COMP 4971C Report

The development of a trading strategy, using the influence of COVID-19 on FANG+ companies

Natarajan Siddarth Jayakumar

Year 2

Supervisor: Professor David Rossiter

Abstract

This study originally aimed to determine the effect of COVID-19 upon FANG+ stocks using reddit post analysis and COVID-19 statistics. However, due to a lack of a good numerical representation the use of reddit post analysis was voided. Although, with the use of COVID-19 statistics and time series forecasting models a profitable trading strategy was developed. The trading strategy built upon single step time series models produced higher profits (up to 250% for period studied) and losses (as low as -55% for period studied) than a buy and hold strategy, while a trading strategy built upon multi-step time series models produced a more stable profit (up to 15% for period studied) and losses (as low as -1% fore period studied) compared to a buy and hold strategy.

Contents

Table of Figures

Table of Tables

1. Introduction

1.1 Motivation

COVID-19 has changed our society to its very core, due to it, the very air we breathe is questioned of its safety on a constant basis. Thus, many of us had to isolate from each other. However, due to our society's reliance on collaboration to grow, technology has become an ever more important part of our lives as it provides a safe way to connect all of us together. As a result, many tech companies have reached levels of wealth that have never been seen before.

From this unimaginable growth and an unspeakable pandemic came the idea to see the extent of this COVID-19 pandemic on the tech industry, and where better to look than FANG+. This is because they are the most coveted and highly traded tech stocks, as of now.

Since these stocks are all sold on NYSE/Nasdaq which is centred in New York, it seemed best to only focus on the COVID-19 statistics in that area. Furthermore, as COVID-19 has pushed more people into chat forums like reddit as a means of socializing it seemed interesting to also consider how future stock evaluations could be determined by people's discussion within the forum. To be specific, the reddit post titles on the subreddit (subsection of the reddit chat forum) r/wallstreetbets. This was chosen because of the infamous GameStop scandal [1] which revealed the potency of the users on this subreddit upon the stock market.

1.2 Hypothesis

The initial hypothesis is that the COVID-19 statistics would provide a reasonable prediction of the stock market data. However, due to the complexity of processing required for the reddit data, it may not serve as a good means to predict future stock evaluations.

2. Data Management

2.1 Data Collection

The first data that was collected was the US COVID-19 statistics. This is because, the range of market and reddit data that would be considered in this study is restricted to the range of time where there were COVID-19 cases in New York. The COVID-19 data (specifically the daily cases, deaths, total cases, and total deaths) was made available, and collected, from the CDC via the SODA API for python (source 1). From the data collected it was found that the first COVID-19 case that occurred in New York was on the 23rd of January 2020, which was then used as start point of reddit and stock data collection. Then 28th May 2021 was chosen as the end date for model training, and validation, as it provided for at least 1 year of data for model training and validation while allowing a long enough range (from 28th May 2021 to 28th August 2021) for testing the trading strategy.

Using this fixed timeline of data, the 'yfinance' python API was used to collect the adjusted closing prices of FANG+ stocks within the stock market (specifically on NYSE and Nasdaq) [2], and was collected at a frequency of one price evaluation an hour. The reason for considering the adjusted closing and not the normal closing price, is that it takes into consider other illicit factors that could have affected the closing price at a given time, such as stock splits. Thus, adjusted closing price values can be viewed as the closing price filtered of certain extraneous variables [3].Moreover, since the stock market has limited working hours of 9 am to 4:30 pm on working weekdays, there is a very limited amount of stock data that can be used to train the model. Thus, to get a larger and more continuous data set pre-market (4 am to 9 30 am) and after-market data (4pm to 8pm) was considered. [4]

For collecting reddit post data, the popular PSAW python API was used. This reddit post data, however, could not be used directly and required more processing. This is because it consisted of a several metadata attributes and requires a numerical representation for any sort of processing to be done upon it.

2.2 Data processing

To obtain a numerical representation of the data, the PSAW API was used to abstract all the metadata attributes for each post. They were the title, the time of posting, the score of the post (which is proportional to the number of people who upvoted the post), and the comments. Using this metadata information, an algorithm to associate a numerical representation of a reddit post was devised. It consisted of applying a semantic analysis upon the post title, and then using the score attribute of the post for further filtering if needed. The semantic analysis of the post title was done using the NLTK library, specifically the pretrained 'vader' semantic analyser. This is due to the immense popularity of the NLTK library, and because the 'vader' semantic analyser was trained upon twitter tweets which follow a similar semantic and syntax to reddit post titles. The semantic analyser returns three key attributes being the positive, negative, and compound. For this case the compound value only was considered, and then upon it the score value of the post was multiplied to get a scaled numerical representation of the reddit post. This numerical representation of the reddit posts in the time frame was plotted together with the covid statistics, and closing prices of stocks, to see if any visible trends could be spotted, shown in Figure 1.

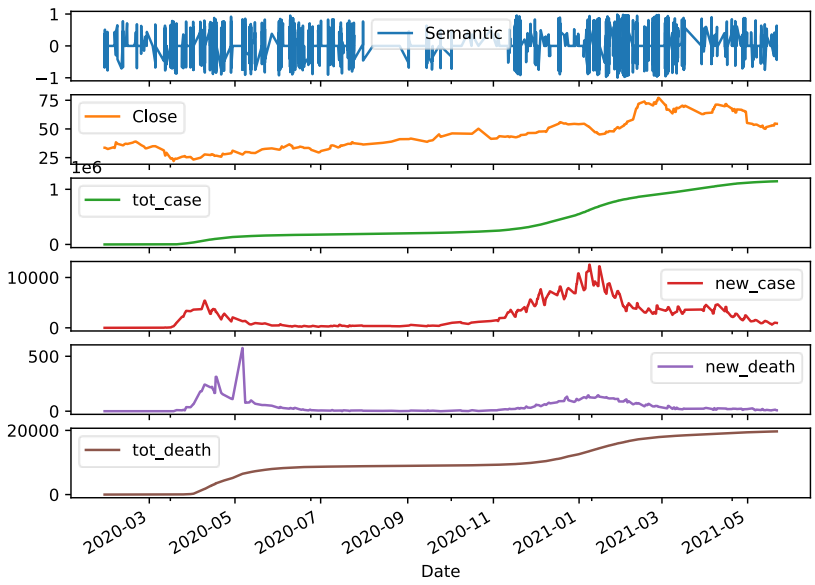

Figure 1: all data in timeline plotted together using Twitter stock as example

As seen from the figure above, there was a very strong similarity in the trend between the COVID-19 statistics and the close price of Twitter stock (used here for only for reference). However, it is also clearly visible that the semantic data is simply too noisy to use, thus further filtering needed to be applied upon it. Therefore, a filtering of semantic data was done to only consider posts with a score value greater than 1, as many posts did not have that. On top of which, a moving average (considering 5 data points at a time) was applied upon the semantic data [5]. A subsection of the result of it is shown in Figure 2.

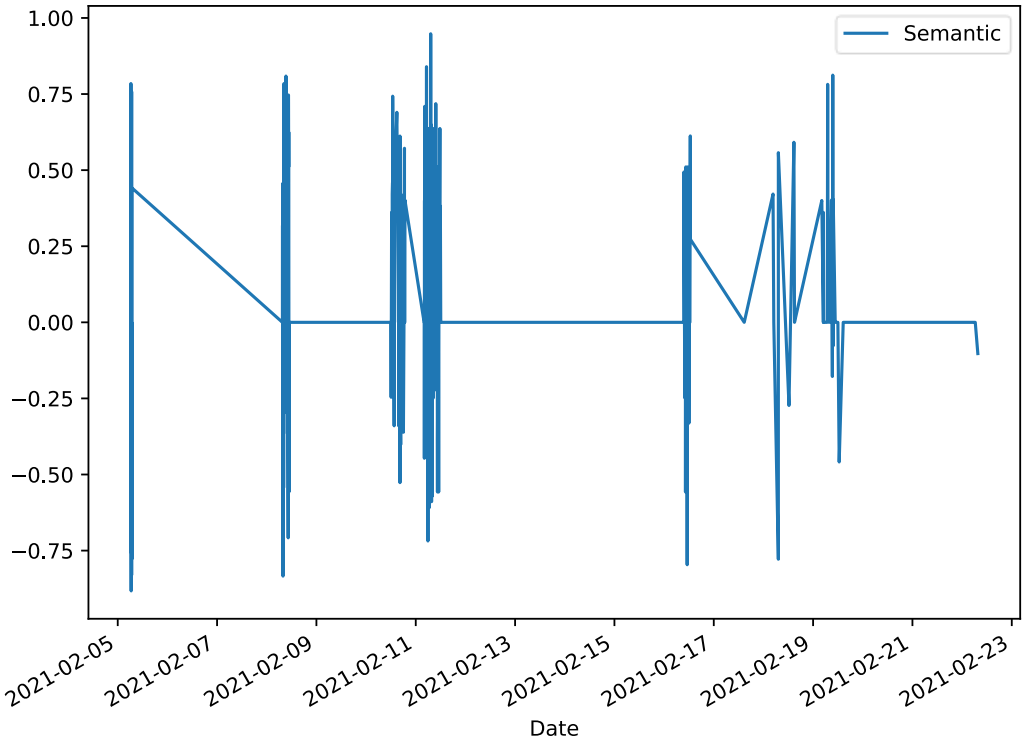

Figure 2: subsection of reddit data post processing

As seen in Figure 2, even by applying the further data filtering and processing upon the reddit post semantic analysis, we still end up having data that is simply too noisy to be used. Thus, as predicted by the initial hypothesis, reddit post data however interesting lacks an accurate numerical representation to be used.

3. Model Development

To make a good trading strategy it is required to first make a model with good predictions. To do this the TensorFlow library was used. This was because, it's user-friendly API allows for quick development of different types of models. Furthermore, when run on Google Collab's TPU (Tensor Processing Unit) accelerator (a google made hardware accelerator designed to process tensor objects more efficiently), the time required for training is decreased significantly. [6]

The models developed for the trading strategies are time series forecasting models, which look at the trend of a data to make predictions. Specifically, two types of time series forecasting models were developed. They are the single step and multi-step models. The single step model works by taking a singular input (set of features), to produce a single output (prediction). The multi-step model works by taking a set of inputs, to produce a sequence of predictions.

The single step models developed were: -

- Linear
- Recurrent Neural Network (Long Short-Term Memory/ LSTM)
- Residual Recurrent Neural Network (LSTM)

The multi-step models developed were: -

- Multi-linear
- Multi-Recurrent Neural Network (using LSTM)
- Auto-Regressive Neural Network (using LSTM)

3.1 Model Preparation

Before training the model, all the data was first normalized and extraneous values were removed, to allow for a better training quality. Furthermore, all the data was verified to be sorted in terms of linear time, as the order in which the data is passed into the model for training can affect the prediction given. In addition, all models underwent a training with 5 epochs (represents the number of times the entire dataset is passed into the model). The epochs were limited to also prevent over-fitting, which would lead to high precision of prediction on trained data values and very low precision of prediction on new input.

Moreover, the comparisons of single step models and multi-step models were done separately for higher effectiveness. On one hand, to determine the effectiveness of the single step models a baseline model, which would have an output equal to its input, was used. On the other hand, a multi-step baseline model, which uses repeats its last input value as the output series, was used to determine the effectiveness of multi-step models. Furthermore, all the single step models were parsed with three random sections of 100 consecutive data points to visually determine its accuracy of prediction. While the multi-step models were provided with three sets of 24 consecutive data points to use as inputs to generate a series of predictions.

3.2 Baseline

An implementation of this model which takes one input to produce one prediction, using Twitter stock. (Note that in the following diagrams normed is short for normalised)

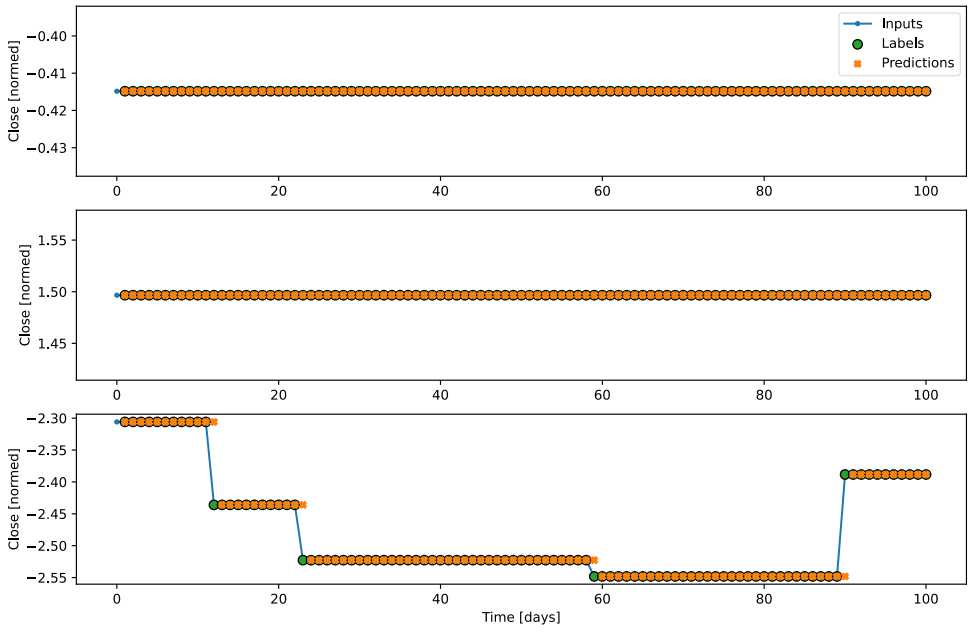

Figure 3: continous single prediction baseline model

3.3 Linear Model

The Linear Model works by trying to predict the very next timestep using the current time-step, and it does this by associating a weight to each of the features (that is the covid-19 statistics). An implementation of it with one prediction on Twitter stock, is shown below.

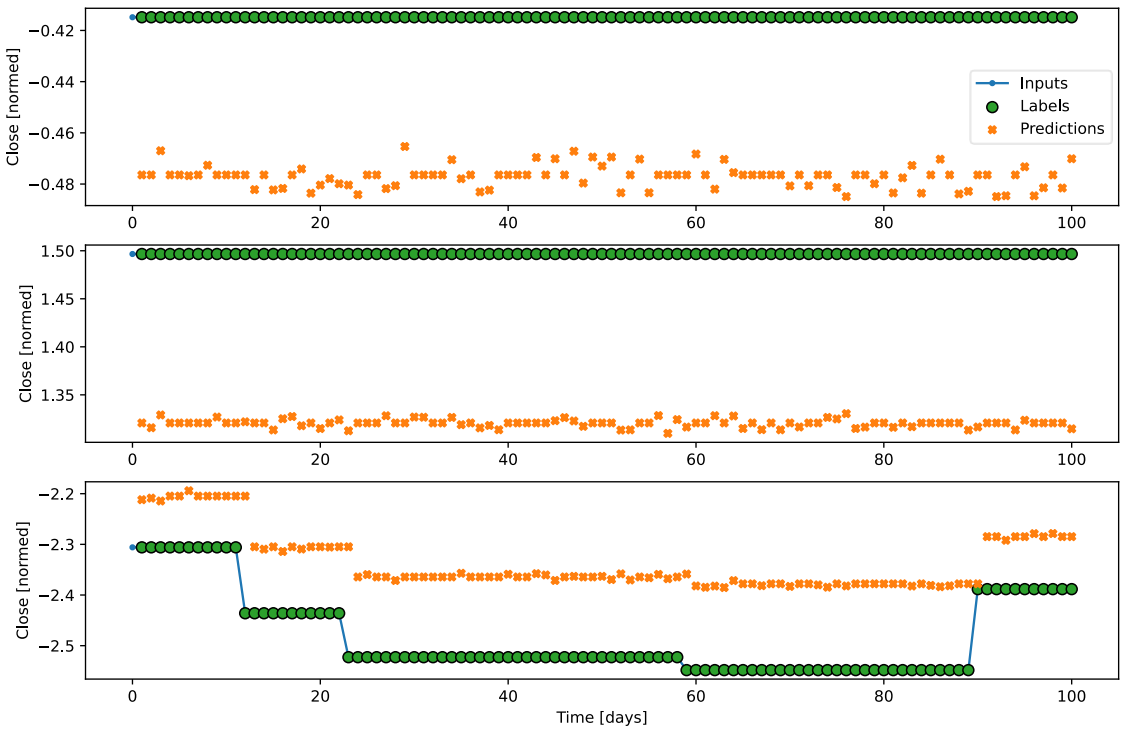

Figure 4: continous single prediction linear model

3.4 RNN (Recurrent Neural Network)

Recurrent Neural Networks, specifically LSTM (Long-Short Term Memory), works by having an internal state that will constantly be updated with each input it takes, and each time an input is provided an output prediction is returned, below is a figure showing an implementation of it on Twitter stock.

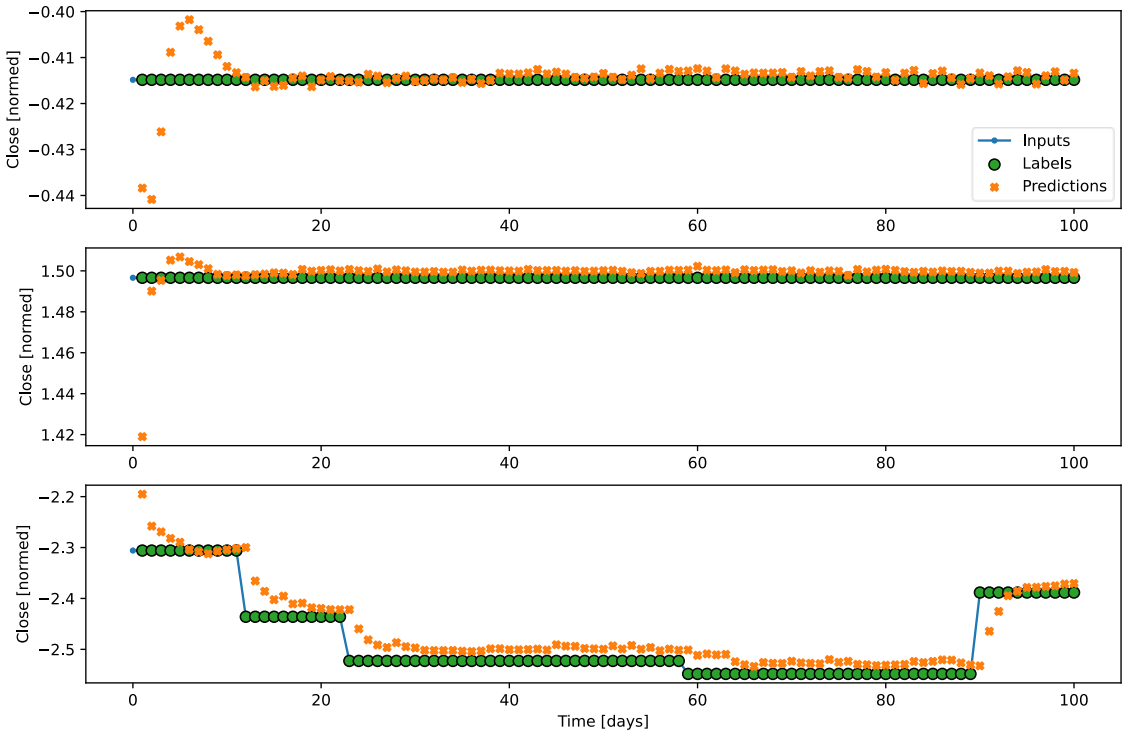

Figure 5: continous single prediction RNN(LSTM) model

3.5 Residual RNN (Recurrent Neural Network)

Furthermore, the Residual RNN is an improved version of the RNN by adding a residual wrapper. The residual wrapper will try to predict the difference of input and prediction between timesteps, instead of generating a new prediction completely. This is done in hopes of producing a more accurate model. An implementation of this on Twitter stock is shown in the figure below.

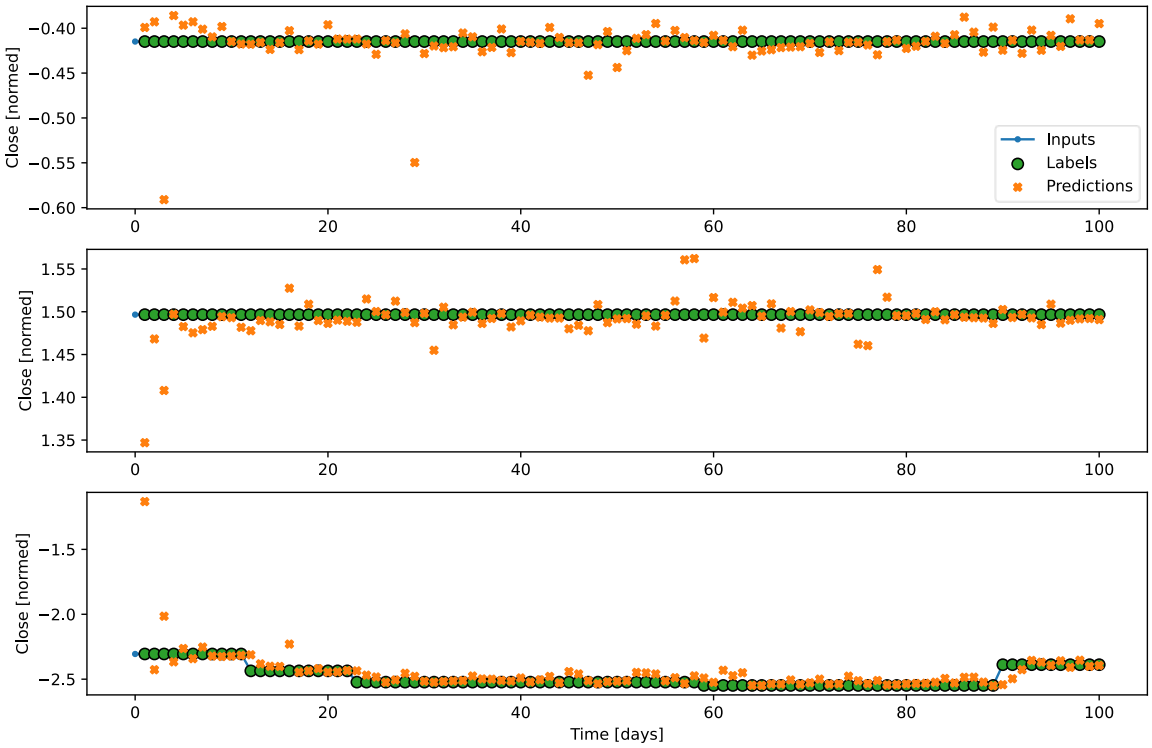

Figure 6: continuous single prediction RNN(LSTM) with residual wrapper

3.6 Multi-Baseline Model

The multi-baseline model works similarly to a Baseline model, as it uses the last input as a prediction/output. However, instead of returning one prediction it returns a sequence of predictions, of the exact same input value. An implementation of this on Twitter stocks is shown below.

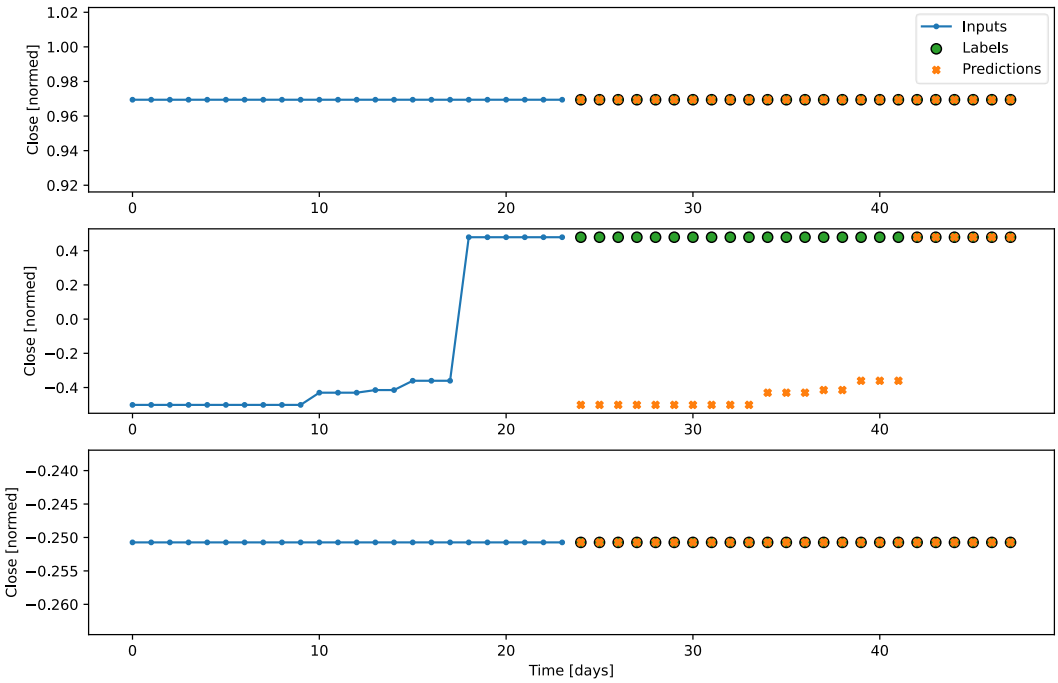

Figure 7: multiple predictions baseline

3.7 Mutli-Linear Model

The multi-linear model is like the single step linear model, in the sense that it uses an input to generate weights, which in turn is used for a prediction. However, it differs by using an additional layer called a Dense Layer. This layer allows the model to condense multiple inputs into one value which is used to make the weights for the prediction. Then the model is made to produce a sequence of predictions based upon these weights, instead of a single prediction like that of the linear model. An implementation of this model on Twitter stocks is shown below.

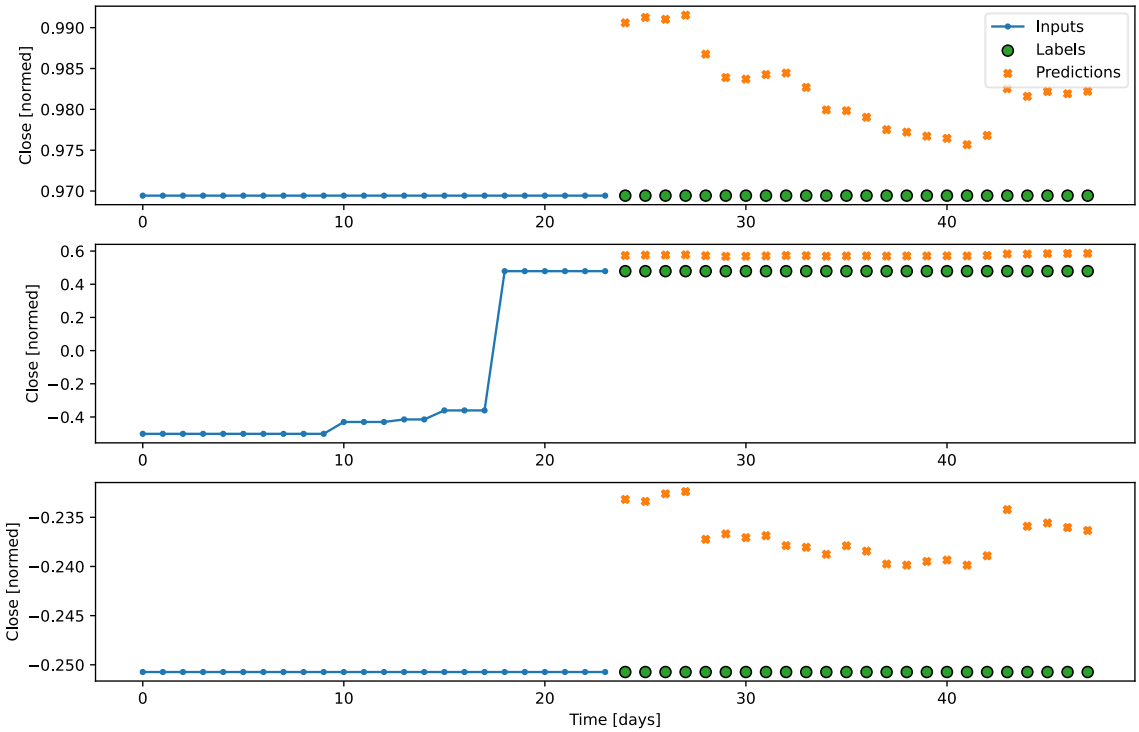

Figure 8: multiple predictions linear model

3.8 Mutli-LSTM Model

Like the difference between the multi-linear model and linear model, the only key difference between the multi-LSTM model and LSTM model is the addition of a Dense layer. This dense layer is used to condense multiple inputs into a single value, which is then used to set the internal state of the LSTM model. From which, this model is made to produce a sequence of predictions, instead of a single prediction like that in the LSTM model. An implementation of this model on Twitter stocks is shown below.

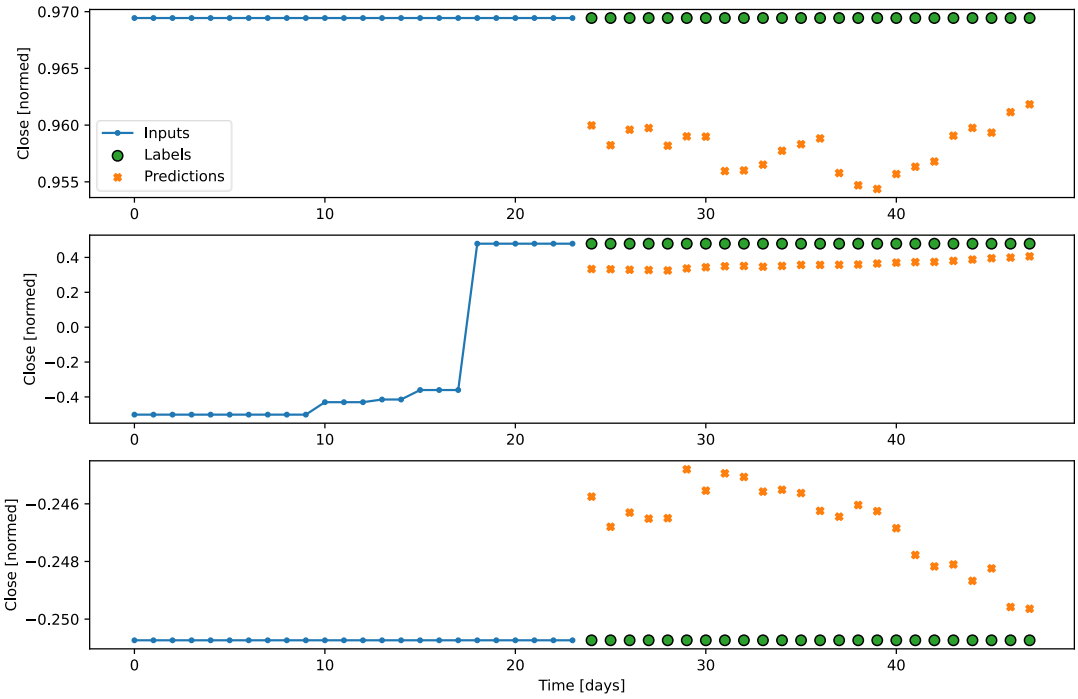

Figure 9: multiple predictions RNN(LSTM) model

3.9 Auto-Regressive Neural Network

Unlike the other models before, the auto-regressive model takes into consideration its last output prediction, and the next incoming input, to generate a more accurate output prediction series. To be more specific, the model used was an Auto Regressive version of the LSTM. This has been implemented on the Twitter stock shown in the figure below.

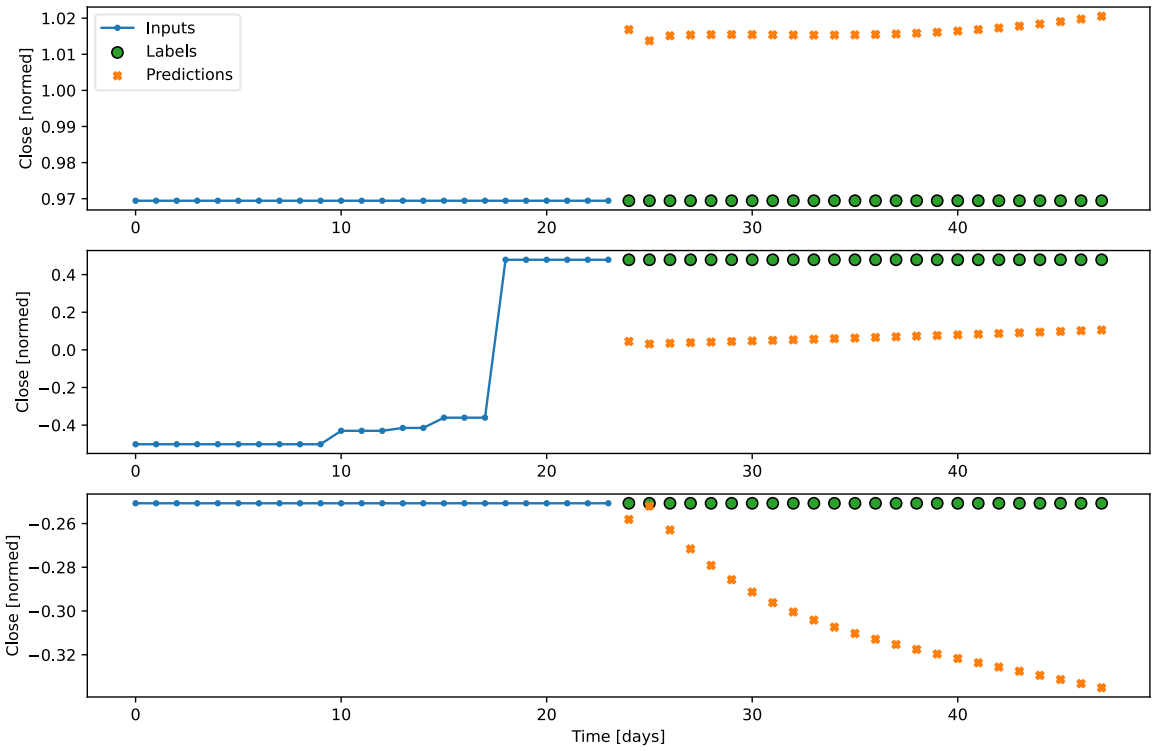

Figure 10: multiple predictions of Auto-Regressive Neural Network, built upon RNN(LSTM)

4. Trading Strategy and Model Evaluation

A simulation of the trading strategy followed certain key constraints. The first was that during one simulation run it could only trade one stock using one model of predictions. The models used were all the time-series models developed except for the baseline and multi-baseline, as the baseline models were only used for reference of comparison of other model predictions and would provide no benefit in trading.

Furthermore, in each simulation run it would start with 100,000 US Dollars and would always stick with the same trading rule. The rule used was a modified of Cabot's loss-limit [7]. Instead of using the 80-buy and 20-sell a 60-buy and 40-sell was used, while a region in between was set for holding. This is because, a limited region of stock data is used for trading, and the general fluctuation in stock value is not high enough for any selling or buying to occur. At the same time during one buy/sell action only one individual stock of that company could be bought or sold.

Moreover, the drawdown in prediction were also recorded. This is because, it is important to consider user willingness of following a trading strategy, which would be unlikely with a high drawdown. Below are the graphs of results from the trading simulation (Tables of Data, and Trades in appendix).

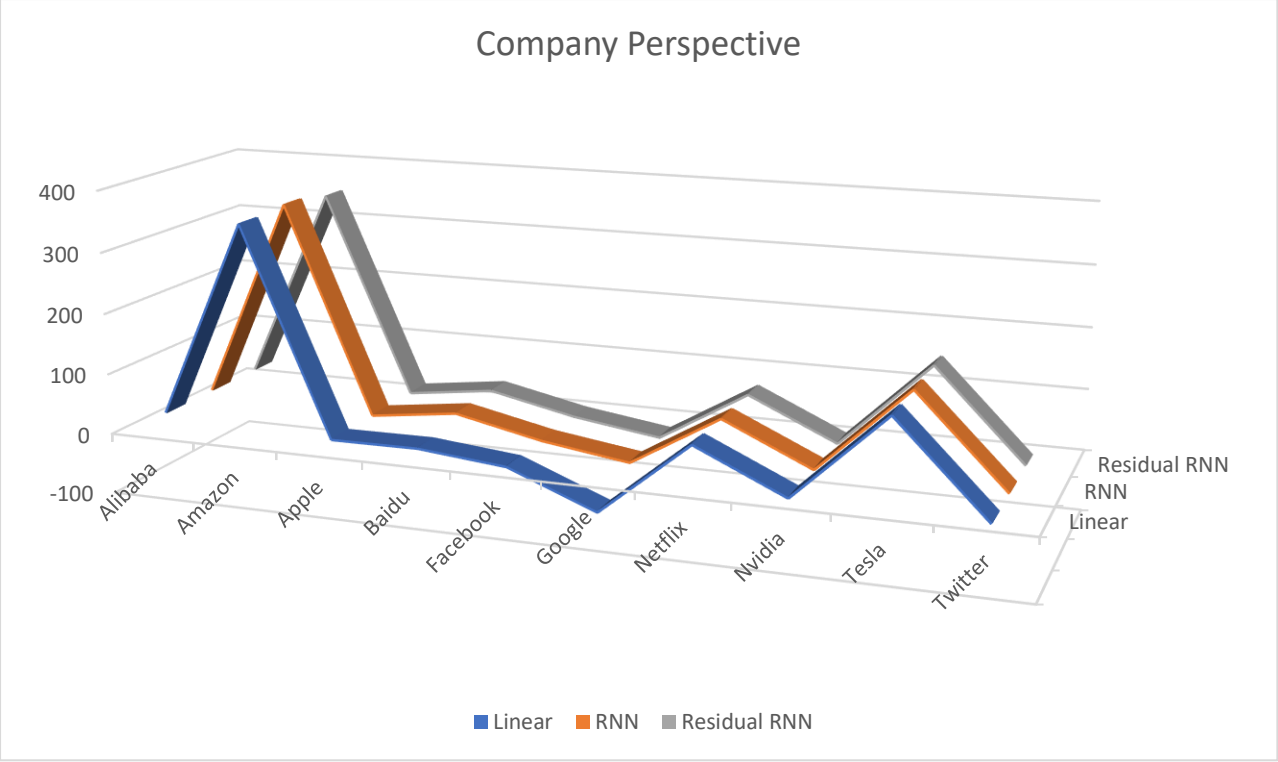

Figure 11: single-step model profit (%) (company perspective) (for period studied)

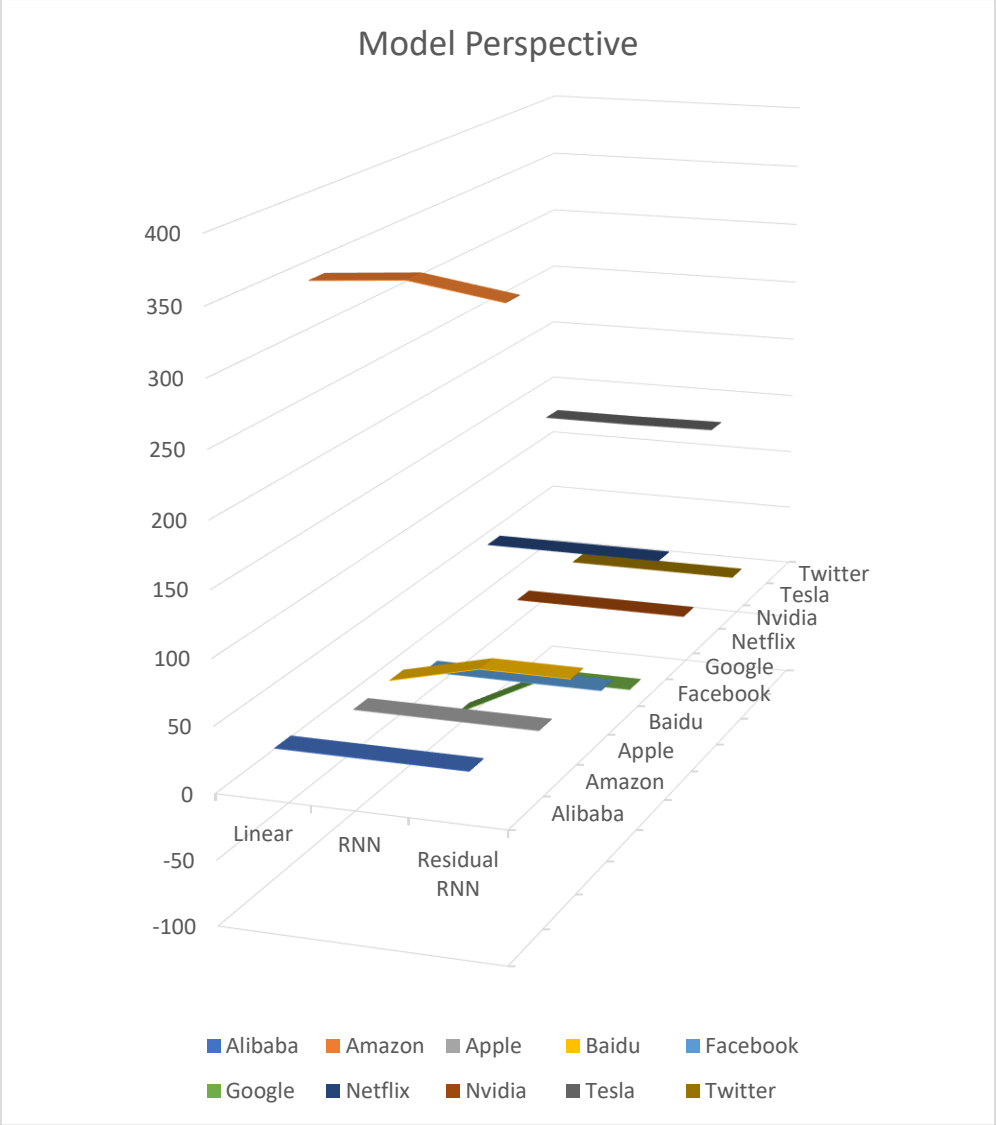

Figure 12: single step model profit (%) (model perspective) (for period studied)

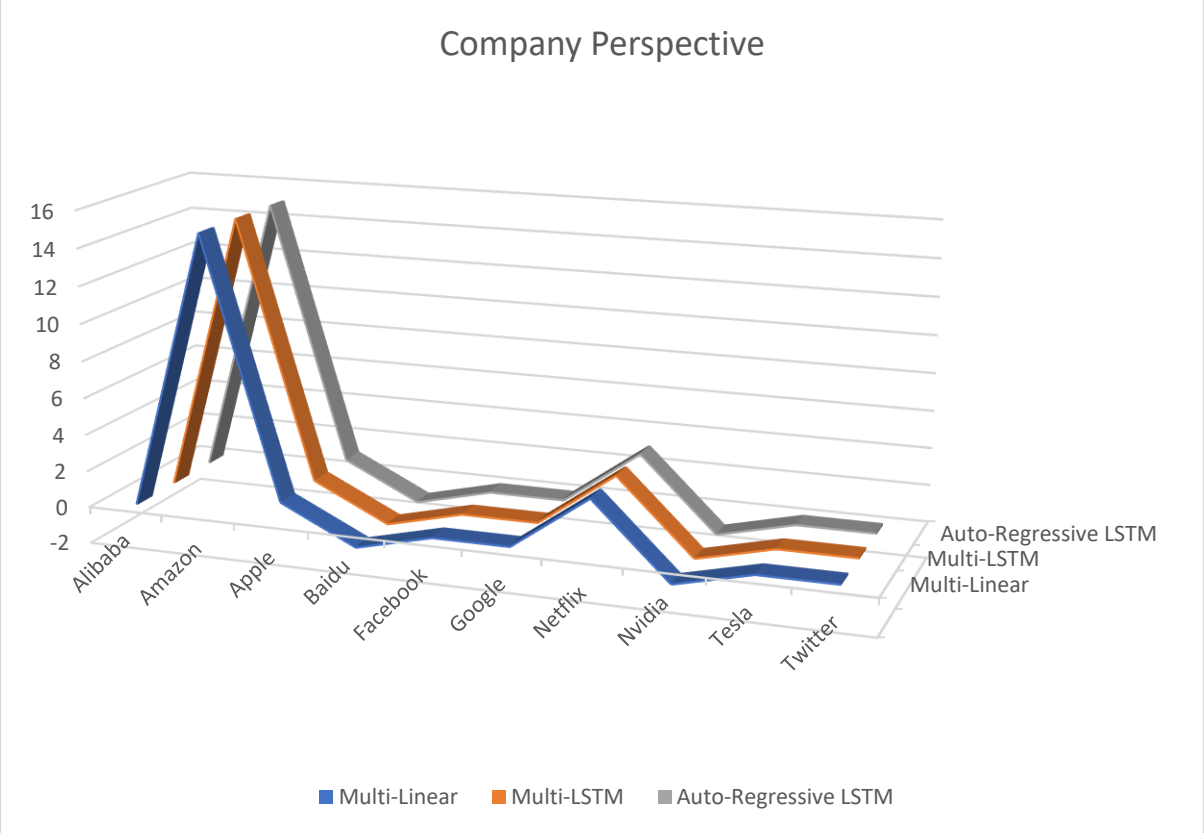

Figure 13: multi-step model profit (%) (company perspective) (for period studied)

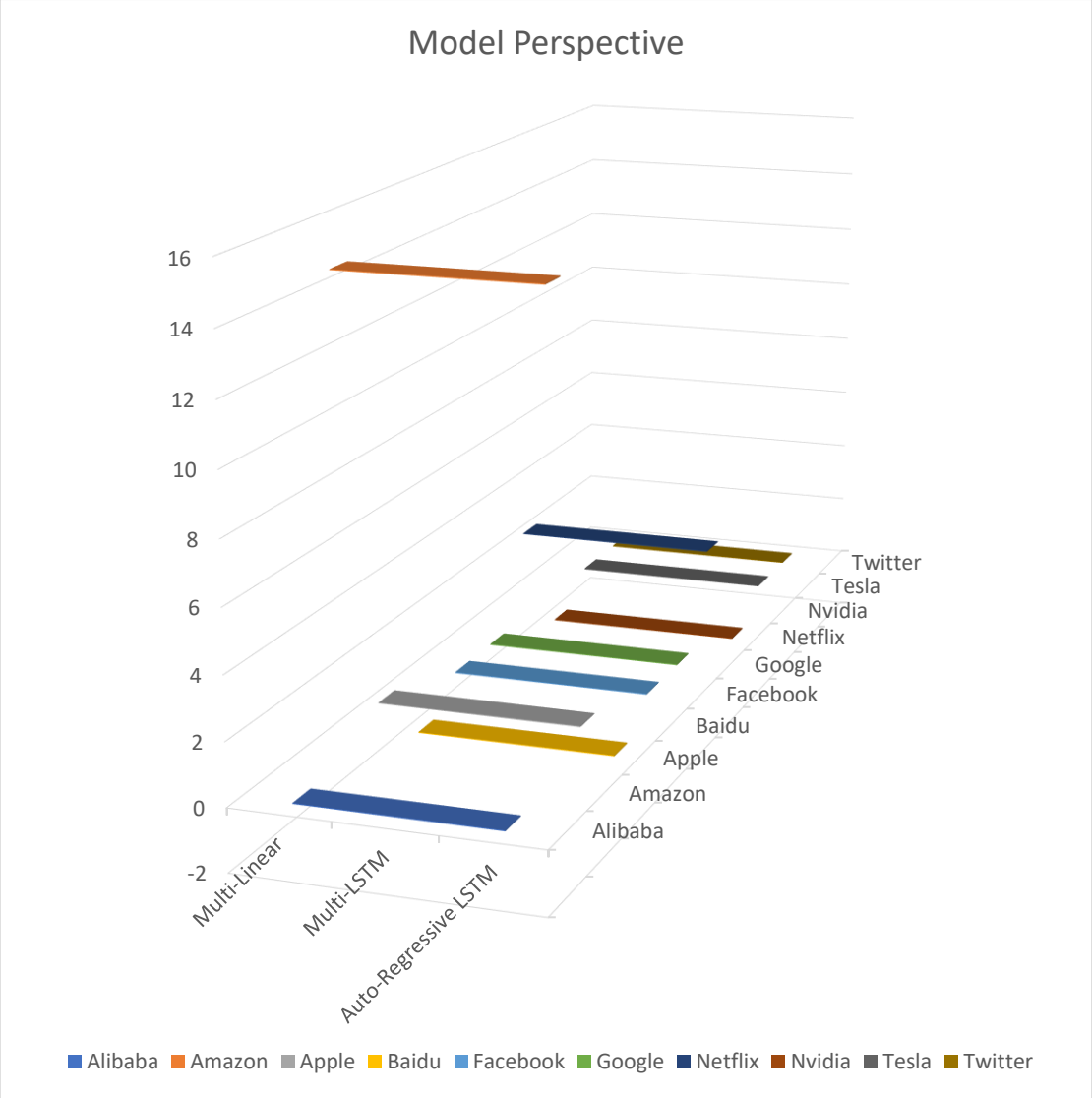

Figure 14: multi-step model profit (%) (model perspective) (for period studied)

From the following figures regardless of the time series-forecasting model chose, may it be single step or multi-step, relatively speaking, the most amount of profit occurred with Amazon and Netflix stocks. However, the profit with the Amazon stocks were far higher than that of Netflix which may indicate that these stocks have a stronger correlation with COVID-19 statistics. Although, after considering the individual stock prices of Netflix and Amazon it becomes clear why the profit was much higher with Amazon stocks compared to Netflix. This is because, during the testing time Amazon stocks were approximately six times higher than that of Netflix stocks, and in the simulations the trading strategy was only allowed to buy one stock in each moment. This meant that a similar amount of each stock of each company was bought during their respective simulation cycles, but the simulation cycle with Amazon had more money invested and thus more money gained in the end.

Table 1: buy and hold FANG+ stocks (profit) during period studied

Company	**Profit (%)**
Alibaba	-25%
Amazon	4%
Apple	19%
Baidu	-22%
Facebook	13%
Google	22%
Netflix	11%
Nvidia	39%
Tesla	14%
Twitter	9%

Furthermore, when comparing the multi-step model-based trading strategy against the single step model trading strategy, a significantly higher margins of profit were noticed. However, the multi-step based-trading strategy provided lower loss compared to the single-step trading strategy. When comparing this trading strategy against a Buy and Hold strategy within the 3-month testing period, the multi-step trading strategy have lower profits and losses compared to Buy and Hold (in Table 1) while single step trading strategy have higher profits and losses compared to Buy and Hold. Thus, indicating that multi-step trading strategy provide a higher stability, while single step trading strategy provide higher profits with more volatility. This is also supported by the higher Annual growth rate values for the single step-based trading strategy compared to the multi-step-based trading strategy.

Table 2: growth rate during period studied

Model	Annual Rate of Return (%)
Linear	228
RNN (LSTM)	251.2
Residual RNN	247.2
Multi-Linear	6.8
Multi-LSTM	6.8
Auto-Regressive LSTM	6.8
MEAN	124.5

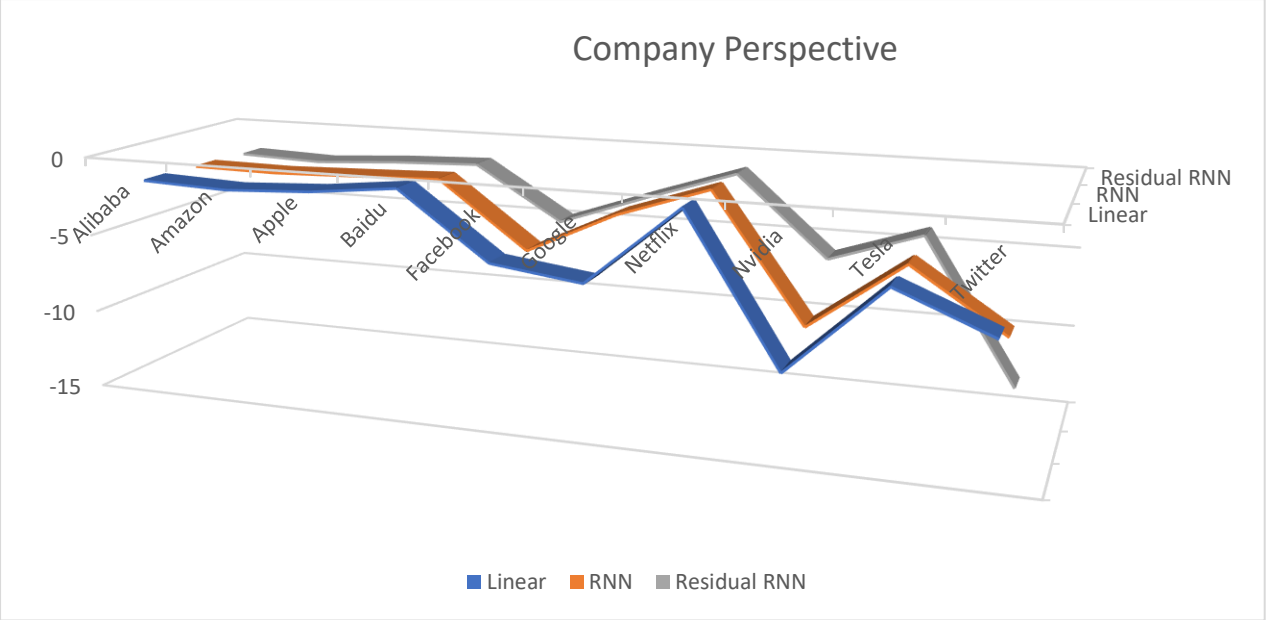

Figure 15: drawdown for single step model (company perspective) (during period studied)

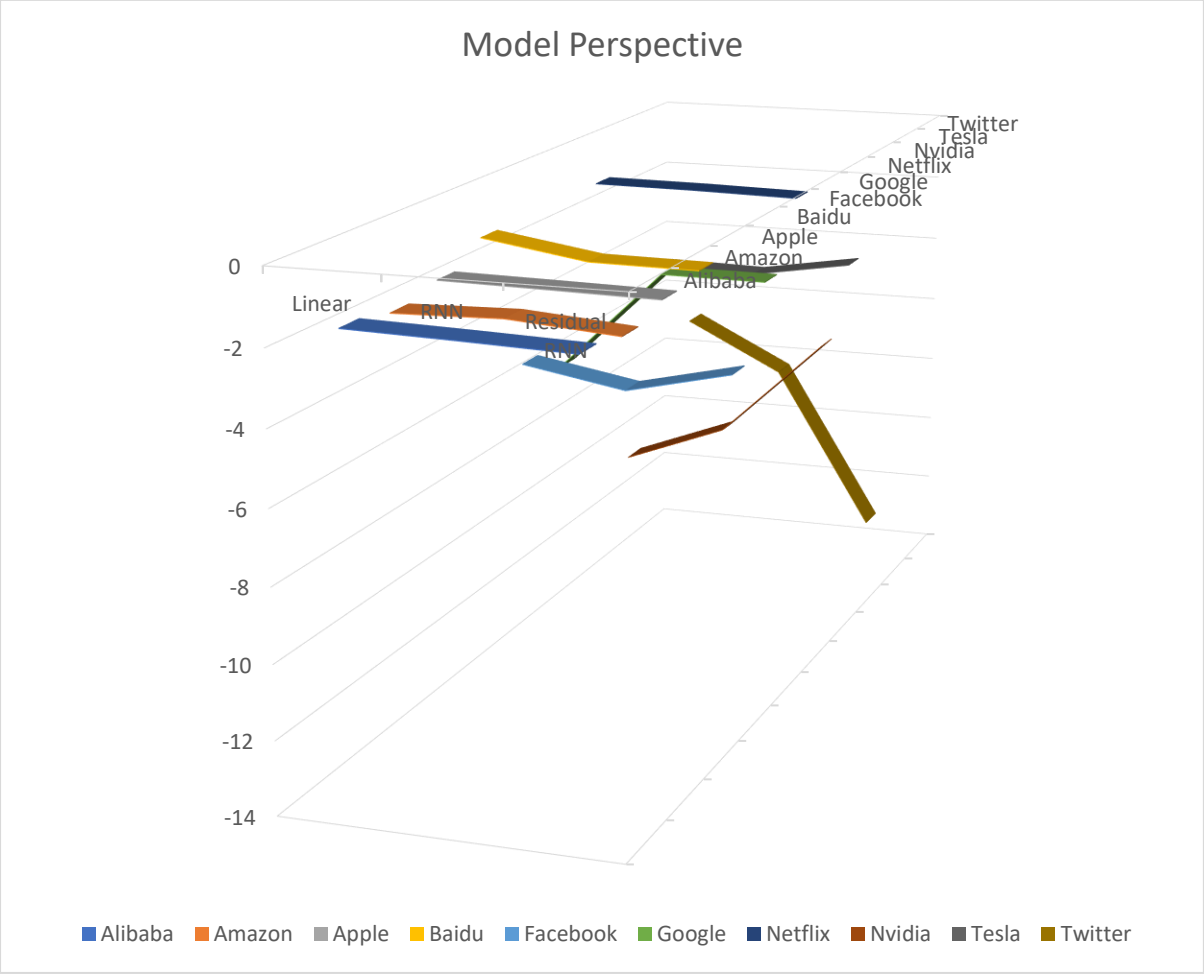

Figure 16: drawdown for single step model (model perspective) (during period studied)

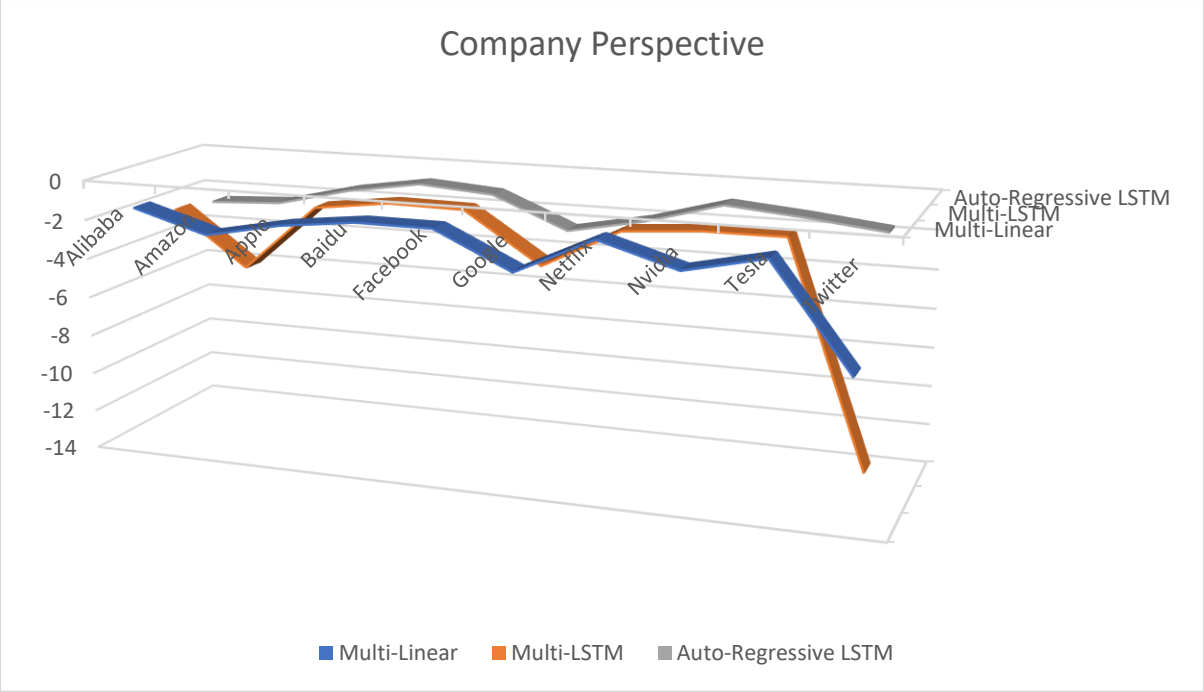

Figure 17: drawdown for multi-step model (company perspective) (during period studied)

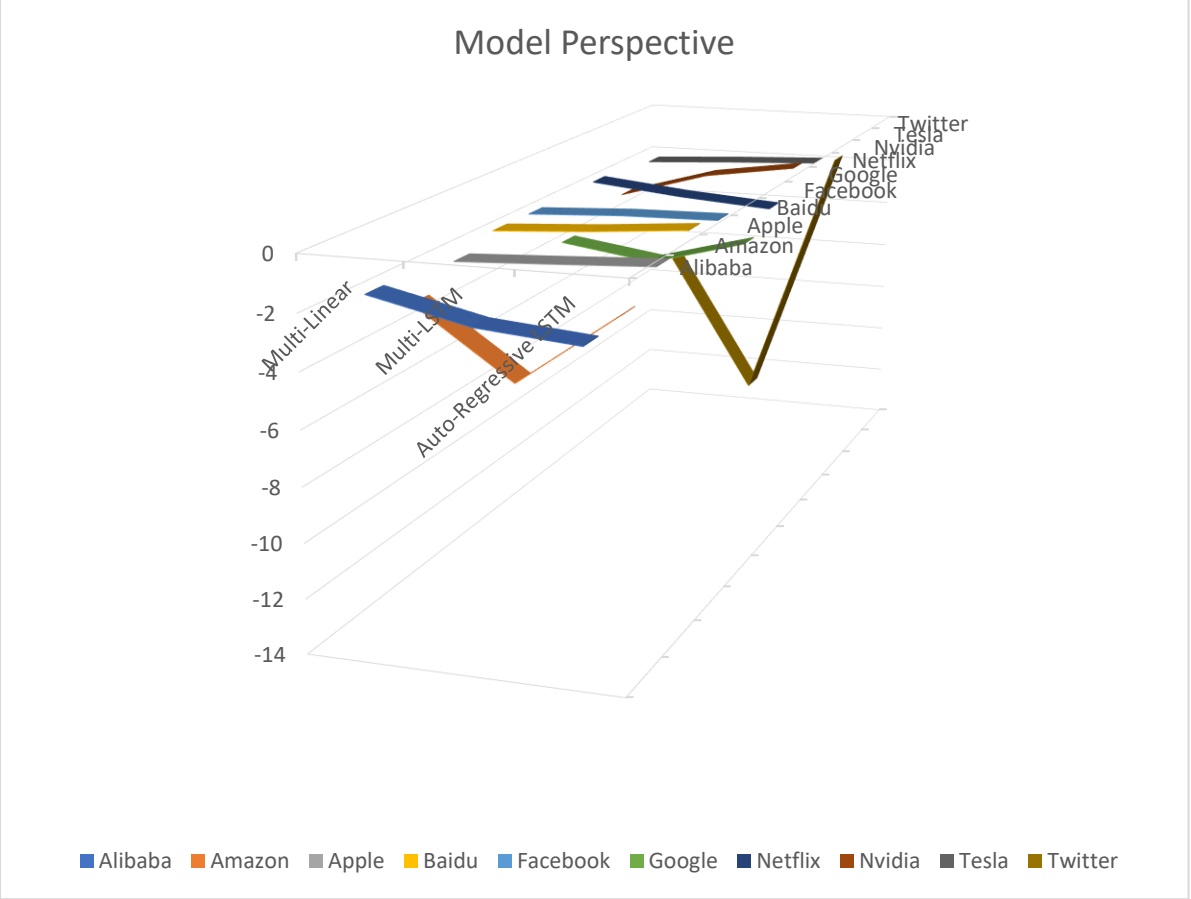

Figure 18: drawdown for multi-step model (model perspective) (during period studied)

Another key factor that needs to be considered when using Trading strategy is the Drawdown within their predictions, which is shown in the figures above. A high Drawdown in predictions would make an investor unlikely to invest within the stock as it is a sign of high volatility. Unlike with the profits, the relative scaling of drawdowns varies drastically across the single step models and multi-step models. However, one stock consistently has one of the highest drawdowns, which is Twitter. This relatively high predicted drawdown for Twitter is also reflected in the actual drawdown of twitter, indicating a possible accuracy in the relative drawdowns of the predictions generated by the models used for the trading strategy. Although, one key difference is that the drawdown predictions for the trading strategies are far lower than the Drawdowns of the actual stocks (in Table 3). This can be indicative of the high number of variables in normal day to day trading which is not fully captured in the simulation.

Table 3: buy and hold FANG+ stocks (drawdown) during period studied

Company	**Drawdown (%)**
Alibaba	-9.52
Amazon	-16.41
Apple	-18.6
Baidu	-59.6
Facebook	-19.71
Google	-17.93
Netflix	-17.29
Nvidia	-15.32
Tesla	-24.65
Twitter	-45.81

5. Conclusion

To summarise, in terms of making a maximal net profit the best strategy is to use the single step prediction trading strategy. However, with this strategy the punishment that arises from failure is high. Thus, for a more stable amount of profit and loss its best to use a multi-step prediction trading strategy.

In addition, for future improvement and extendibility of the project it may be a good idea to consider the new variants of COVID-19 as they may have more interesting effects upon the stock market. Moreover, an alternative to reddit post data

that could be done, with a possible more fruitful outcome, is the semantics behind financial YouTube videos/comments as there is a numerical representation of views and likes that can be used. Furthermore, this project could be expanded to consider several different types of trading rules and to also try to consider stock markets and covid cases across several countries.

However, the future validity of these results is highly uncertain. This is due to the key assumption that the COVID-19 pandemic will play a strong role in our actions, but the everchanging nature of society and this pandemic may make the extendibility of this project implausible.

6. Appendix

Table 4: annual profits of single step models (%) during period studied

	Linear	RNN	Residual RNN
Alibaba	31	31	31
Amazon	353	358	347
Apple	15	15	15
Baidu	16	33	33
Facebook	2	2	2
Google	-55	-19	-19
Netflix	68	68	68
Nvidia	1	1	1
Tesla	143	143	144
Twitter	-4	-4	-4
Mean per Model	57	62.8	61.8

Table 5: profits of multi-step models (%) during period studied

Profits	Multi-Linear	Multi-LSTM	Auto-Regressive LSTM
Alibaba	0	0	0
Amazon	15	15	15
Apple	1	1	1
Baidu	-1	-1	-1
Facebook	0	0	0
Google	0	0	0
Netflix	3	3	3
Nvidia	-1	-1	-1
Tesla	0	0	0
Twitter	0	0	0
Mean per Model	1.7	1.7	1.7

Table 6: drawdown of single step models during period studied

Drawdown	Linear	RNN	Residual RNN
Alibaba	-1.598	-1.67352	-1.76635
Amazon	-1.82158	-1.7453	-1.96803
Apple	-1.55074	-1.59333	-1.64369
Baidu	-0.95838	-1.42036	-1.45367
Facebook	-5.08401	-5.59951	-4.91153
Google	-5.81489	-2.8758	-2.89391
Netflix	-0.90572	-0.9319	-0.99046
Nvidia	-10.0263	-8.87799	-6.02233
Tesla	-4.48235	-4.54501	-4.05657
Twitter	-6.87494	-8.36107	-13.1957
Mean per Model	-3.91169	-3.76238	-3.89022

Table 7: drawdown of multi-step models during period studied

	Multi-Linear	Multi-LSTM	Auto-Regressive LSTM
Alibaba	-1.46787	-2.28767	-2.60405
Amazon	-2.58458	-5.10308	-2.4063
Apple	-1.83244	-1.66595	-1.48671
Baidu	-1.39318	-1.17623	-0.88319
Facebook	-1.40845	-1.23682	-1.17844
Google	-3.23663	-3.77965	-2.79563
Netflix	-1.35823	-1.75207	-2.02931
Nvidia	-2.52618	-1.46842	-0.9363
Tesla	-1.65236	-1.45532	-1.29173
Twitter	-6.82321	-12.6658	-1.74169
Mean per Model	-2.42831	-3.2591	-1.73534

7. Work Cited

[1] VOA, “Controversy Over GameStop’s Stock Market Saga Explained,” VOA News, 30 January 2021. [Online]. Available: https://www.voanews.com/a/economy-business_controversy-over-gamestops-stock-market-saga-explained/6201422.html.

[2] ETF Database, “NYSE FANG+ Index (+300%) – ETF Tracker,” ETF Database, [Online]. Available: https://etfdb.com/index/nyse-fangtm-index/. [Accessed 8 December 2021].

[3] AngelOne, “HOW IS THE ADJUSTED CLOSING PRICE DIFFERENT FROM THE CLOSING PRICE?,” AngelOne, 13 November 2020. [Online]. Available: https://www.angelone.in/knowledge-center/share-market/difference-between-closing-price-and-adjusted-closing-price#:~:text=While%20closing%20price%20merely%20refers,accurate%20measure%20of%20stocks'%20value.

[4] S. Silberstein, “How After-Hours Trading Affects Stock Prices,” Investopedia, 17 October 2021. [Online]. Available: https://www.investopedia.com/ask/answers/05/saleafterhours.asp.

[5] C. Potter, “How to Use a Moving Average to Buy Stocks,” Investopedia, 28 April 2021. [Online]. Available: https://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp.

[6] U. Malapaka, “Using TPUs on Google Colab with Keras,” Towards Data Science, 18 May 2020. [Online]. Available: https://towardsdatascience.com/using-tpus-on-google-colab-966239d24573.

[7] C. Heritage, “The 80-20 Rule,” Nasdaq, 9 Aug 2012. [Online]. Available: https://www.nasdaq.com/articles/80-20-rule-2012-08-09.