STOCK ANALYSIS AND FORECASTING BASED ON TWEETS

Group 09

- Charita Tummala - 16338814
- Prasanna Vetapalem-16333633
- Abhinay Reddy Polimera-16334737
- Hari kiran Reddy Muthyala-16335918
- Sivendra Pammi-16327143
- Divya Sai Gummadi-16330818

OVERVIEW:

- Worked on tweets dataset which are linked with ticker symbols of the companies and performed sentiment analysis to check whether the tweets has an influence on the real-time stock price of the companies.

- The Real-Time stock market dataset of a company and forecasted the future trend for 30 days based on the past stock price using LSTM model.

NumPy

PySpark

matplotlib

pandas

About the Dataset

- We have collected four datasets from Kaggle related to company, company ticker symbols, tweets, and stock prices.
- We have collected tweet data of top companies like Apple, Tesla, Google, Amazon
- Over 1.6 million tweets are collected for over 10 years period of time from 2011 to 2021.

Dataset

Tweets data loaded into DataFrames and displayed

tweet_id	writer	post_date	body	comment_num	retweet_num	like_num
550441509175443456	VisualStockRSRC	1420070457	lx21 made $10,008...	0	0	1
550441672312512512	KeralaGuy77	1420070496	Insanity of today...	0	0	0
550441732014223360	DozenStocks	1420070510	S&P100 #Stocks Pe...	0	0	0
550442977802207232	ShowDreamCar	1420070807	$GM $TSLA: Volksw...	0	0	1
550443807834402816	i_Know_First	1420071005	Swing Trading: Up...	0	0	1
550443808606126081	aaplstocknews	1420071005	Swing Trading: Up...	0	0	1
550443809700851716	iknowfirst	1420071005	Swing Trading: Up...	0	0	1
550443857142611968	Cprediction	1420071016	Swing Trading: Up...	0	0	1
550443857595600896	iknowfirst_br	1420071017	Swing Trading: Up...	0	0	1
550443857692078081	Gold_prediction	1420071017	Swing Trading: Up...	0	0	1
550443858010861568	IKFResearch	1420071017	Swing Trading: Up...	0	0	1
550444112328261632	GetAOM	1420071077	$UNP $ORCL $QCOM ...	0	0	0
550444969924653056	AppleNewsAAPL	1420071282	$AAPL Apple goes ...	0	0	1
550444970738335744	espositooooo	1420071282	“@WSJ: Apple is b...	0	0	0
550445066444369921	Bidnessetc	1420071305	Apple filed for i...	0	0	0
550445850170642432	JorelLaraKalel	1420071492	@CNBC 15 Top #tra...	0	0	2
550447574285418497	btcgemini	1420071903	We searched throu...	0	0	0
550447850857828352	JorelLaraKalel	1420071969	Top 10 searched #...	0	0	2
550447998577426433	UPBOptionMil	1420072004	2014 The Year in ...	0	2	2
550448085789200384	MacHashNews	1420072025	Give your brain a...	0	0	0

We collected the ticker symbols and its associated companies along with the related tweet ids.

ticker_symbol	company_name
AAPL	apple
GOOG	Google Inc
GOOGL	Google Inc
AMZN	Amazon.com
TSLA	Tesla Inc
MSFT	Microsoft

tweet_id	ticker_symbol
550803612197457920	AAPL
550803610825928706	AAPL
550803225113157632	AAPL
550802957370159104	AAPL
550802855129382912	AAPL
550802745737768960	AAPL
550797494188142592	AAPL
550797275786518528	AAPL
550797272686923776	AAPL
550796617444765696	AAPL
550795512899960832	AAPL
550795254102638593	AAPL
550795167318700033	AAPL
550795088821886976	AAPL
550793298357391360	AAPL
550793247669231616	AAPL
550793108242198528	AAPL
550791919815892992	AAPL
550791232738590720	AAPL
550790423888035840	AAPL

PRE-PROCESSING:

- Joining the two dataframes using Tweet_id column.
- Converted the Post_date to required date format.
- Calculated the total_engagement column by adding the comments, retweets and likes.
- Dropped the unwanted columns.

tweet_id	writer	body	comment_num	retweet_num	like_num	ticker_symbol	total_engagements	t_date
692169663577485315	ValaAfshar	Apple has $216 bi...	42	984	677	AAPL	1703.0	2016-01-27
770310550991605760	cnntech	Apple's next iPho...	11	729	918	AAPL	1658.0	2016-08-29
575014851363405824	RANsquawk	Loving my Apple W...	66	882	654	AAPL	1602.0	2015-03-09
816359802733555712	DavidSchawel	Sometimes hard to...	14	646	900	AMZN	1560.0	2017-01-03
854690001866686464	philstockworld	Will We Hold It W...	0	969	520	AMZN	1489.0	2017-04-19
85[illegible]	[illegible]	[illegible] We Hold It W...	0	969	520	TSLA	1489.0	2017-04-19
102[illegible]	[illegible]	[illegible]s - I'm beside...	207	317	899	TSLA	1423.0	2018-07-23
87[illegible]	[illegible]	[illegible]ZN has no stor...	40	509	837	AMZN	1386.0	2017-06-16
86[illegible]	[illegible]	[illegible]ergate Wednesd...	1	971	400	TSLA	1372.0	2017-05-10
61[illegible]	[illegible]	[illegible]d last of our ...	153	671	533	AAPL	1357.0	2015-06-24
1054728662786826240	CitronResearch	$TSLA dropping ea...	148	308	861	TSLA	1317.0	2018-10-23
1018938697415315457	epichedge	Live view of $AMZ...	7	366	927	AMZN	1300.0	2018-07-16
1020077355346169857	vincent13031925	"Tesla Spokespers...	38	256	986	TSLA	1280.0	2018-07-19
1199424478536753155	AlexSibila	~Tesla feature re...	563	48	662	TSLA	1273.0	2019-11-26
1135604016015060993	willchamberlain	FACEBOOK, GOOGLE ...	58	389	826	GOOG	1273.0	2019-06-03
1045404879341137921	Reuters	SEC files lawsuit...	56	630	585	TSLA	1271.0	2018-09-27
1167316598283071495	TeslaNY	.@Tesla Model 3 p...	36	279	952	TSLA	1267.0	2019-08-30
1020036769629143040	Microsoft	$MSFT Q4 EARNINGS...	26	333	896	MSFT	1255.0	2018-07-19
1209424426904801280	YCalenge	Last night, $TSLA...	234	172	835	TSLA	1241.0	2019-12-24
1050135192109760525	charliebilello	% Below 52-week h...	34	415	783	AMZN	1232.0	2018-10-10

Total Engagement: sum of tweet's comments, retweets and likes.

Sentiment Analysis is the process of classifying the text into positive, negative, or, neutral. Sentiment analysis is contextual mining of words which indicates the social sentiment of a brand and also helps the business to determine whether the product which they are manufacturing is going to make a demand in the market or not.

Sentiment analysis tries to gain is to analyze people's opinion in a way that it can help the businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.

we shall now make a function that takes in text and returns the sentiment of that text.

```
[16] def getSentiment(body):

         from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
         analyzer = SentimentIntensityAnalyzer()

         assert body is not None
         vs = analyzer.polarity_scores(body)
         score = vs['compound']

         if (score >= 0.05):
             return "Positive"

         elif (score < 0.05 and score > -0.05):
             return "Neutral"

         elif (score <= -0.05):
             return "Negative"

         print(score)
```

- Vader Sentiment is used to determine sentiment of each post.
- Tweets are classified into three types: Positive, Negative, Neutral.

- We categorized the tweets into positive, negative and neutral.

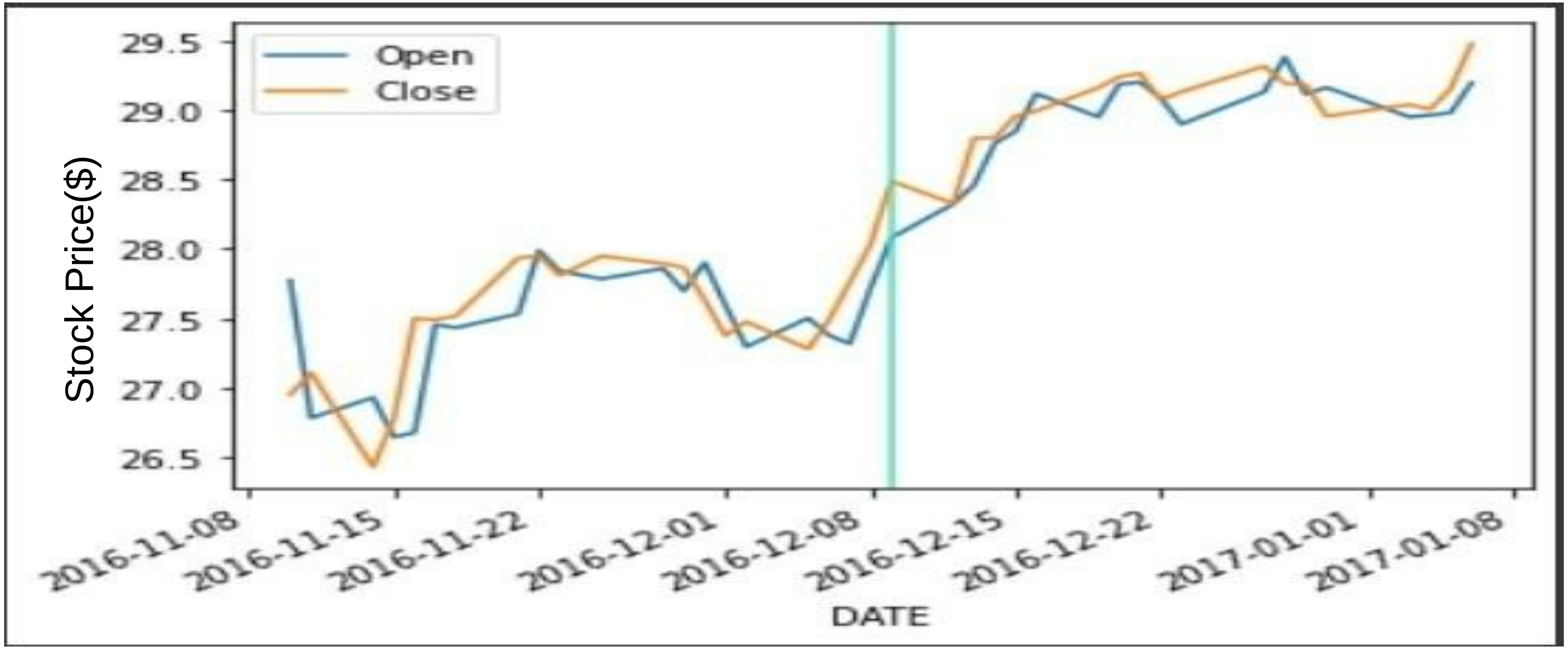
tweet_id	writer	body	comment_num	retweet_num	like_num	ticker_symbol	total_engagements	t_date	Score
3577485315	ValaAfshar	Apple has $216 bi...	42	984	677	AAPL	1703.0	2016-01-27	Neutral
0991605760	cnntech	Apple's next iPho...	11	729	918	AAPL	1658.0	2016-08-29	Neutral
1363405824	RANsquawk	Loving my Apple W...	66	882	654	AAPL	1602.0	2015-03-09	Positive
2733555712	DavidSchawel	Sometimes hard to...	14	646	900	AMZN	1560.0	2017-01-03	Negative
1866686464	philstockworld	Will We Hold It W...	0	969	520	AMZN	1489.0	2017-04-19	Neutral
1866686464	philstockworld	Will We Hold It W...	0	969	520	TSLA	1489.0	2017-04-19	Neutral
8403382272	QTRResearch	Guys - I'm beside...	207	317	899	TSLA	1423.0	2018-07-23	Negative
7003791362	SJosephBurns	$AMZN has no stor...	40	509	837	AMZN	1386.0	2017-06-16	Negative
3010203648	philstockworld	Watergate Wednesd...	1	971	400	TSLA	1372.0	2017-05-10	Neutral
7219076096	Carl_C_Icahn	Sold last of our ...	153	671	533	AAPL	1357.0	2015-06-24	Positive
2786826240	CitronResearch	$TSLA dropping ea...	148	308	861	TSLA	1317.0	2018-10-23	Negative
7415315457	epichedge	Live view of $AMZ...	7	366	927	AMZN	1300.0	2018-07-16	Neutral
5346169857	vincent13031925	"Tesla Spokespers...	38	256	986	TSLA	1280.0	2018-07-19	Negative
6015060993	willchamberlain	FACEBOOK, GOOGLE ...	58	389	826	GOOG	1273.0	2019-06-03	Negative
8536753155	AlexSibila	~Tesla feature re...	563	48	662	TSLA	1273.0	2019-11-26	Positive
9341137921	Reuters	SEC files lawsuit...	56	630	585	TSLA	1271.0	2018-09-27	Negative
8283071495	TeslaNY	.@Tesla Model 3 p...	36	279	952	TSLA	1267.0	2019-08-30	Neutral
9629143040	Microsoft	$MSFT Q4 EARNINGS...	26	333	896	MSFT	1255.0	2018-07-19	Negative
6904801280	YCalenge	Last night, $TSLA...	234	172	835	TSLA	1241.0	2019-12-24	Positive
2109760525	charliebilello	% Below 52-week h...	34	415	783	AAPL	1232.0	2018-10-10	Neutral

Real-Time Stock Market Data

ticker_symbol	Date	Open	High	Low	Close	Adj Close	Volume
AAPL	11/4/2022	142.089996	142.669998	134.380005	138.380005	138.380005	140716700
AAPL	11/3/2022	142.059998	142.800003	138.75	138.880005	138.650009	97918500
AAPL	11/2/2022	148.949997	152.169998	145	145.029999	144.78981	93604600
AAPL	11/1/2022	155.080002	155.449997	149.130005	150.649994	150.400497	80379300
AAPL	10/31/2022	153.160004	154.240005	151.919998	153.339996	153.086044	97943200
AAPL	10/28/2022	148.199997	157.5	147.820007	155.740005	155.482086	164762400
AAPL	10/27/2022	148.070007	149.050003	144.130005	144.800003	144.560196	109180200
AAPL	10/26/2022	150.960007	151.990005	148.039993	149.350006	149.102661	88194300
AAPL	10/25/2022	150.089996	152.490005	149.360001	152.339996	152.087708	74732300
AAPL	10/24/2022	147.190002	150.229996	146	149.449997	149.202484	75981900
AAPL	10/21/2022	142.869995	147.850006	142.649994	147.270004	147.026108	86464700
AAPL	10/20/2022	143.020004	145.889999	142.649994	143.389999	143.152527	64522000
AAPL	10/19/2022	141.690002	144.949997	141.5	143.860001	143.62175	61758300
AAPL	10/18/2022	145.490005	146.699997	140.610001	143.75	143.511932	99136600
AAPL	10/17/2022	141.070007	142.899994	140.270004	142.410004	142.174164	85250900
AAPL	10/14/2022	144.309998	144.520004	138.190002	138.380005	138.150833	88512300
AAPL	10/13/2022	134.990005	143.589996	134.369995	142.990005	142.753204	113224000
AAPL	10/12/2022	139.130005	140.360001	138.160004	138.339996	138.110886	70433700
AAPL	10/11/2022	139.899994	141.350006	138.220001	138.979996	138.749832	77033700
AAPL	10/10/2022	140.419998	141.889999	138.570007	140.419998	140.187439	74899000

Graphing the trend

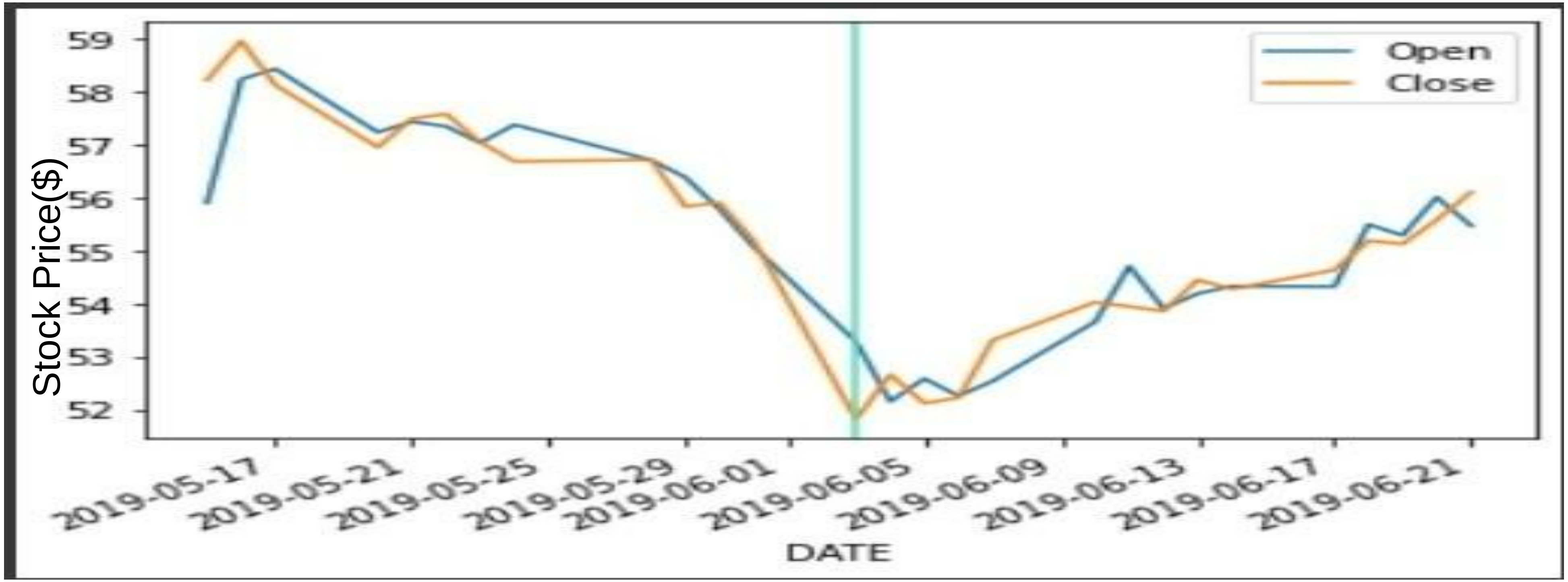
```
[ ]  df3_f.filter(df3_f.t_date == '2016-12-09').filter(df3_f.ticker_symbol == 'AAPL').show()

+------------------+-------------+------------------+-----------+-----------+--------+-------------+------------------+----------+--------+
|          tweet_id|       writer|              body|comment_num|retweet_num|like_num|ticker_symbol|total_engagements|    t_date|   Score|
+------------------+-------------+------------------+-----------+-----------+--------+-------------+------------------+----------+--------+
|807233825659097088|philstockworld|Thank Trump It's ...|          1|        911|      33|         AAPL|             945.0|2016-12-09|Positive|
+------------------+-------------+------------------+-----------+-----------+--------+-------------+------------------+----------+--------+
```



Graphing the trend

```
[ ]  df3_f.filter(df3_f.t_date == '2019-06-03').filter(df3_f.ticker_symbol == 'GOOG').show()
```

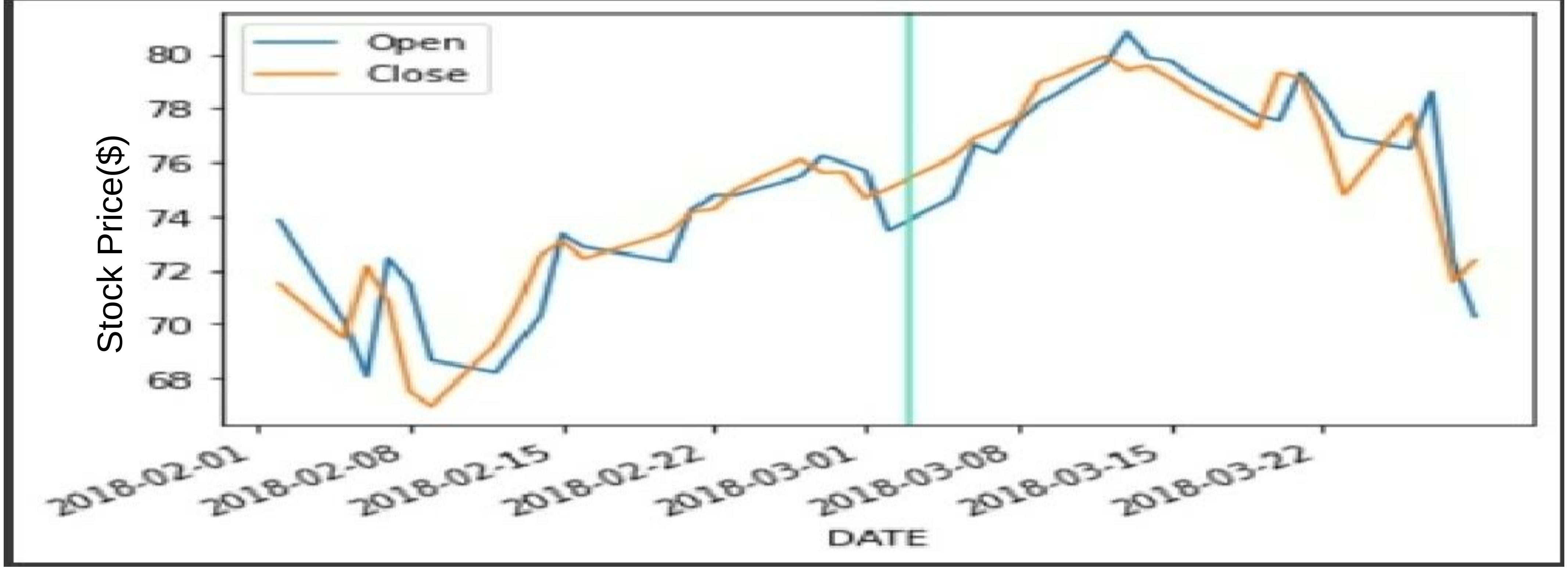
tweet_id	writer	body	comment_num	retweet_num	like_num	ticker_symbol	total_engagements	t_date	Score
1135604016015060993	willchamberlain	FACEBOOK, GOOGLE ...	58	389	826	GOOG	1273.0	2019-06-03	Negative
1135658420785336320	matthewstoller	The House Judicia...	8	38	93	GOOG	139.0	2019-06-03	Negative
1135638699754938368	GretaLWall	#MarketWrap Major...	2	16	47	GOOG	65.0	2019-06-03	Negative

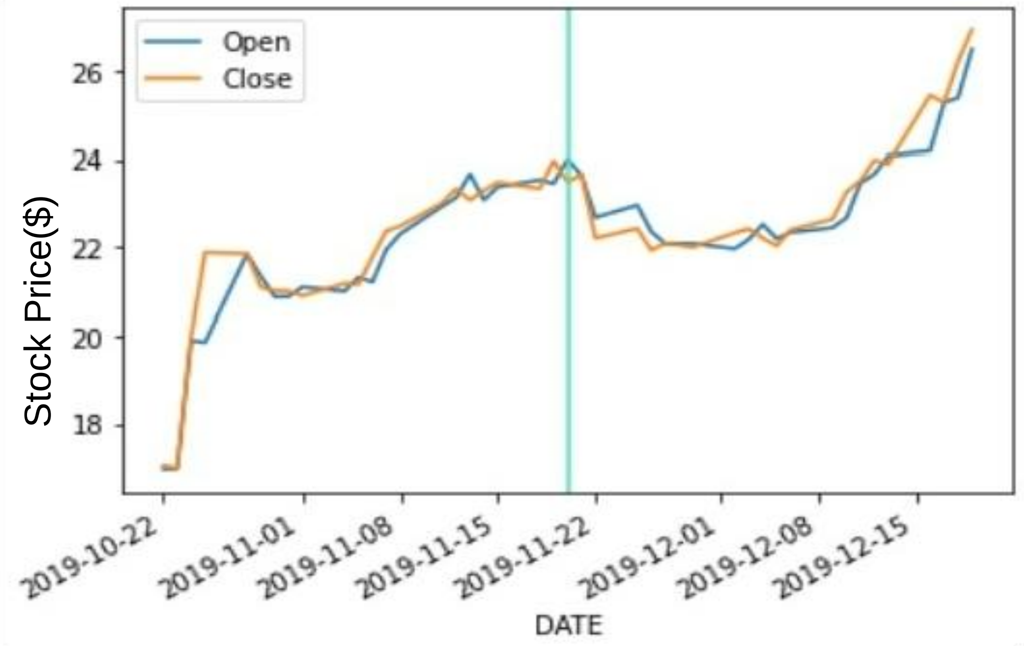


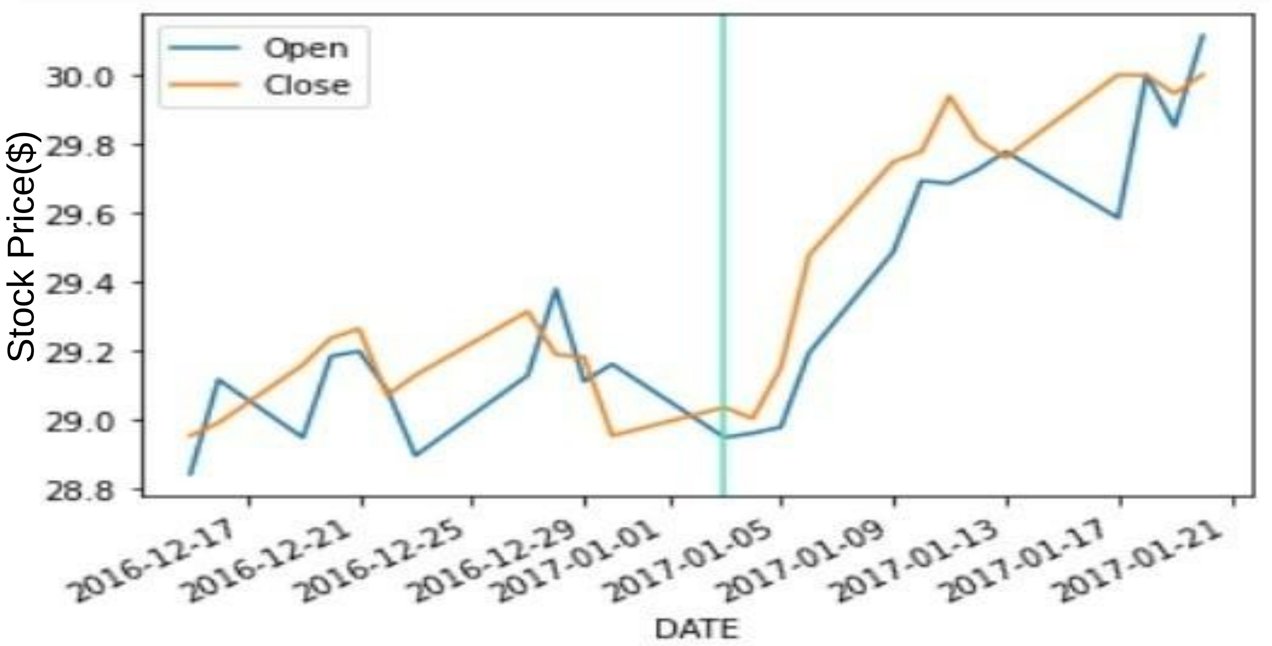
Graphing the trend

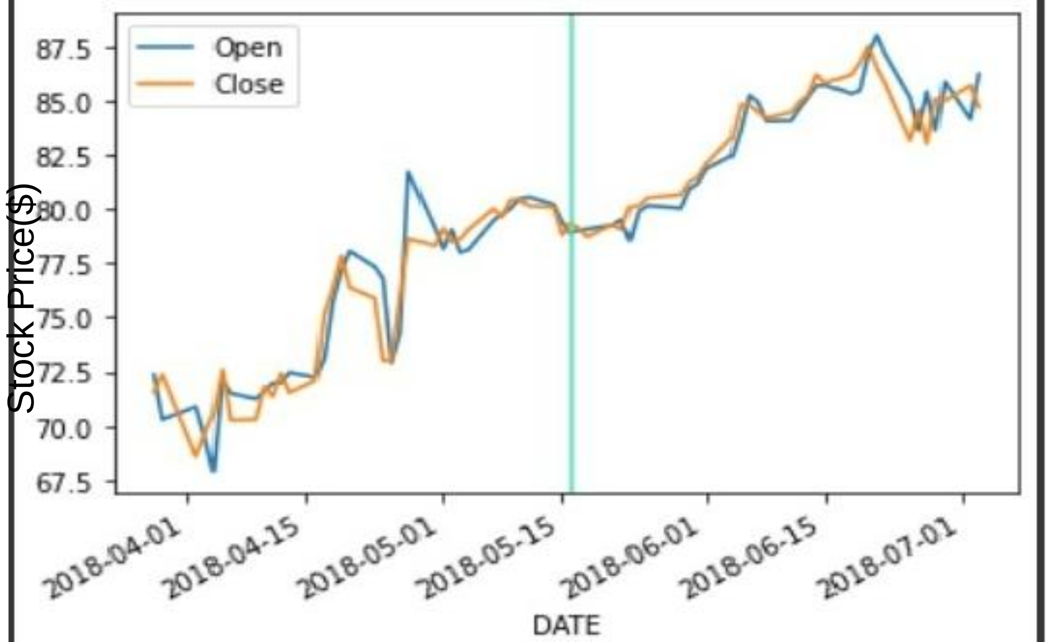
```
[ ]  df3_f.filter(df3_f.t_date == '2018-03-03').filter(df3_f.ticker_symbol == 'AMZN').show()
```

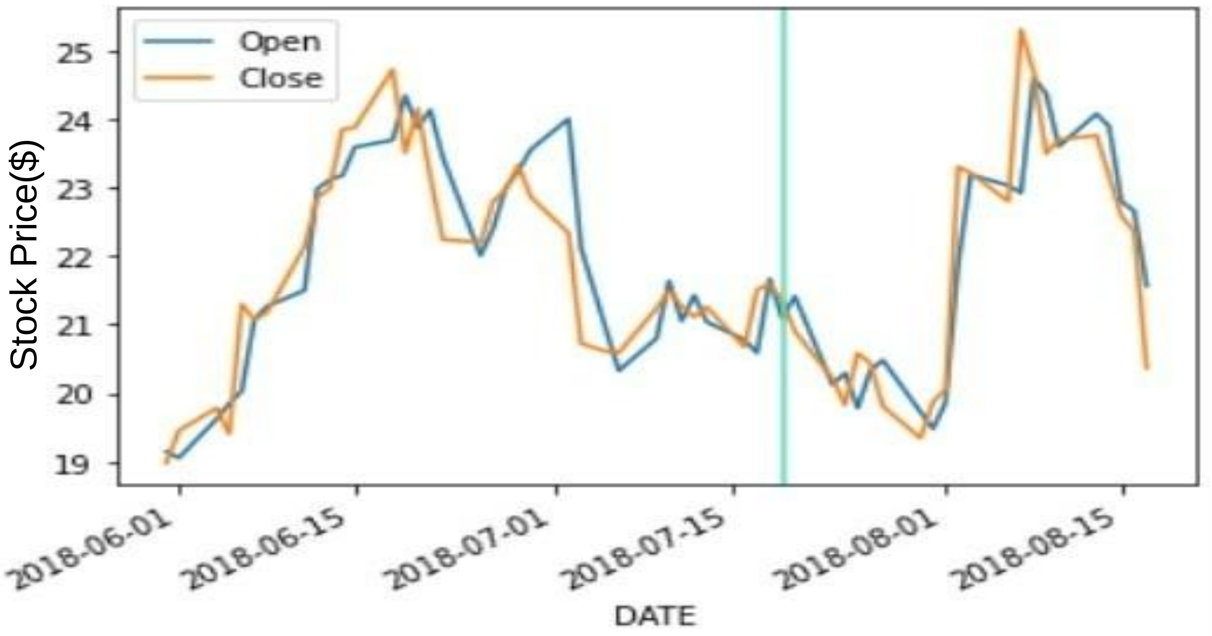
tweet_id	writer	body	comment_num	retweet_num	like_num	ticker_symbol	total_engagements	t_date	Score
969982565787545601	victoriouscake	Mutual funds are ...	2	8	61	AMZN	71.0	2018-03-03	Positive




Open
Close
Stock Price($)
DATE


Open
Close
Stock Price($)
DATE


Open
Close
Stock Price($)
DATE


Open
Close
Stock Price($)
DATE

AAPL STOCK DATA

	Date	Low	Open	Volume	High	Close	Adjusted Close
0	12-12-1980	0.128348	0.128348	469033600.0	0.128906	0.128348	0.100922
1	15-12-1980	0.121652	0.122210	175884800.0	0.122210	0.121652	0.095657
2	16-12-1980	0.112723	0.113281	105728000.0	0.113281	0.112723	0.088636
3	17-12-1980	0.115513	0.115513	86441600.0	0.116071	0.115513	0.090830
4	18-12-1980	0.118862	0.118862	73449600.0	0.119420	0.118862	0.093463
...	...	...	...	...	...	...	...
10167	12-04-2021	130.630005	132.520004	91420000.0	132.850006	131.240005	131.240005
10168	13-04-2021	131.929993	132.440002	91266500.0	134.660004	134.429993	134.429993
10169	14-04-2021	131.660004	134.940002	87222800.0	135.000000	132.029999	132.029999
10170	15-04-2021	133.639999	133.820007	89347100.0	135.000000	134.500000	134.500000
10171	16-04-2021	133.279999	134.300003	84818500.0	134.669998	134.160004	134.160004

10172 rows × 7 columns

- Long Term Short Memory, says that it not only processes single data points but also entire sequences of data(such as speech, images and video).
- Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning which is proposed in 1997.
- Unlike standard feed-forward neural networks, LSTM has feedback connections.
- “Long Short- Term Memory (LSTM) networks are a modified version of recurrent neural networks, which makes it easier to remember past data in memory”.

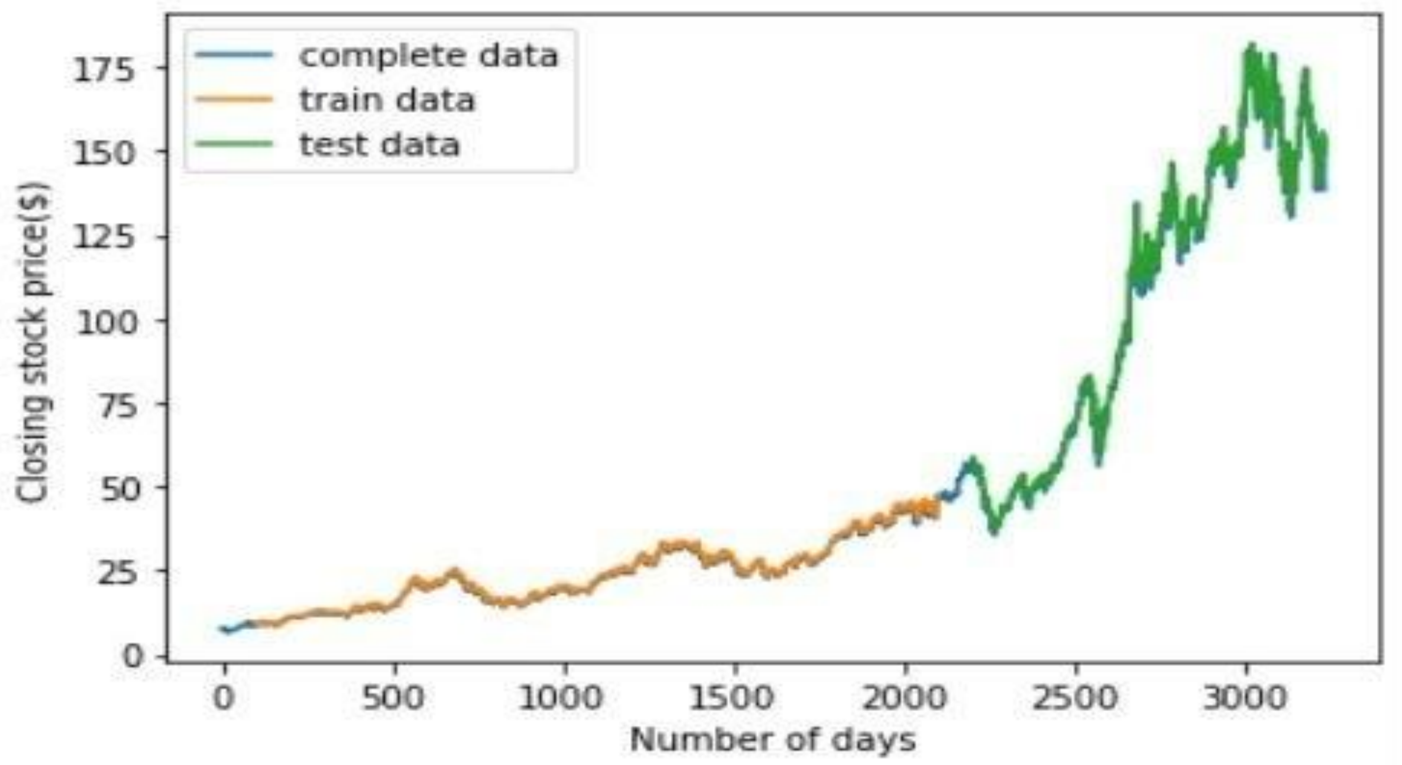
WHY LSTM?

LONG SHORT TERM MEMEORY

LSTM could not process a single data point. it needs a sequence of data for processing and able to store historical information. LSTM is an appropriate algorithm to make prediction and process based-on time-series data.

The stock market has enormously historical data that varies with trade date, which is time-series data, but the LSTM model predicts future price of stock within a short-time period with higher accuracy when the dataset has a huge amount of data.

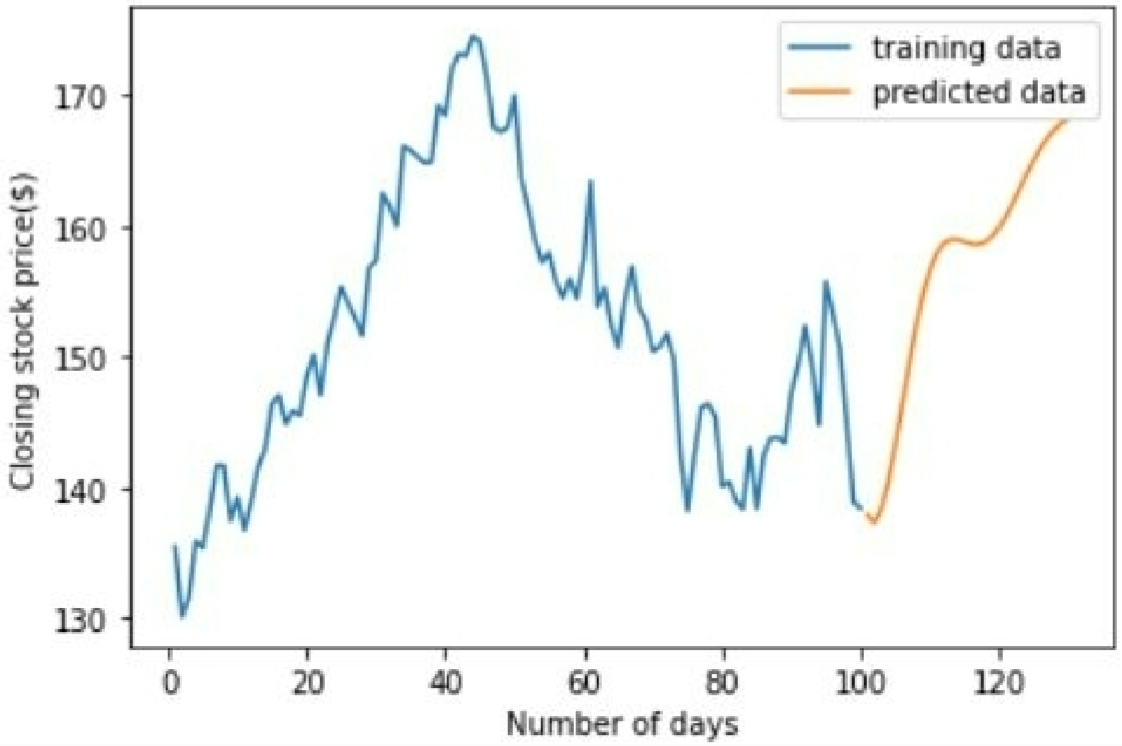
```
[ ] model.summary()

Model: "sequential_5"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_15 (LSTM)               (None, 100, 50)           10400
_________________________________________________________________
lstm_16 (LSTM)               (None, 100, 50)           20200
_________________________________________________________________
lstm_17 (LSTM)               (None, 50)                20200
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 51
=================================================================
Total params: 50,851
Trainable params: 50,851
Non-trainable params: 0
_________________________________________________________________
```



We are using a sequential model and adding the layers of the LSTM

- Green indicates the Test Data
- Blue indicates the Complete Data
- Orange indicates the Train Data

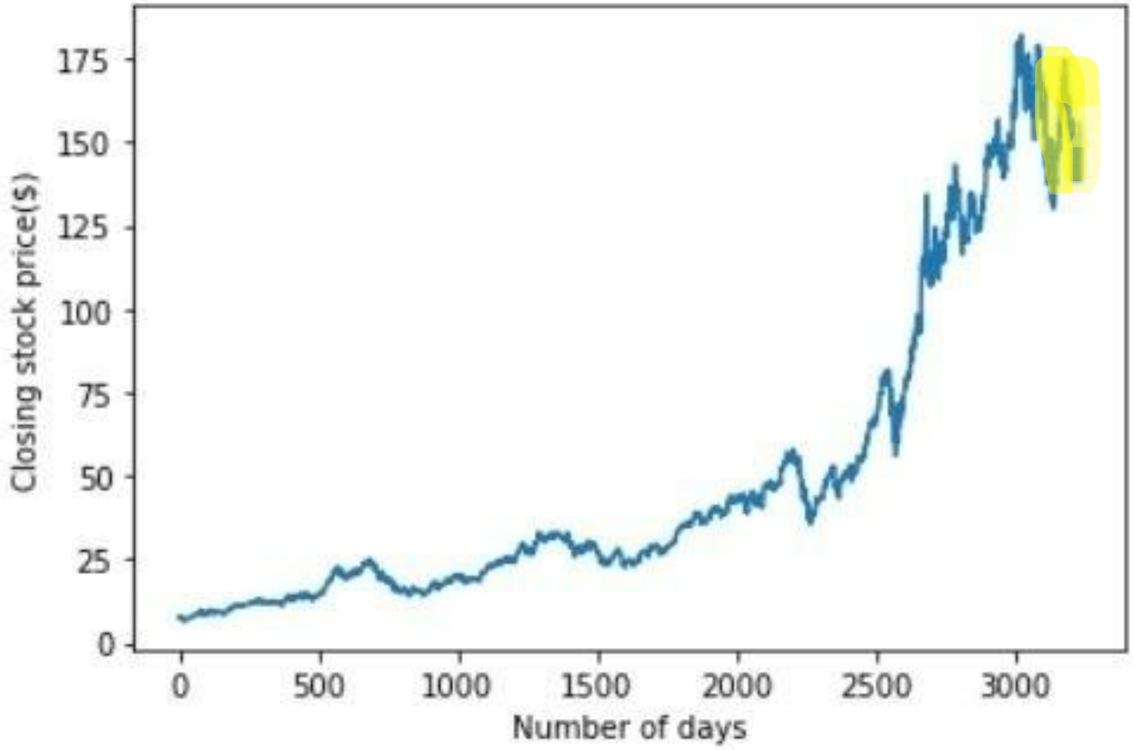
PREDICTION FOR THE NEXT 30 DAYS



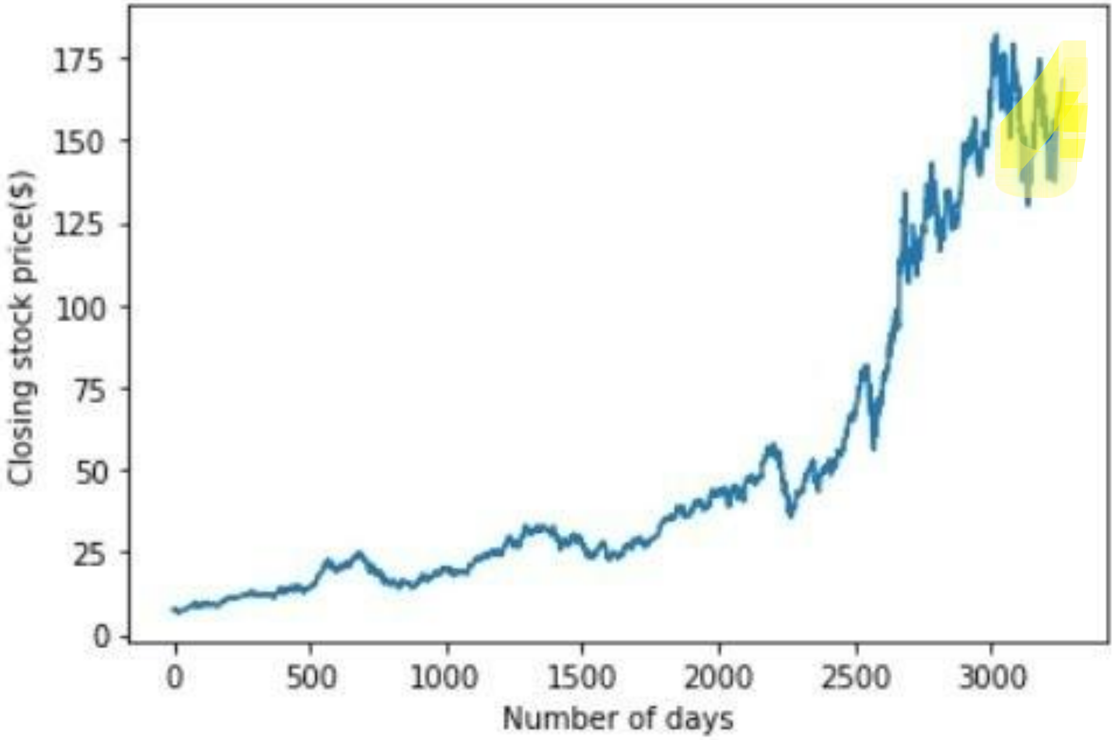
The orange color line in the plot represents the value of stock price for the next 30 days.

Before and After Modelling

Original graph



Extended graph



CONCLUSION:

SENTIMENT ANALYSIS:

- Out of 8 random tweets only 4 (50%) influenced the stock prices when compared with the opening and closing values.
- So, in conclusion, **Yes, twitter does have an effect on the stock market,** but it is not necessary that if your tweet does go viral, based on emotion behind tweet it will influence the stock prices of company.

LSTM MODELLING TIME FORECASTING:

- While the exact price points from our predicted price weren't always close to the actual price, our model did still indicate overall trends such as going up or down. This project teaches us how the LSTMs can be effective in times series forecasting.

Thank You