



Tweet Sentiment Analysis Effect on Stock Price

Submitted by:-

Ankit Gupta(181IT107)
Ayush Rahangdale(181IT109)
Naman Vijayvargiya(181IT129)
Ashok Bhobhiya(181IT154)



INTRODUCTION

- Stock exchange is a subject that is highly affected by economic, social, and political factors.
- There are several factors e.g. external factors or internal factors which can affect and move the stock market.
- Stock prices rise and fall every second due to variations in supply and demand.
- “Stock Price Prediction Using Twitter Sentiment Analysis” a method for predicting stock prices is developed using Twitter steps articles.



OBJECTIVE

- Twitter opinions play an important role about what brand value in today and this can be a metric for deciding stock market variation.
- To make a prediction model for finding and analysing correlation between contents of tweets and stock prices and then making predictions for future prices by using machine learning.
- To show that how stock market depends on various political and economic factors like twitter.
- Check whether the positive sentiment helps a company in increasing in stock as more people will invest in it and vice-versa.
- Implement the model with sentiments and sentiments+past stock data and compare the models amongst themselves.



DATASET

1. Stock price Dataset (RAW)

Dataset is collected for apple company

Source-<https://finance.yahoo.com/quote/AAPL/history/>

Instances-565 and No of features is-6

2. Twitter Dataset for AAPL(RAW)

Source-<https://archive.org/search.php?query=collection%3Atwitterstream&sort=-publicdate>

Instances-20668 and No of features is 2

3. After Preprocessing total no of instances is 17650 and no of features is 7



METHODOLOGY

- Data Collection : Tweets on AAPL are extracted from twitter API. The tweets were collected using Twitter API and filtered using keywords like #Apple etc.
- Data Pre-processing :
 - Cleaning the data
 - Dealing with missing values
- Sentiment Analysis : Tweets are classified as positive, negative and neutral based on the sentiment present. Compound sentiment is found using positive and negative sentiment.
- Finding Correlation Between Sentiments and Stock Price
- Feature Scaling
- Model Training :

Data Collection:-

RAW stock data for Apple company are collected from yahoo finance along with Tweets for the company which are collected from internet archives.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2013-12-31	19.791786	20.045713	19.785713	20.036428	17.849323	223084400
1	2014-01-02	19.845715	19.893929	19.715000	19.754642	17.598297	234684800
2	2014-01-03	19.745001	19.775000	19.301071	19.320715	17.211735	392467600
3	2014-01-06	19.194643	19.528570	19.057142	19.426071	17.305593	412610800
4	2014-01-07	19.440001	19.498571	19.211430	19.287144	17.181829	317209200
...
560	2016-03-23	26.620001	26.767500	26.475000	26.532499	24.678110	102814000
561	2016-03-24	26.367500	26.562500	26.222500	26.417500	24.571148	104532000
562	2016-03-28	26.500000	26.547501	26.264999	26.297501	24.459534	77645600
563	2016-03-29	26.222500	26.947500	26.219999	26.920000	25.038527	124760400
564	2016-03-30	27.162500	27.605000	27.150000	27.389999	25.475679	182404400

565 rows × 7 columns

2	2014-01-01	#iPhone users are more intelligent than #Samsu...
3	2014-01-01	2013 Wrap-Up And Trading Set Review - Part III...
4	2014-01-01	Apple Screwed Up Big Time http://t.co/Q2Pzk2VO...
...
17645	2014-01-11	\$AAPL Apple to open new store in Brisbane CBD ...
17646	2014-01-12	\$AAPL What's Behind The Swift Rise In Apple St...
17647	2014-01-12	RT @SupremeSees: \$OXB Monday Gapper? http://t...
17648	2014-01-12	Apple Inc. (AAPL): What's Behind The Swift Ris...
17649	2014-01-12	Disruptive innovation \$AAPL \$HPQ \$GRMN http://...

17650 rows × 2 columns

Data Pre-processing :

Data collected from Tweepy contains emoji, links, HTML tag etc which is not required for getting sentiments from text. They are removed and tweets are preprocessed.

Since date is used to map stock closing data with Tweets thus stock market is close on weekends but we get tweets so required to fill those stock for training.

Missing Values= (prev day stock+ next day stock)/2

2	2014-01-01	#iPhone users are more intelligent than #Samsu...
3	2014-01-01	2013 Wrap-Up And Trading Set Review - Part III...
4	2014-01-01	Apple Screwed Up Big Time http://t.co/Q2Pzk2VO...
...
17645	2014-01-11	\$AAPL Apple to open new store in Brisbane CBD ...
17646	2014-01-12	\$AAPL What's Behind The Swift Rise In Apple St...
17647	2014-01-12	RT @SupremeSees: \$OXB Monday Gapper? http://t...
17648	2014-01-12	Apple Inc. (AAPL): What's Behind The Swift Ris...
17649	2014-01-12	Disruptive innovation \$AAPL \$HPQ \$GRMN http://...

17650 rows × 2 columns

3	2014-01-01	iPhone users are more intelligent than Samsung...
4	2014-01-01	2013 WrapUp And Trading Set Review Part III h...
...
17645	2016-03-31	REVIEW This is Apples best iPad AAPL http://t.co/L...
17646	2016-03-31	Apple now collecting some ResearchKit data fro...
17647	2016-03-31	AAPL Just got this email from AppleWhy Apples ...
17648	2016-03-31	RT businessinsider This guy found a hidden way...
17649	2016-03-31	Disney Infinity drops support for its Apple TV...

17650 rows × 2 columns

Before dealing with missing values

	Date	Tweets	Prices
0	2013-12-31	RT philstockworld Summary of Yesterdays Webcas...	20.0364
1	2014-01-01	RT philstockworld Summary of Yesterdays Webcas...	
2	2014-01-01	iTV Will Boost Apple httpco8dup4cQc08 AAPL APPLE	
3	2014-01-01	iPhone users are more intelligent than Samsung...	
4	2014-01-01	2013 WrapUp And Trading Set Review Part III h...	
...
17645	2016-03-31	REVIEW This is Apples best iPad AAPL httpstcoL...	
17646	2016-03-31	Apple now collecting some ResearchKit data fro...	
17647	2016-03-31	AAPL Just got this email from AppleWhy Apples ...	
17648	2016-03-31	RT businessinsider This guy found a hidden way...	
17649	2016-03-31	Disney Infinity drops support for its Apple TV...	

17650 rows × 3 columns

After dealing with missing values

	Date	Tweets	Prices
0	2013-12-31	RT philstockworld Summary of Yesterdays Webcas...	20.0364
1	2014-01-01	RT philstockworld Summary of Yesterdays Webcas...	19.8955
2	2014-01-01	iTV Will Boost Apple httpco8dup4cQc08 AAPL APPLE	19.8955
3	2014-01-01	iPhone users are more intelligent than Samsung...	19.8955
4	2014-01-01	2013 WrapUp And Trading Set Review Part III h...	19.8955
...
17645	2016-03-31	REVIEW This is Apples best iPad AAPL httpstcoL...	27
17646	2016-03-31	Apple now collecting some ResearchKit data fro...	27
17647	2016-03-31	AAPL Just got this email from AppleWhy Apples ...	27
17648	2016-03-31	RT businessinsider This guy found a hidden way...	27
17649	2016-03-31	Disney Infinity drops support for its Apple TV...	27

17650 rows × 3 columns

Sentiment Analysis:-

This is the most crucial step in the process. Sentiments like negative, Positive, Neutral that arise from the tweet are analysed and used as features in training for predicting the stock price. Corresponding to each text in the tweet sentiment score is evaluated and average out for the entire tweet.

These sentiments can have a direct influence on how people will invest on a stock of the company and thus play a vital role in closing stock price.

	Date	Tweets	Prices	Comp	Negative	Neutral	Positive
0	2013-12-31	RT philstockworld Summary of Yesterdays Webcas...	20.0364				
1	2014-01-01	RT philstockworld Summary of Yesterdays Webcas...	19.8955				
2	2014-01-01	iTV Will Boost Apple httpco8dup4cQc08 AAPL APPLE	19.8955				
3	2014-01-01	iPhone users are more intelligent than Samsung...	19.8955				
4	2014-01-01	2013 WrapUp And Trading Set Review Part III h...	19.8955				
...
17645	2016-03-31	REVIEW This is Apples best iPad AAPL httpstcoL...	27				
17646	2016-03-31	Apple now collecting some ResearchKit data fro...	27				
17647	2016-03-31	AAPL Just got this email from AppleWhy Apples ...	27				
17648	2016-03-31	RT businessinsider This guy found a hidden way...	27				
17649	2016-03-31	Disney Infinity drops support for its Apple TV...	27				

17650 rows × 7 columns

	Date	Tweets	Prices	Comp	Negative	Neutral	Positive
0	2013-12-31	RT philstockworld Summary of Yesterdays Webcas...	20.0364	0	0	1	0
1	2014-01-01	RT philstockworld Summary of Yesterdays Webcas...	19.8955	0	0	1	0
2	2014-01-01	iTV Will Boost Apple httpco8dup4cQc08 AAPL APPLE	19.8955	0.4019	0	0.69	0.31
3	2014-01-01	iPhone users are more intelligent than Samsung...	19.8955	0.5095	0	0.798	0.202
4	2014-01-01	2013 WrapUp And Trading Set Review Part III h...	19.8955	0	0	1	0
...
17645	2016-03-31	REVIEW This is Apples best iPad AAPL httpstcoL...	27	0.6369	0	0.741	0.259
17646	2016-03-31	Apple now collecting some ResearchKit data fro...	27	0	0	1	0
17647	2016-03-31	AAPL Just got this email from AppleWhy Apples ...	27	0	0	1	0
17648	2016-03-31	RT businessinsider This guy found a hidden way...	27	0.4767	0	0.86	0.14
17649	2016-03-31	Disney Infinity drops support for its Apple TV...	27	0.4019	0	0.816	0.184

17650 rows × 7 columns



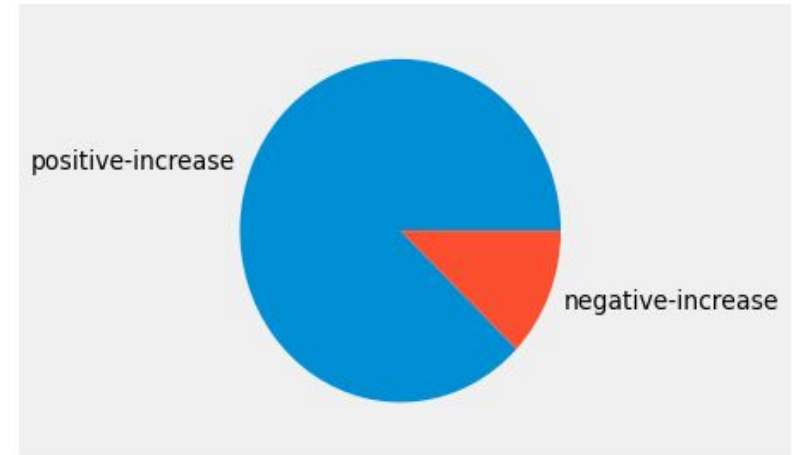
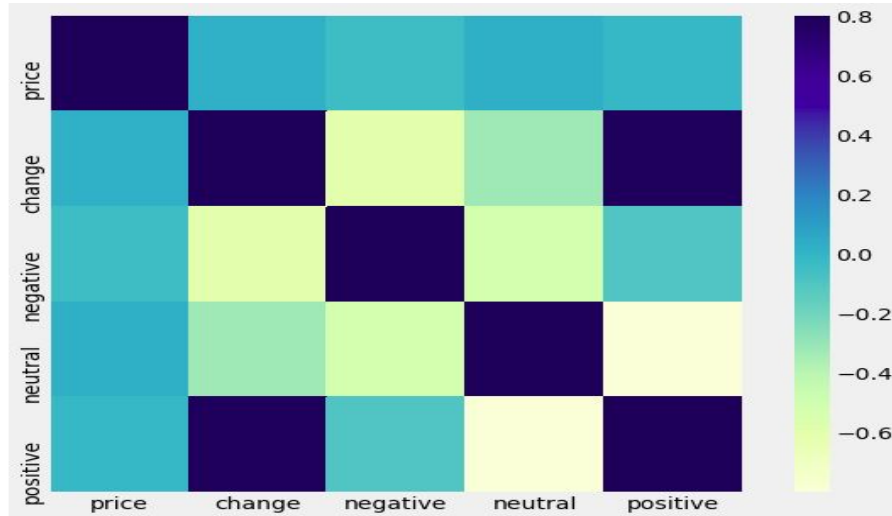
Negative sentiments words-



Positive Sentiments words-

Correlation Between Sentiments and Stock Price:-

Correlation is measured on a scale that varies from -1 to +1. This is used to assess a possible linear association between two continuous variables. Here we are finding between sentiment score and stock price. That will give an idea how the sentiments are related to stock price.






Getting Model Features for Training

Our model contains 17k instances but the problem is each day we get many tweets and thus we need to get one to one mapping between date, closing stock price and sentiments. We average all the sentiments of a particular date corresponding to that stock price.

$$Comp = |pos - neg| / (pos + neg)$$

This is used as a feature in training the model along with the past 60 days. The past 60 days stock is pass into model as a feature because when only sentiments were used in predicting the stock then the model was just remembering the data and thus was overfitting on training data and getting very high validation loss, thus it become quite important to pass past stock values along with sentiments in predicting the stock.

Before feeding the stock data, the data is normalised between 0-1 using Min-Max Scalar. Normalization is generally required as we are dealing with attributes on a different scale, otherwise, it may lead to a dilution in effectiveness of an equally important attribute(on lower scale) because of other attributes having values on larger scale.



```
▶ scaler = MinMaxScaler(feature_range=(0,1))
  scaled_data = scaler.fit_transform(dataset)
  scaled_data
```



```
[0.41374472],
[0.40635873],
[0.37470442],
[0.40506009],
[0.40911834],
[0.36723711],
[0.37527251],
[0.37527251],
[0.38330792],
[0.38298313],
[0.37129536],
[0.36204254],
[0.36675012],
[0.38825889],
```

After taking average the total data stock points are 782 and thus data is divided as 80% training data and 20% training data.

$X_{train}.shape = (627*61)$ where 60 features are the last 60 days stock price and 1 is the comp sentiment of the predicted day.

$Y_{train}.shape = 627$ which is the predicted stock price.

Test data comprises over 80 days of stock data.



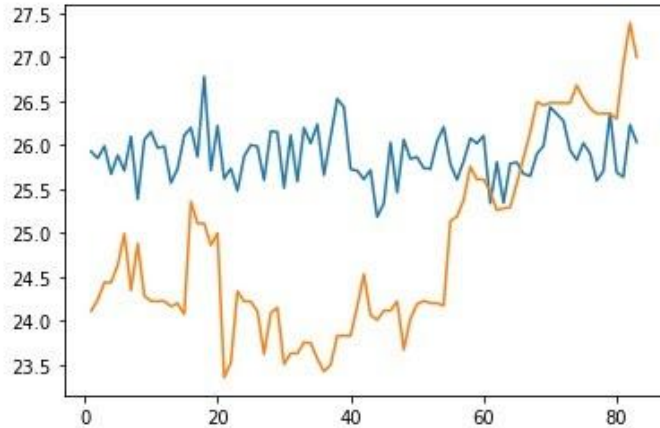
Models Used

Model is trained considering sentiments as the features and close value as the prediction. 90% is training dataset and 10% is testing dataset

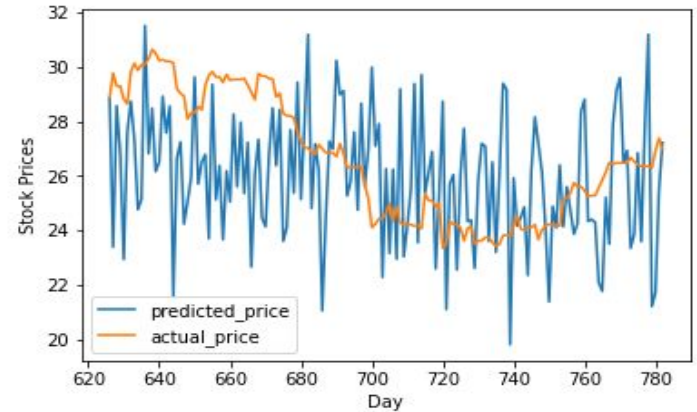
- **Training only with sentiments:**
 - LSTM
 - Random Regression
- **Training considering sentiments and past stock data:**
 - LSTM
 - Random Forest Regressor
 - KNN (with $k=5,10,15$)
 - SVM
 - MLP
 - XGBoost

Training only with sentiments

LSTM

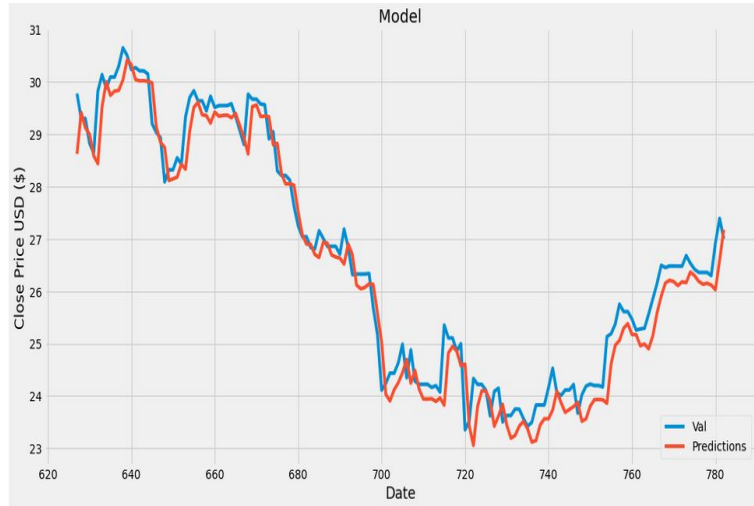


Random Forest Regressor



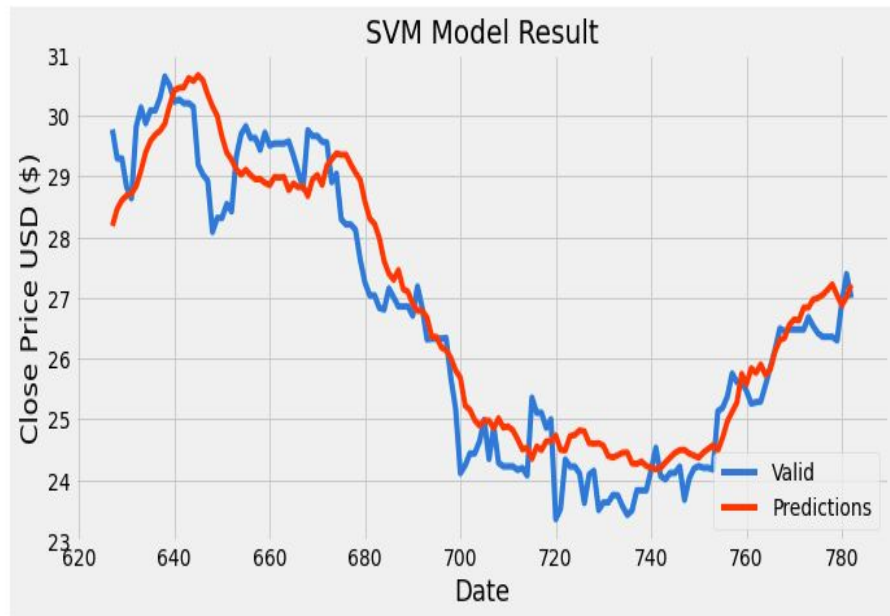
The figure shows the prediction of the stocks based on just sentiment scores. The model does not perform better because stock price does not only depend on sentiments and these sentiments may help in predicting uprising/ down of the stock price but not helpful in predicting the actual price. The model as seen is overfitting on train data and just remembering on prediction. Thus along with sentiments past stock data should also be used in predicting the stock price.

Training considering past data and sentiments

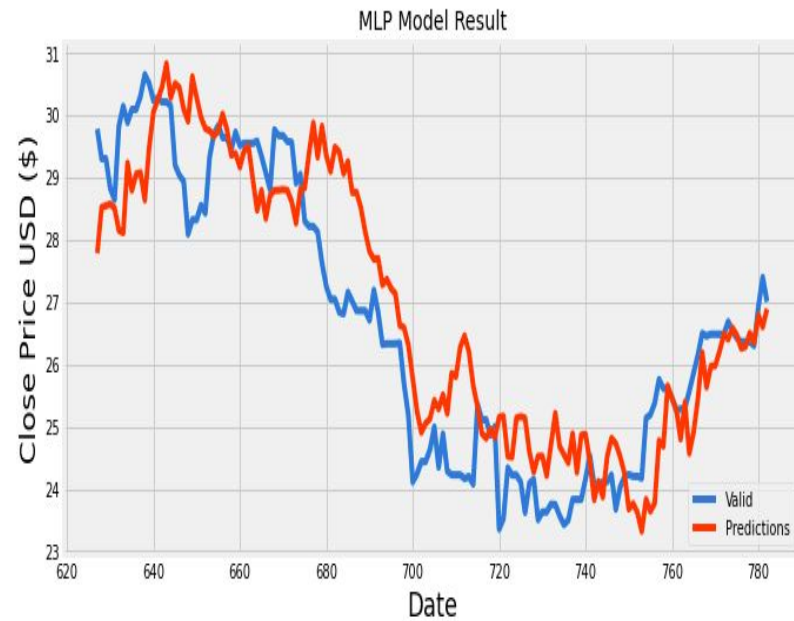


As seen, LSTM performs best in all the models used. LSTM is known well to handle the problem of vanishing gradients and thus performs very well. It is known for remembering the past data. The model is trained for 20, 50 and 100 epochs.

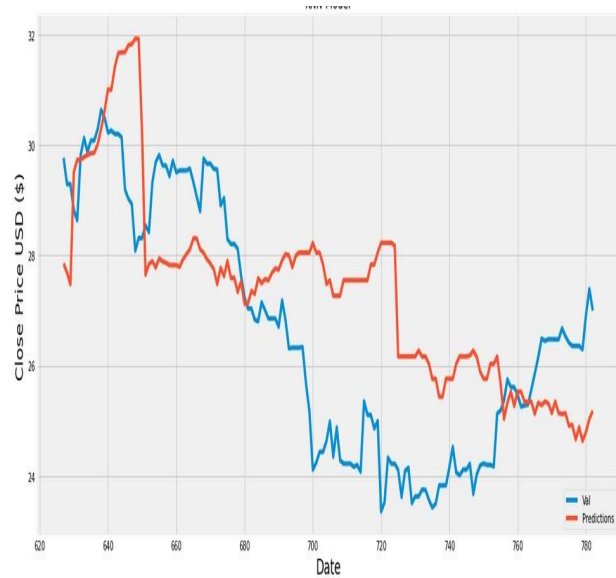
LSTM Result



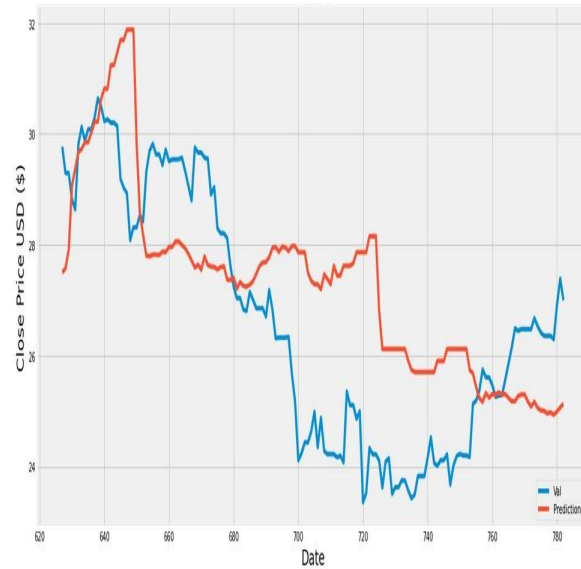
SVM result



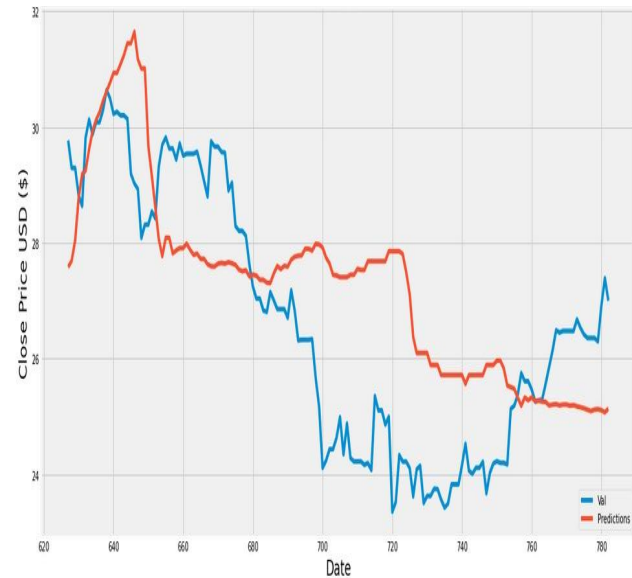
MLP result



KNN with k=5

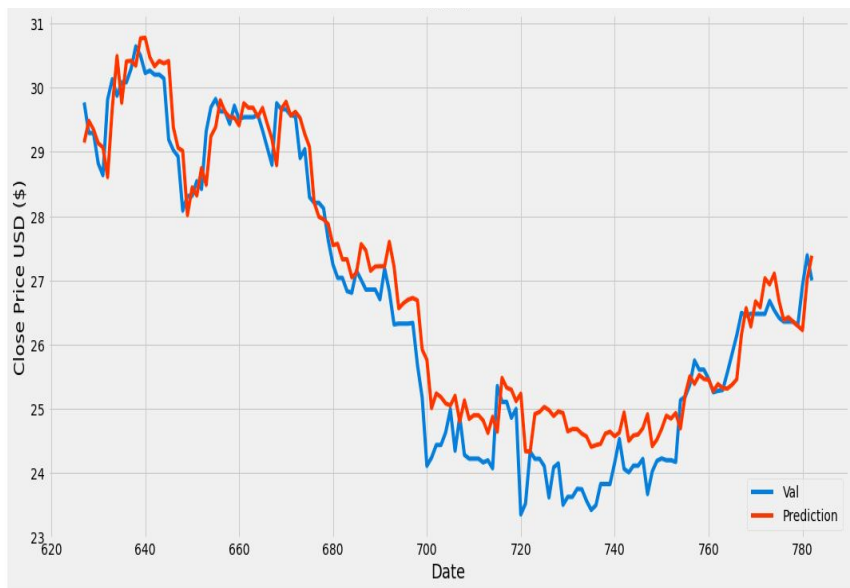


KNN with k=10

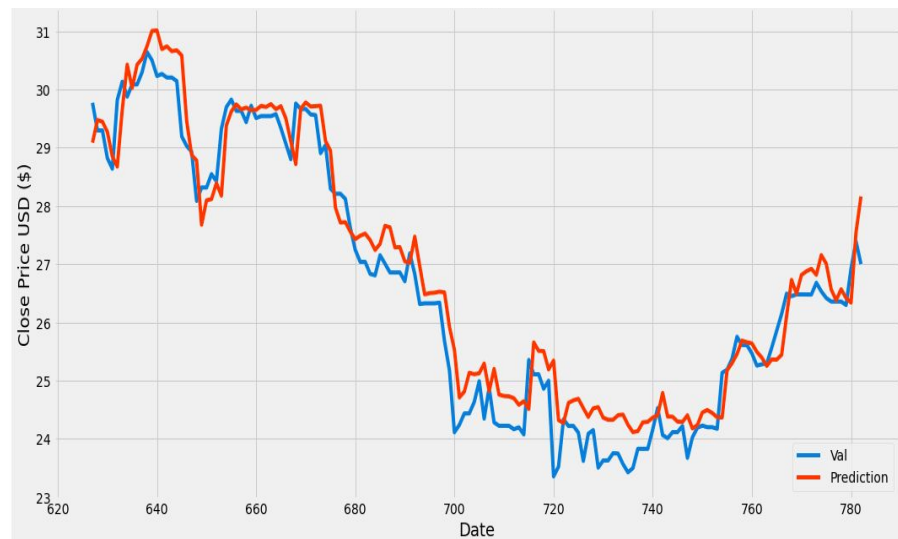


KNN with k=15

As seen while using KNN model as the value of K increases the predicted values tend to smooth.



XGBoost result



✓ 0s completed at 08:44

Random forest regressor result



Performance Evaluation

Model	MSE
LSTM	0.19469020816498622
Random Forest Regressor	0.2777
XGBoost	0.33997595595072805
KNN with K=5	4.016973119704898
KNN with K=10	3.82613058425248
KNN with K=15	3.5881210918559385
SVR	0.4849353561475266
MLP Regressor	1.1419813891059811



Conclusion

- We found that that strong correlation exists between stock price and tweet sentiments, but with this data we can only predict whether the stock will rise or fall, to predict the actual price we need to consider the past 60 days stock price data.
- We also found that LSTM is the best model to predict the stock price because it solves the problem of vanishing gradient and helpful in remembering the past data. Also, we found that mean square error (mse) is least while using LSTM.



Thank you