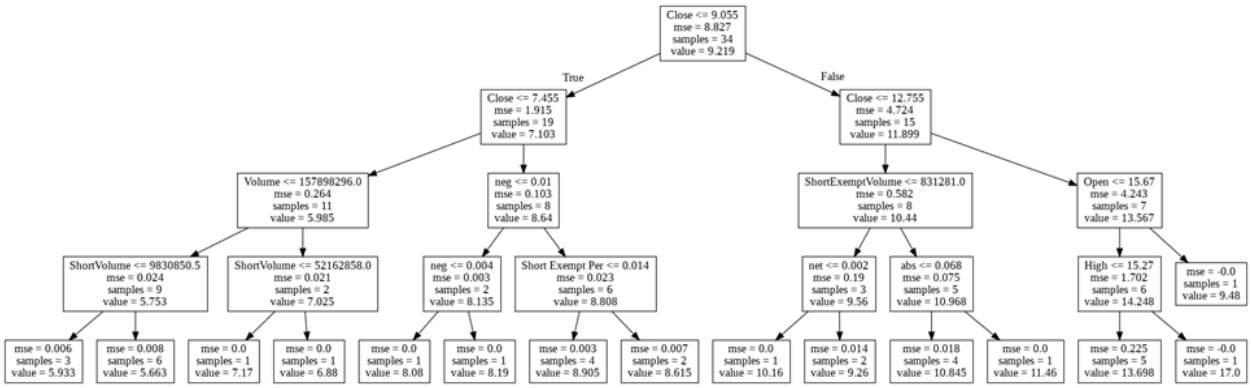


PLTR	
Train	0.99513
Validation	-1.1960

Stock prediction: 20210326	
Prediction	Next Open
23.2999	22.48

Up or Down?	
Prediction	Actual
Up	Down

Decision tree: PLTR

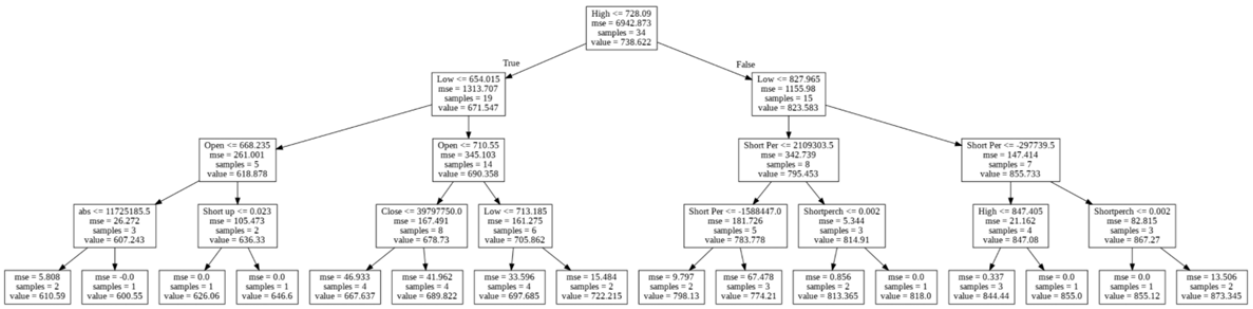


AMC	
Train	0.9956
Validation	0.3240

Stock prediction: 20210326	
Prediction	Next Open
9.38	10.32

Up or Down?	
Prediction	Actual
Down	Down

Decision tree: AMC



TSLA	
Train	0.9966
Validation	-1.5518

Stock prediction: 20210326	
Prediction	Next Open
600.55	615.64

Up or Down?	
Prediction	Actual
Down	Down

Decision tree: TLSA

Figure 14: Results of decision tree analysis