Supervised Machine Learning: Application to synthesis of Market News Narratives & Sentiments to Predict Future Market Movements

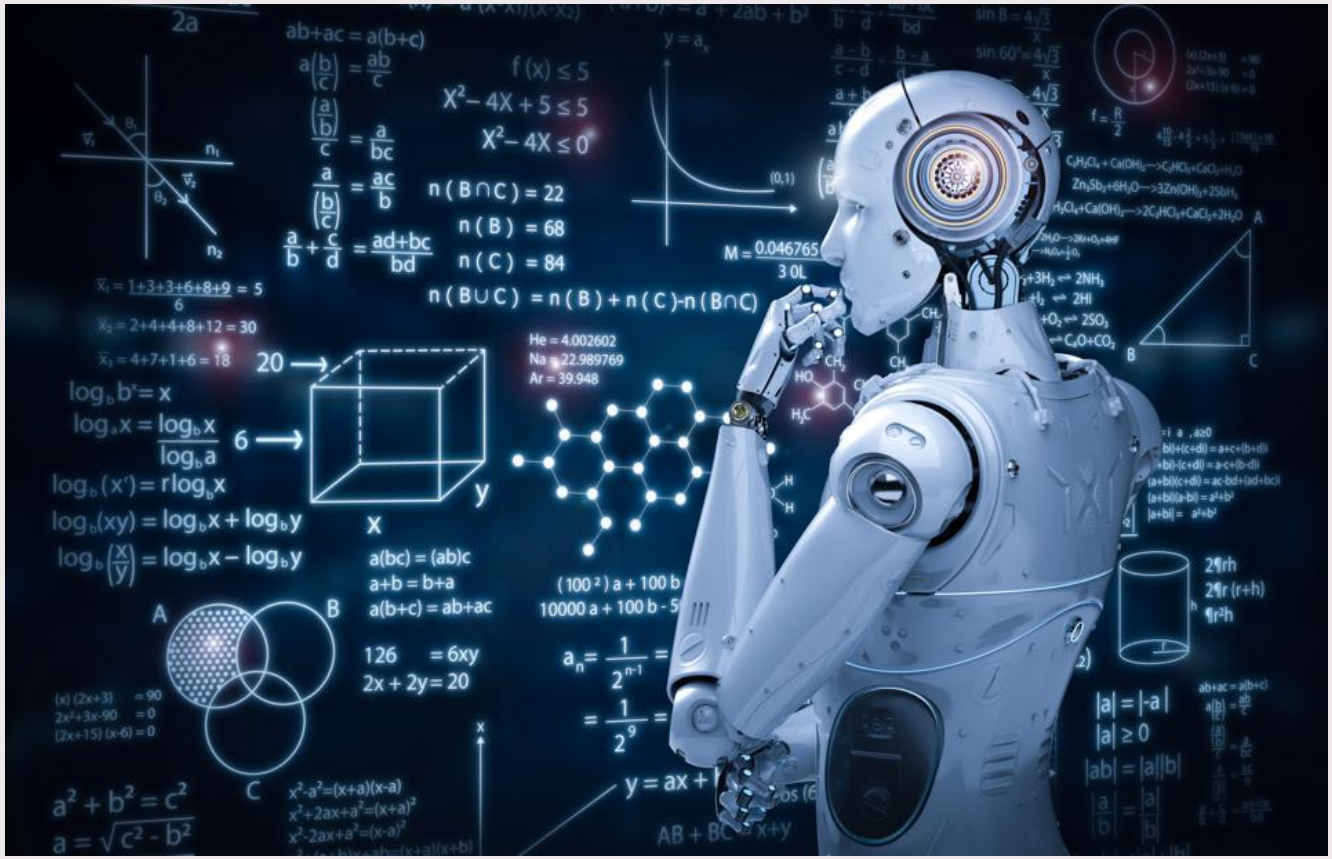
Samuel M Nuugulu

Department of Computing, Mathematical & Statistical Sciences
University of Namibia

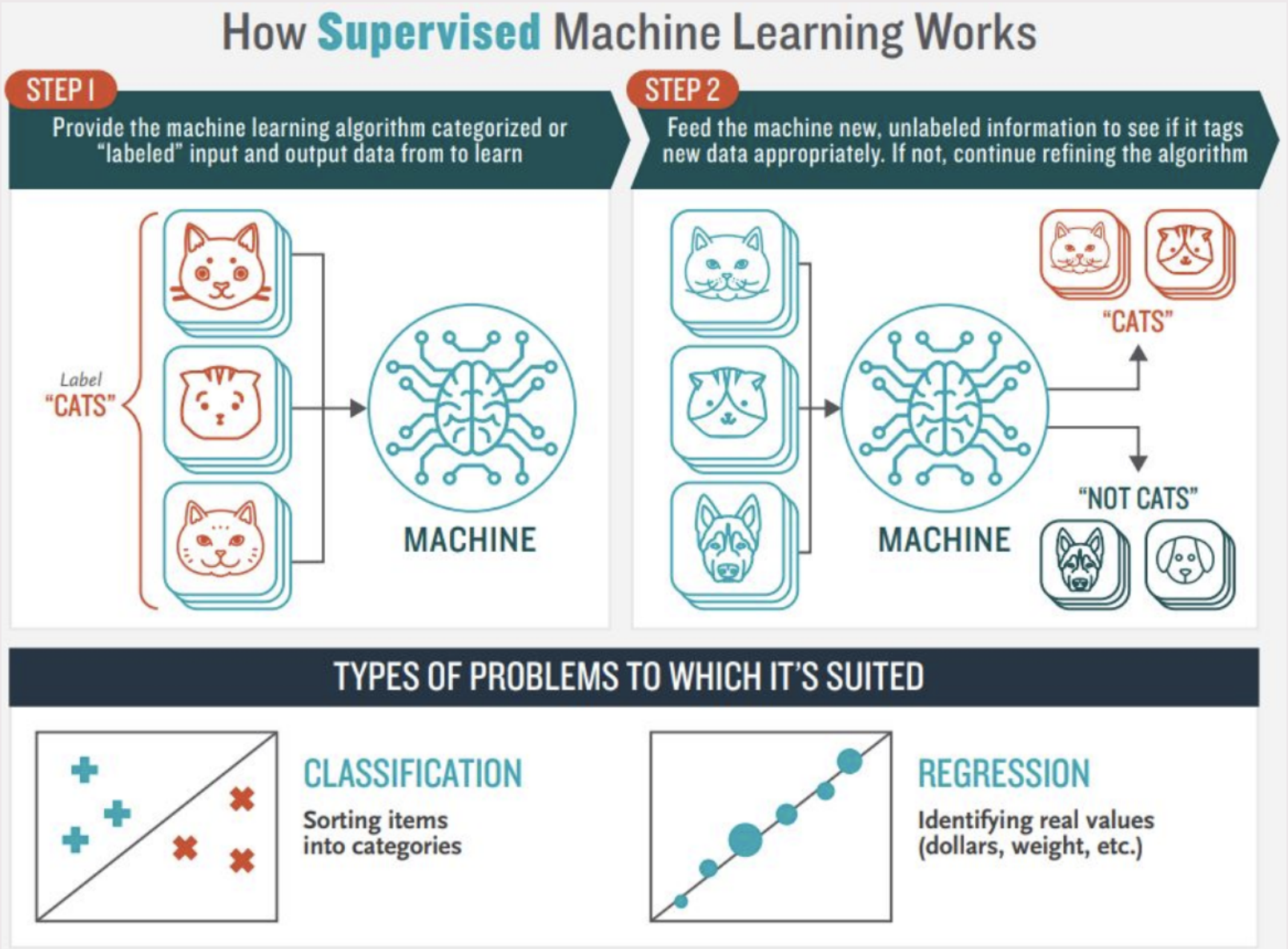
October 26, 2022

What is Machine Learning?

Use of data and algorithms to imitate the way that humans learn

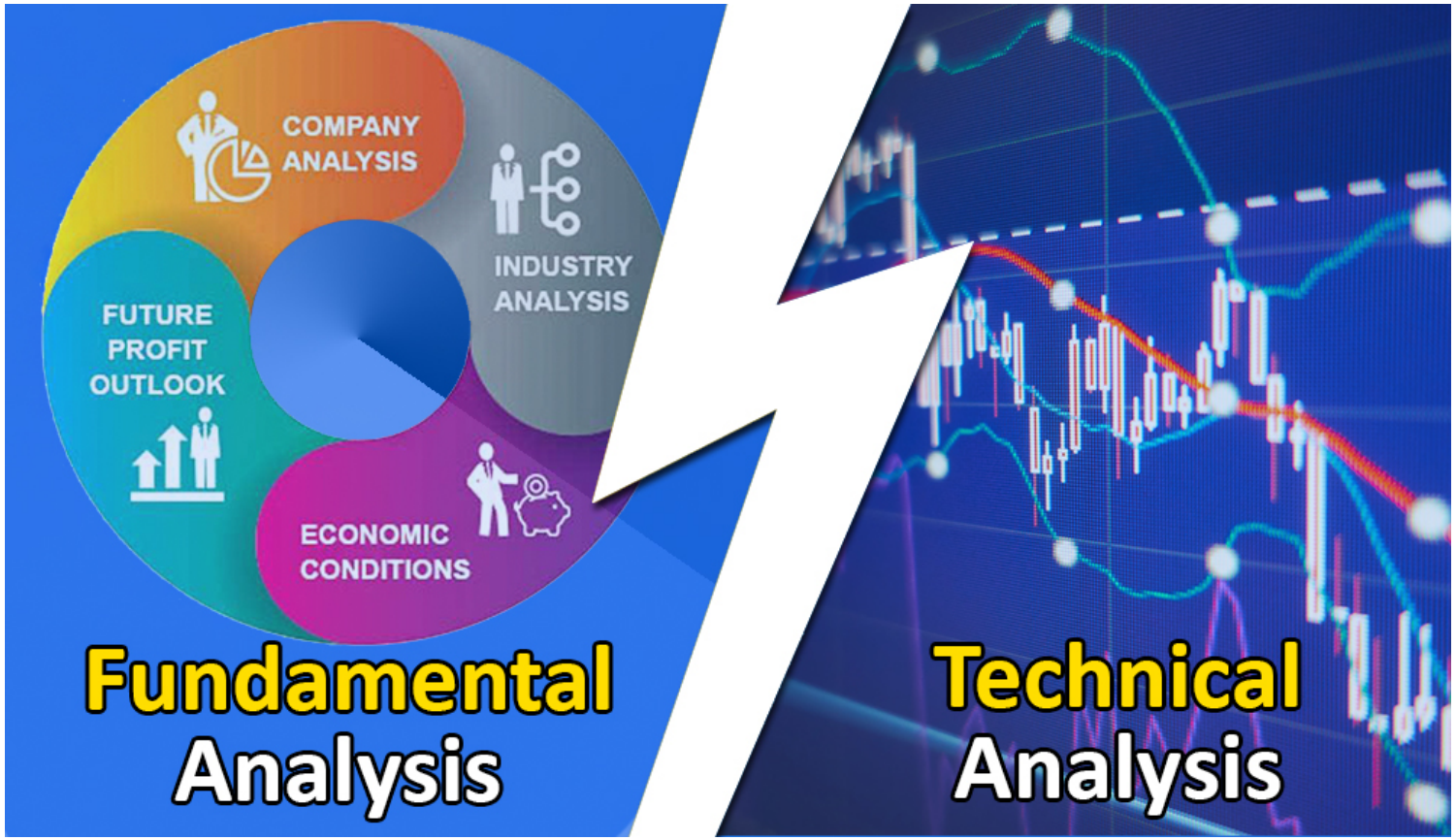


What is Supervised Machine Learning?



What drive the stock market?

Stock Markets are driven by many factors: **fundamentals** (*news*) or **technicals** (*supply and demand*).



Advent of financial technology $\Longrightarrow$ **high volume of trading data**

Challenge

Almost impossible to make a rational investment/trading decisions at face value.

Solution

Machine Learning?

Proposed technique

Aggregating **stock price data**, **stock news sentiments** and **narratives** on a particular stock $\Longrightarrow$ a futuristic **sentiment score**.

Data: Stock News Data

Collect news articles on Tesla, Facebook and Twitter using yahoo finance API for python. We will use the VADER Natural Language Processing Technique (for Text Mining).

Sample news articles on Tesla


```
<tr><td align="right" width="130">09:26AM</td><td align="left"><div class="news-link-container"><div class="news-link-left"><a class="tab-
link-news" href="https://www.barrons.com/articles/cathie-wood-buying-tesla-stock-51666358707?siteid=yhoof2"
onclick="trackAndOpenNews(event, 'Barrons.com', 'https://www.barrons.com/articles/cathie-wood-buying-tesla-stock-
51666358707?siteid=yhoof2');" target="_blank">Cathie Wood Is Buying Tesla Stock. Shes Been Right About It This Year.</a></div><div
class="news-link-right"><span style="color:#aa6dc0;font-size:9px"> Barrons.com</span></div></div></td></tr>
<tr><td align="right" width="130">09:07AM</td><td align="left"><div class="news-link-container"><div class="news-link-left"><a class="tab-
link-news" href="https://finance.yahoo.com/news/exxonmobil-hasbro-highlighted-zacks-bull-130701284.html" onclick="trackAndOpenNews(event,
'Zacks', 'https://finance.yahoo.com/news/exxonmobil-hasbro-highlighted-zacks-bull-130701284.html');" target="_blank">ExxonMobil and Hasbro
have been highlighted as Zacks Bull and Bear of the Day</a></div><div class="news-link-right"><span style="color:#aa6dc0;font-size:9px">
Zacks</span></div></div></td></tr>
<tr><td align="right" width="130">08:42AM</td><td align="left"><div class="news-link-container"><div class="news-link-left"><a class="tab-
link-news" href="https://www.marketwatch.com/story/weekend-reads-how-the-strong-dollar-can-affect-your-financial-health-
11666356159?siteid=yhoof2" onclick="trackAndOpenNews(event, 'MarketWatch', 'https://www.marketwatch.com/story/weekend-reads-how-the-
strong-dollar-can-affect-your-financial-health-11666356159?siteid=yhoof2');" target="_blank">How the strong dollar can affect your
financial health</a></div><div class="news-link-right"><span style="color:#aa6dc0;font-size:9px"> MarketWatch</span></div></div></td></tr>
<tr><td align="right" width="130">08:20AM</td><td align="left"><div class="news-link-container"><div class="news-link-left"><a class="tab-
link-news" href="https://www.investors.com/market-trend/stock-market-today/dow-jones-futures-fall-as-yields-keep-rising-twitter-tumbles-
on-latest-elon-musk-twist/?src=A00220" onclick="trackAndOpenNews(event, 'Investor\u0027s Business Daily', 'https://www.investors.com
/market-trend/stock-market-today/dow-jones-futures-fall-as-yields-keep-rising-twitter-tumbles-on-latest-elon-musk-twist/?src=A00220');"
target="_blank">Dow Jones Futures Fall As Yields Keep Rising; Twitter Skids On Latest Elon Musk Twist</a></div><div class="news-link-
right"><span style="color:#aa6dc0;font-size:9px"> Investor's Business Daily</span></div></div></td></tr>
<tr><td align="right" width="130">08:08AM</td><td align="left"><div class="news-link-container"><div class="news-link-left"><a class="tab-
link-news" href="https://www.barrons.com/articles/tesla-apple-stock-price-earnings-dollar-51666280976?siteid=yhoof2"
onclick="trackAndOpenNews(event, 'Barrons.com', 'https://www.barrons.com/articles/tesla-apple-stock-price-earnings-dollar-
51666280976?siteid=yhoof2');" target="_blank">The Bad News for Apple in Teslas Earnings</a></div><div class="news-link-right"><span
style="color:#aa6dc0;font-size:9px"> Barrons.com</span></div></div></td></tr>
<tr><td align="right" width="130">07:58AM</td><td align="left"><div class="news-link-container"><div class="news-link-left"><a class="tab-
link-news" href="https://finance.yahoo.com/news/elon-musk-says-already-recession-115826344.html" onclick="trackAndOpenNews(event,
'Fortune', 'https://finance.yahoo.com/news/elon-musk-says-already-recession-115826344.html');" target="_blank">Elon Musk says were already
in a recession that could last until Spring 2024 and only the strong will survive</a></div><div class="news-link-right"><span
style="color:#aa6dc0;font-size:9px"> Fortune</span></div></div></td></tr>
```

News Data preprocessing

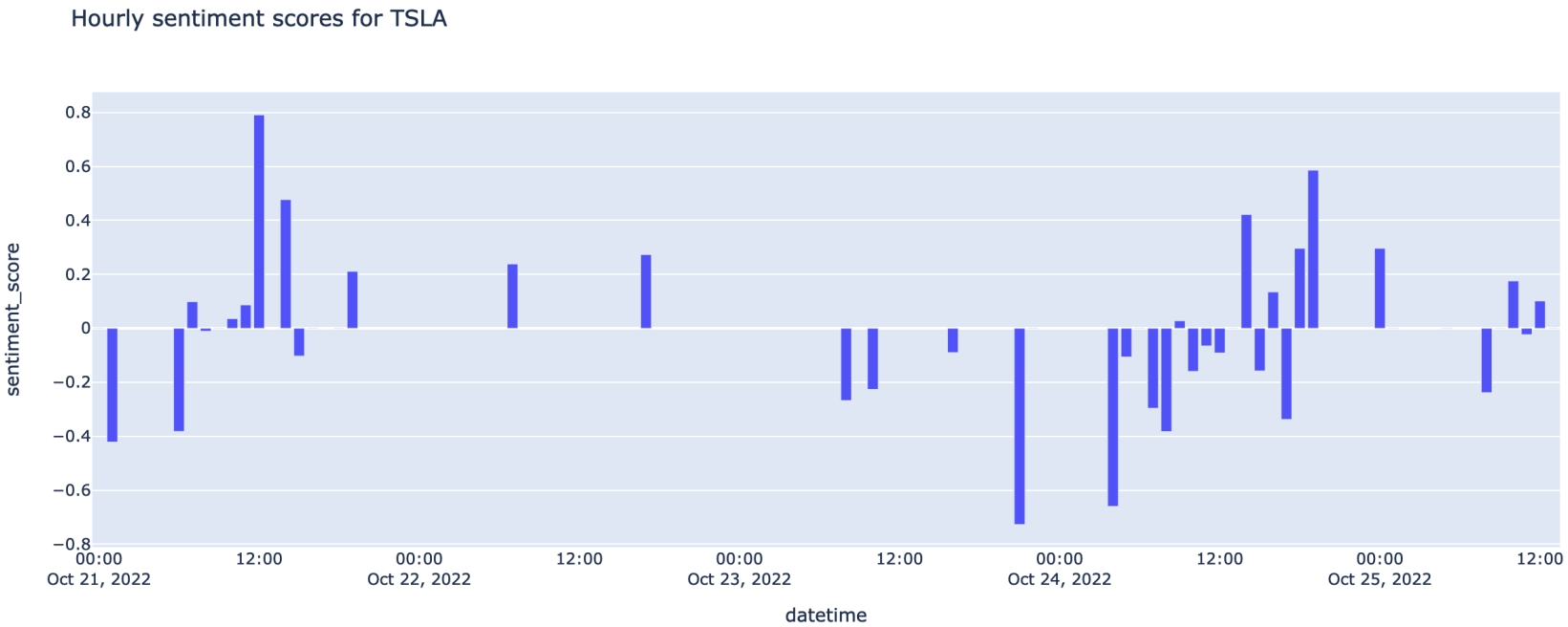
- Remove all regular expressions, verbs, and do a lower cases conversion
- Tokenization each article into word vectors
- Parse the data into a dataframe with three columns; 'date', 'time', 'headline'

	date	time	headline	datetime
0	Oct-25-22	01:11PM	GM Earnings: General Motors Affirms Outlook, A...	2022-10-25 13:11:00
1	Oct-25-22	12:27PM	Why Tesla, Rivian, and Nio Stocks All Popped T...	2022-10-25 12:27:00
2	Oct-25-22	12:10PM	PACCAR (PCAR) Q3 Earnings Top Estimates, Jump ...	2022-10-25 12:10:00
3	Oct-25-22	11:47AM	Cathie Wood Goes Bargain Hunting: 3 Stocks She...	2022-10-25 11:47:00
4	Oct-25-22	11:40AM	Elon Musk Could Buy Twitter This Week. Heres W...	2022-10-25 11:40:00
...	...	...	...	...
95	Oct-21-22	07:30AM	Tesla Stock Just Made a New 52-Week Low. Here'...	2022-10-21 07:30:00
96	Oct-21-22	06:41AM	Twitter tells staff no layoffs are planned, fo...	2022-10-21 06:41:00
97	Oct-21-22	01:36AM	Musk says recession could last until 2024	2022-10-21 01:36:00
98	Oct-21-22	01:27AM	Musk says recession could last until 2024	2022-10-21 01:27:00
99	Oct-21-22	12:03AM	Exclusive-Automakers to double spending on EVs...	2022-10-21 00:03:00

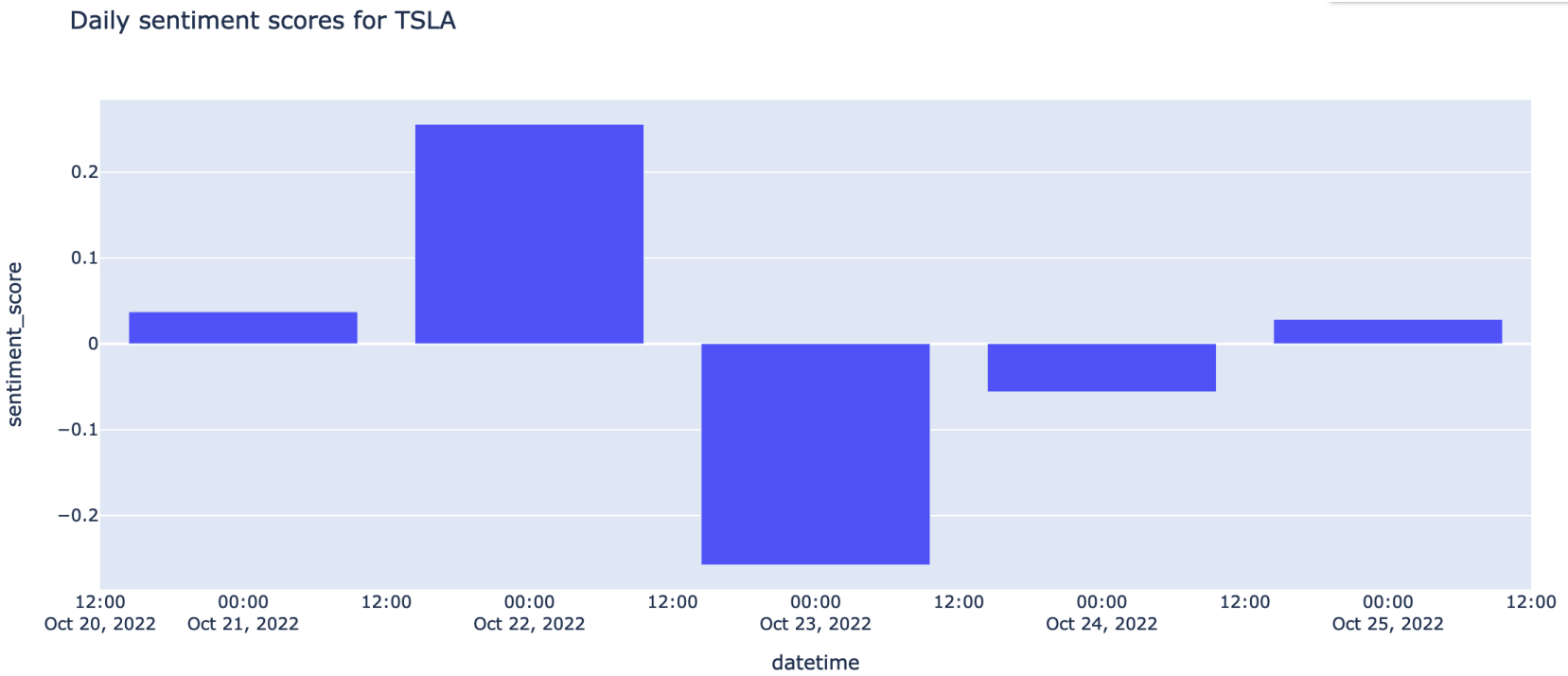
Use Natural Language Tool Kit(NLTK) to derive the polarity scores of the news.

datetime	headline	neg	neu	pos	sentiment_score
2022-10-25 13:11:00	GM Earnings: General Motors Affirms Outlook, A...	0.000	1.000	0.000	0.0000
2022-10-25 12:27:00	Why Tesla, Rivian, and Nio Stocks All Popped T...	0.000	1.000	0.000	0.0000
2022-10-25 12:10:00	PACCAR (PCAR) Q3 Earnings Top Estimates, Jump ...	0.000	0.816	0.184	0.2023
2022-10-25 11:47:00	Cathie Wood Goes Bargain Hunting: 3 Stocks She...	0.000	0.816	0.184	0.2023
2022-10-25 11:40:00	Elon Musk Could Buy Twitter This Week. Heres W...	0.000	1.000	0.000	0.0000
...	...	...	...	...	...
2022-10-21 07:30:00	Tesla Stock Just Made a New 52-Week Low. Here'...	0.149	0.851	0.000	-0.2732
2022-10-21 06:41:00	Twitter tells staff no layoffs are planned, fo...	0.217	0.783	0.000	-0.3818
2022-10-21 01:36:00	Musk says recession could last until 2024	0.318	0.682	0.000	-0.4215
2022-10-21 01:27:00	Musk says recession could last until 2024	0.318	0.682	0.000	-0.4215
2022-10-21 00:03:00	Exclusive-Automakers to double spending on EVs...	0.000	1.000	0.000	0.0000

Hourly Polarity Scores



Daily Polarity Scores



Data: Stock Price Data

Download data from yahoo finance using the API for python.

Feature Engineering

Additional features: current trend (C_t), stochastic strength index (RSI_t), and target- future trend (F_t).

$$RSI_t = 100 - \frac{100}{1 + RS_t} \tag{1}$$

where

$$RS_t = \frac{\text{Average Gain}_t}{\text{Average Loss}_t}$$

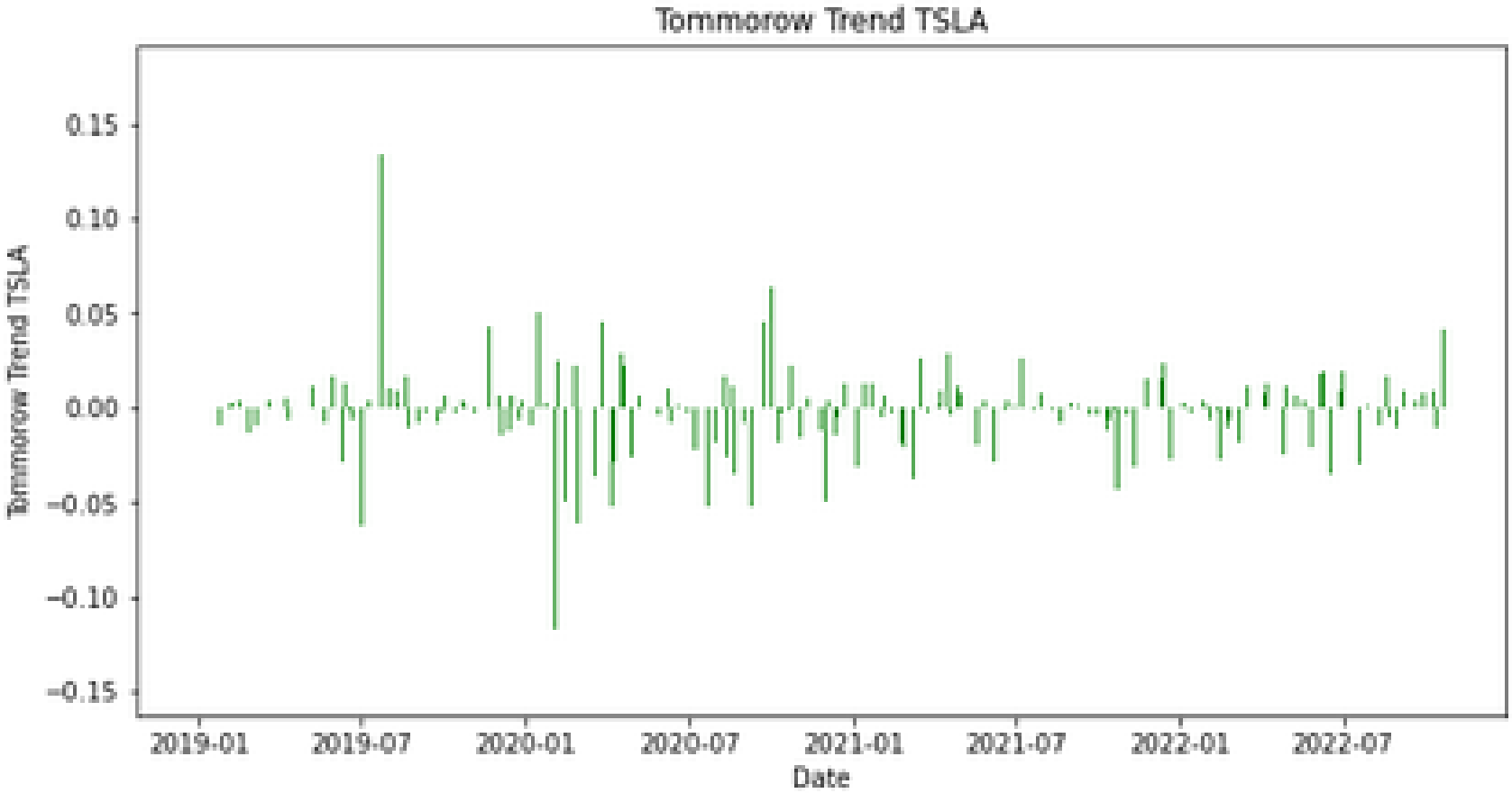
$$C_T = \begin{cases} \text{Up} & \text{if } P_c - P_o \geq 0 \\ \text{Down} & \text{if } P_c - P_o < 0 \end{cases} \tag{2}$$

$$F_T = \begin{cases} \text{Up} & \text{if } P_{f,c} - P_o \geq 0 \\ \text{Down} & \text{if } P_{f,c} - P_o < 0 \end{cases} \tag{3}$$

	Date	High	Low	Open	Close	Volume	Adj Close	RSI	prevClose	T_change	F_change
0	2019-01-02	28.990000	27.870001	28.260000	28.809999	15053700	28.809999	NaN	NaN	0.000000	0.015152
1	2019-01-03	29.180000	27.940001	28.379999	27.990000	19051700	27.990000	NaN	28.809999	0.015152	-0.014089
2	2019-01-04	30.100000	28.309999	28.389999	29.950001	23412600	29.950001	NaN	27.990000	-0.014089	-0.008278
3	2019-01-07	31.379999	29.770000	30.200001	31.340000	19917800	31.340000	NaN	29.950001	-0.008278	-0.011356
4	2019-01-08	32.049999	30.910000	31.700001	31.799999	18915200	31.799999	NaN	31.340000	-0.011356	0.000000
—	—	—	—	—	—	—	—	—	—	—	—
95	2019-05-20	37.730000	36.919998	37.119999	37.150002	9411900	37.150002	48.982162	37.500000	0.010237	-0.008540
96	2019-05-21	37.860001	37.330002	37.470001	37.470001	8861400	37.470001	50.832810	37.150002	-0.008540	0.001604
97	2019-05-22	39.320000	37.240002	37.410000	38.580002	21105400	38.580002	56.700214	37.470001	0.001604	0.011271
98	2019-05-23	38.290001	36.799999	38.150002	37.189999	18396800	37.189999	48.840169	38.580002	0.011271	-0.007473
99	2019-05-24	37.849998	37.270000	37.470001	37.410000	9303900	37.410000	50.021091	37.189999	-0.007473	0.000000


Price (Adj Close) TSLA
Tesla
Date
0
50
100
150
200
250
300
350
400
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01


RSI TSLA
0
20
40
60
80
100
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07


Tommorow Trend TSLA
Tommorow Trend TSLA
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
Date

ML Models

Training models: Linear Regression (LR), K-Nearest Neighbours (KN), Gradient Boosting (GB) and Random Forest (RF).

Linear Regression

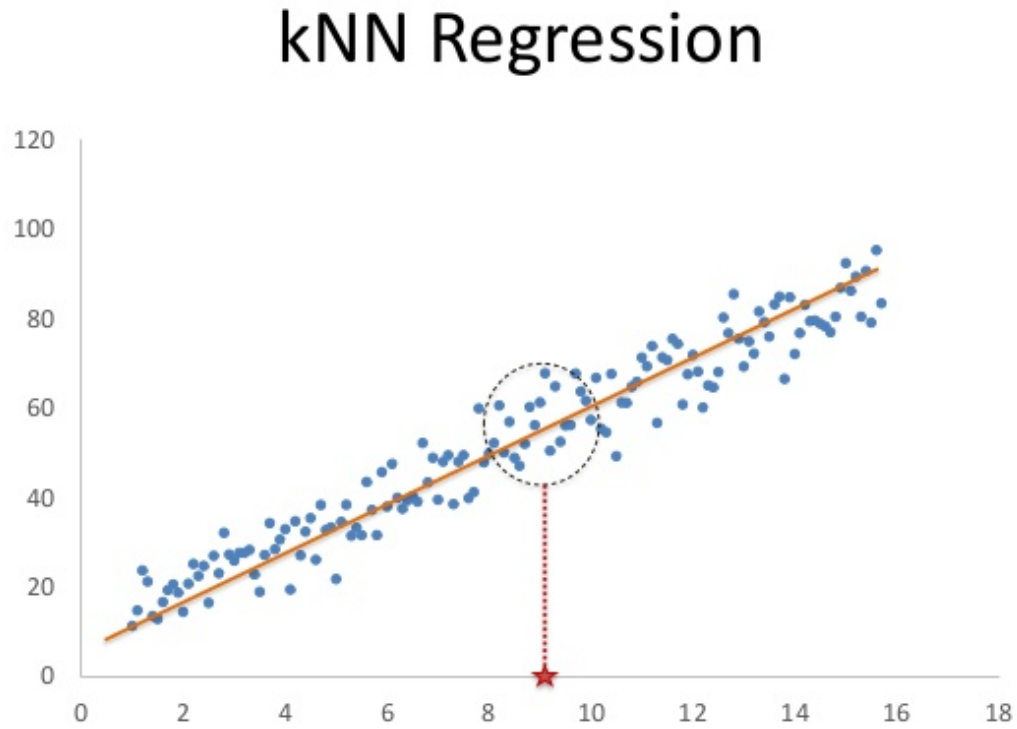
The model assume a linear relationship between input variables and the output variable. The mathematical representation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon, \tag{4}$$

K-Nearest Neighbors Regression

The K-Nearest Neighbors (KNN) regression is a non parametric supervised machine learning technique.

- Given a value for K and a prediction point x_0, the KNN regressor first identifies K points (observations) which are closest to the point x_0, denoted by N_0.



the functional evaluation $f(x_0)$ is given by

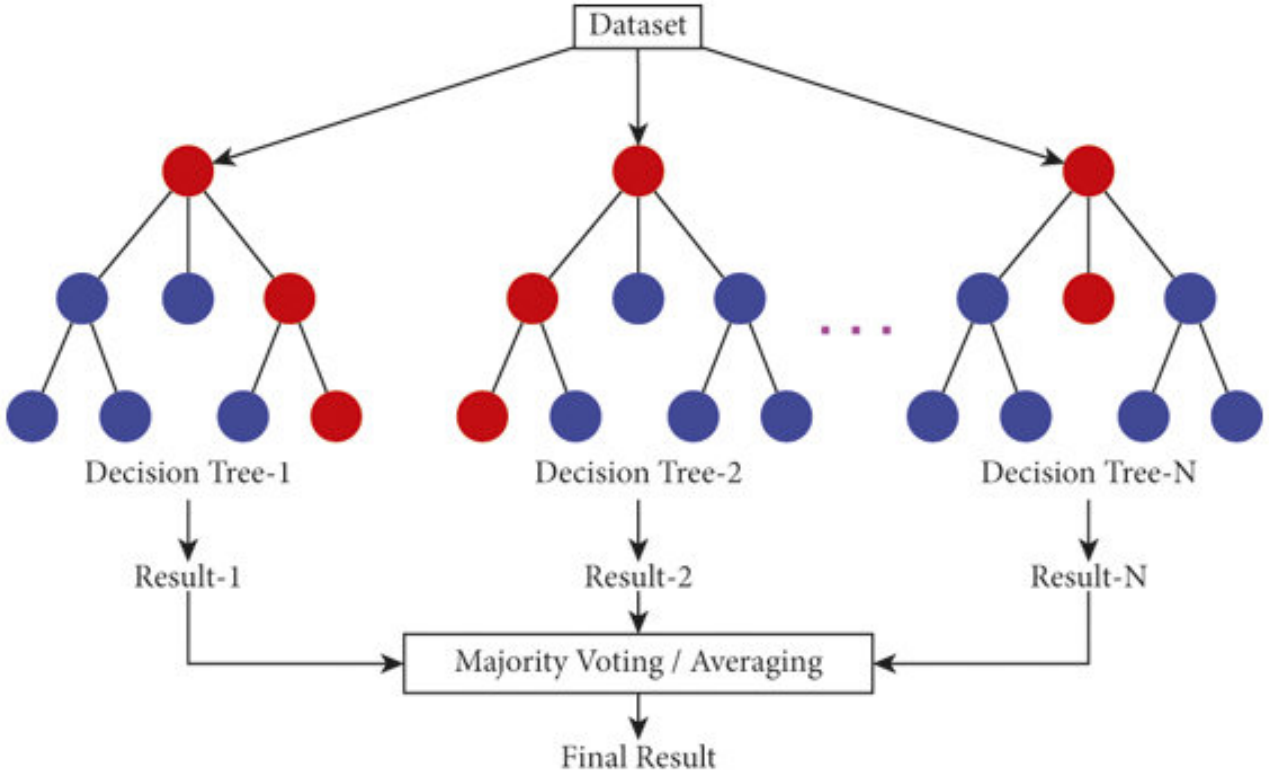
$$f(x_0) = \frac{1}{K} \sum_{y_i \in N_0} y_i, \tag{5}$$

- Optimal value of K: the algorithm seek to optimise the tradeoff between bias and variance of its estimations as presented in eq. (5).

Random Forest Regression

Tree based algorithm, it uses bootstrap and bagging to train weak tree learners.

Suppose our dataset is represented by $\{x_i, y_i\}_{i=1}^{n}$



- The algorithm bags repeatedly N times.
- At each bagging iteration a random sample (x_b, y_b) is selected with replacement from the training data.

- The model is fit on each of the samples to obtain f_b. After training, predictions on unseen data (x') is made by averaging the predictions from all learner trees using the formula

$$\hat{f} = \frac{1}{N} \sum_{b=1}^{N} f_b(x') \tag{6}$$

Gradient Boosting Regression

A tree based algorithm which combines a series of base learner models into a strong one.

- Suppose $h(x_i)$'s are base learner models, equiped with a $Softmax$ loss function

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_i}}, \ i = 1, \cdots, N. \tag{7}$$

The algorithm is as follow

- Step 1: Initialize the model with

$$F_0(x) = \arg\min_{\beta} \sum_{i=1}^{N} L(y_i, \beta). \tag{8}$$

- Step 2: Per iteration $m = 1 : M$, the gradient direction of residuals are computed as follow

$$y_i^* = \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)-F_{m-I(x)}}, i = \{1, 2, \cdots, N\} \quad (9)$$

- Step 3: The base learners are then fitted to the data to get the initial model using the least square method to obtain the following coefficient

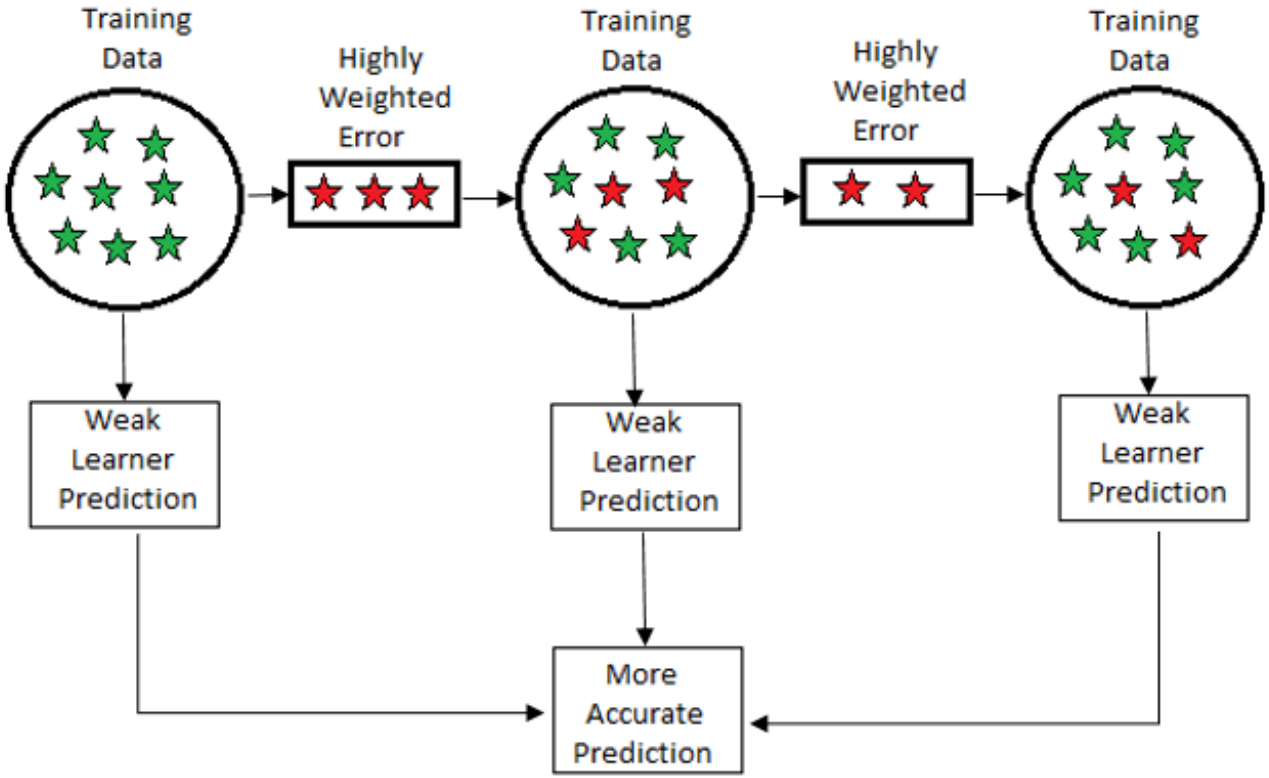
$$\alpha_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \left[y_i^* - \beta h(x_i; a)\right]^2. \quad (10)$$

- Step 4: The loss function is minimised and current model weights are recalculated using

$$\beta_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} L(y_i, F_{m-1}(x) + \beta h(x_i; \alpha)) \tag{11}$$

- Step 5: The model is finally updated using the following relation

$$F_m(x) = F_{m-1}(x) + \beta h(x_i; \alpha) \tag{12}$$



Model Evaluation and Validation

We splitted the data into training (80%) and testing (20%).

	Models	FB	TSLA	TWTR
NMSE	LR	−0.0177	−0.0195	−0.0173
	KN	−0.0140	−0.0176	−0.0161
	RF	−0.0134	−0.0182	−0.0186
	GB	−0.0142	−0.0172	−0.0165
R^2	LR	16.7 %	12 %	32.8 %
	KN	15.9 %	28.9 %	6.6%
	RF	84.7 %	84.5 %	81.2 %
	GB	99.1 %	98.5 %	99.1 %

Table: Model Performance

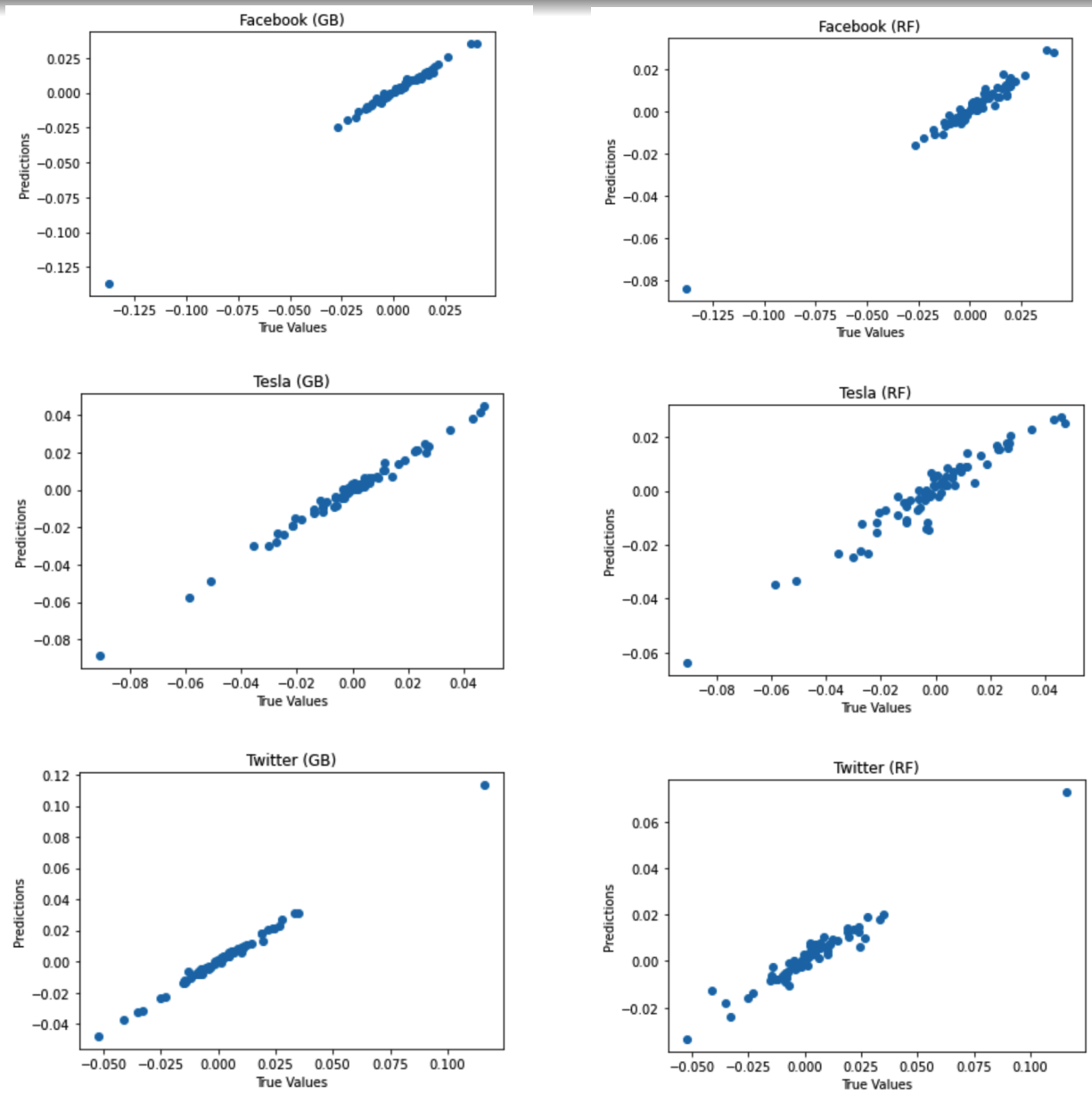


Figure: Predictions of the GB & RF for all stocks