Analyzing the Effects of Twitter Sentiments on Stock Market (2015 to 2020)

Abstract

Stock price prediction is a complex and challenging task given the large number of factors involved, such as economic circumstances, political events, and others, that may affect the stock prices directly or indirectly. However, studies have shown that the stock markets are affected by public sentiments considerably. While more people spend time online voicing their opinions, social media platforms have become a primary source of public sentiment mining. Twitter especially, has caught the interest of researchers working on analyzing public opinions and sentiments in the stock market domain, as it is a rich source of real-time information including societal and personal opinions. In this study, the potential of Twitter sentiments as a key driving element in stock price forecasting was investigated. According to our research, we found that Twitter sentiments are an influential factor in stock prices movement and prediction of some of the major companies listed on the NASDAQ stock market exchange. It was also observed that the measure of 'positivity', 'negativity', and 'subjectivity' in tweets was significantly correlated with stock price movements thus contributing majorly towards the predictive stock price model. As the stock market data is a time-series data having continuous information, our results indicated that the Long-Short Term Memory neural network (LSTM) model gives state-of-theart results to forecast the stock prices using tweets sentiments and historical stock prices data.

Our analysis resulted in confident predictive stock price models for Apple Inc., Amazon and Google as their stock prices showed a high correlation with Twitter Sentiments.

Keywords: Sentiment Analysis, LSTM, Stock Market Prediction, Vader, Subjectivity, TextBlob

Analyzing the Effects of Twitter Sentiments on Stock Market (2015 to 2020)

The stock market functions in a highly volatile environment in which external elements, such as the influence of social media, have a noteworthy impact on stock market values. Stock market

predictions based on public sentiments expressed on Twitter has been an intriguing field of research. Previous studies have shown that it is possible to predict stock price movement to some extent using sentiment analysis data of Twitter. Researchers have examined the correlations between Twitter sentiments and stock price movements. For example, one research concluded that "The daily number of Tweets mentioning S&P 500 is significantly correlated with S&P 500 stock indicators and can be used to predict movements at market and sector level with 68% accuracy" (Mao, Wei, Wang, & Benyuan, 2012, p. 1). Sentiments regarding a company drive the demand for stocks, consequently impacting its price. For example, "positive news, and tweets on social media about a company would encourage people to invest in the stocks of that company and as a result, the stock price of that company would increase" (Pagolu, Challa, Panda, & Majhi, 2016, p. 1). Hence, understanding and interpretation of public sentiments on Twitter is crucial for various stock market stakeholders.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in the source text. In our study, we perform sentiment analysis using Vader.

Sentiment Intensity Analyzer to acquire the measure of 'positivity', 'negativity', 'neutrality', 'compound', and TextBlob to acquire 'polarity' and 'subjectivity'. The analysis is performed on the last five years (2015 to 2020) of tweets data associated with Apple, Amazon, Google, Tesla, and Microsoft, which are among the largest companies listed on the NASDAQ stock exchange market. Since these companies do not have the same products and services, heterogeneity is ascertained in our research and the predictive modelling results can be considered reliable. These companies have strong financial foundations and therefore, the correlation and forecasting results for these companies can be used as benchmarks by investors, entrepreneurs and even the companies themselves for studying the impact of twitter emotions on stock market dynamics. For

example, based on the approach discussed in this paper, Investors can develop risk mitigation strategies, entrepreneurs can adopt novel marketing strategies for promoting their start-ups and the established companies can devise modern product development strategies taking into account the voice of their customers. In today's dynamic and volatile world, it is essential to grasp people's sentiments, which transitions from feelings to opinions, because once the opinions solidify it is difficult to change them. Therefore, before it becomes a generic opinion, feelings should be captured to better understand customers and people by monitoring the live social media platforms like Twitter. Hence, our proposed model can be used to track such millions of sentiments for better forecasting of a company's stock prices.

This research aims to understand and address three crucial questions. First, it intends to analyse the correlation between different sentiment values of twitter data with different stock price attributes of top companies like Apple, Amazon, Google, Microsoft, and Tesla. Examining this aspect helps us in understanding how Twitter sentiments affect the stock market prices of different companies. Subsequently, we can figure out which sentiment values are significantly correlated with the stock market data. Second, it aims to discover how the same day aggregated tweets sentiment values, the previous day aggregated tweets sentiment values and the last threeday aggregated tweets sentiment values were related to the current day's stock prices values. This helps in comparing the effects of sentiments of immediate tweets and the effects of sentiments of tweets accumulated over time. Lastly, it aims to investigate how accurately we can predict the next day's stock prices of a company with the help of the previous day's stock prices combined with associated Twitter sentiment values.

To test our hypothesis and see how the stock prices of top companies have been impacted by the tweets, we aggregate the same day's tweets, previous day's tweets and last three-days' tweets and

combine these with the stock market data to perform statistical tests. After testing our hypothesis, we predict the next day's stock price values utilizing 'positivity', 'negativity', 'subjectivity' and last day's stock price as feature.

Stock price is a sequential time-series data, which means that the value of the current stock price depends substantially on value of the previous stock price. Recurrent Neural Networks are suitable for such problems, except they are prone to experience vanishing gradient problem.

> The memory of the original RNN will reduce its influence due to the number of complex layers after multiple recursions. Therefore, the Long Short-Term Memory (LSTM) neural network concept is introduced. It is a special and important RNN model which can memorize long-term or short-term values and flexibly allow the neural network only to retain the necessary information (Ko & Chang, 2021, p. 2).

Hence, our analysis focuses on the implementation of LSTM models for stock price forecasting. The paper further explains this aspect in the 'Methodology' section.

In today's competitive world, it is crucial to have a cutting-edge methodology that can incorporate public sentiments with stock market data to understand and predict future trends. This research tries to develop one such technique and it can be further explored and refined by researchers.

Literature Review

In the past decades, the stock markets have expanded remarkably, leading to more and more people getting engaged in stock price forecasting research. Recently, researchers have started exploring the domain of Natural Language Processing for stock forecasting. The research by (Tetlock, 2007) examined probable relationships between the media and the stock market using Wall Street Journal data and discovers that strong negativity generates downward pressure on market prices. A study by (Mao, Counts, & Bollen, 2011) uses a wide range of online data like Twitter feeds, news

headings and utilized sentiment tracking measures to predict financial market values. Researchers have also discovered that “Sentiments can drive short-term market fluctuations which in turn causes a disconnect between the stock price and the true value of a company’s share” (Shah, Isah, & Zulkernine, 2019, p. 2). Thus, the above studies support our assumption that sentiments of social media, like Twitter sentiment, are a significant predictor of daily market returns.

One of the publications in this area is by Bollen et al.(2011) who investigated “whether measurements of collective mood states derived from large scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA, stock market index) over time” (Bollen, Mao, & Zeng, 2011, p. 1). They implemented Granger causality analysis and Self-Organizing Fuzzy Neural Networks to predict the changes in the DJIA closing values. Their research resulted in “an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA” (Bollen, Mao, & Zeng, 2011, p. 1). They have used ‘OpinionFinder’ and ‘Google Profile of Mood States (GPOMS)’ for sentiment analysis and have examined the correlation of lag values of sentiments with the stock prices. In our study, we will use ‘VADER (Valence Aware Dictionary and Sentiment Reasoner)’ and ‘TextBlob’ sentiment analysis tools. Also, rather than utilizing lag values, we have inspected the results of the same day, the previous day, and the last three-day aggregated tweets sentiments since we assume that the current day’s Twitter sentiments have some effects on the current day’s stock prices.

The researchers Pagolu et al. (2016) and Kordonis et al. (2016) have tried to use Twitter sentiment analysis to predict stock market movements and both papers found a strong correlation between Twitter sentiment and the changes in companies’ stock prices. These studies inspired us to use Twitter sentimental analysis in predicting the stock market values and we found a significant correlation between these values to move forward with the forecasting models.

Recently, Brian et al. investigated “the correlation of sentiments of the public with stock increase and decreases using Pearson correlation coefficient for stocks” (Dickinson & Hu, 2015). However, we took a different approach to find the correlation between Twitter sentiment values and stock prices and used the Spearman correlation tests as it does not assume the normal distribution of the data.

The study by Namini et al. (2018) investigated whether the deep learning algorithms for forecasting time series data, such as “Long Short-Term Memory (LSTM)”, are superior to the traditional algorithms like “Autoregressive Integrated Moving Average (ARIMA)”. Their study concluded that LSTM outperformed the ARIMA model with an average reduction in error rate of 84%. Hence, the literature review emphasized the use of the Long Short-Term Memory (LSTM) model instead of the Regressor models or ARIMA model. We further experimented with LSTM model configurations to observe any notable difference in the results with respect to hyperparameter tuning.

Datasets

Datasets Source: Kaggle

Dataset Files: Company.csv, Company_Tweet.csv, Tweet.csv, CompanyValues.csv

The datasets were published by Mustafa et al. (2020) and they are a part of the ‘Speculator and Influencer Evaluation in Stock Market by Using Social Media’ conference paper which was published in the 2020 IEEE International Conference on Big Data under the 6th Special Session on Intelligent Data Mining track.

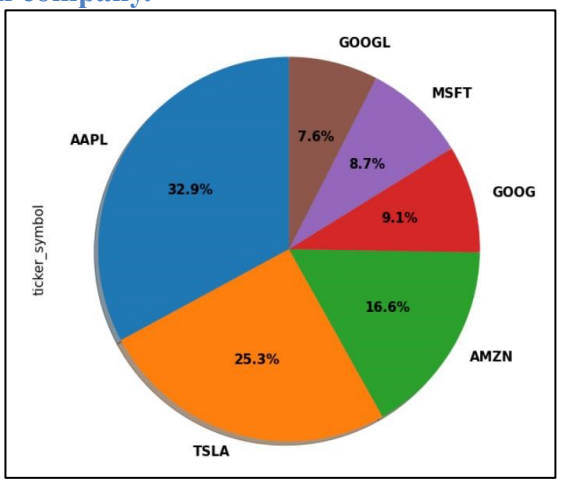
‘Company.csv’ contains the ticker symbols of the companies. Ticker symbols are the symbols used in NASDAQ to represent a company, for example, Apple is represented by

'AAPL'.

'Company_Tweet.csv' contains the unique tweet id associated with a tweet and the companies linked with the tweet id. There are a total of 3717964 unique tweet ids in this dataset.

Figure 1: The ticker symbols associated with each company and the pie chart depicting the percentage of tweets associated with each company.

	ticker_symbol	company_name
0	AAPL	apple
1	GOOG	Google Inc
2	GOOGL	Google Inc
3	AMZN	Amazon.com
4	TSLA	Tesla Inc
5	MSFT	Microsoft



'Tweet.csv' contains the different features of tweets. It contains the tweet id, the author of the tweet, the date on which the tweet was posted, the text contained in the tweet, the number of comments and likes on the tweet, and the number of times the tweet was retweeted.

Table 1: Attributes associated with each tweet

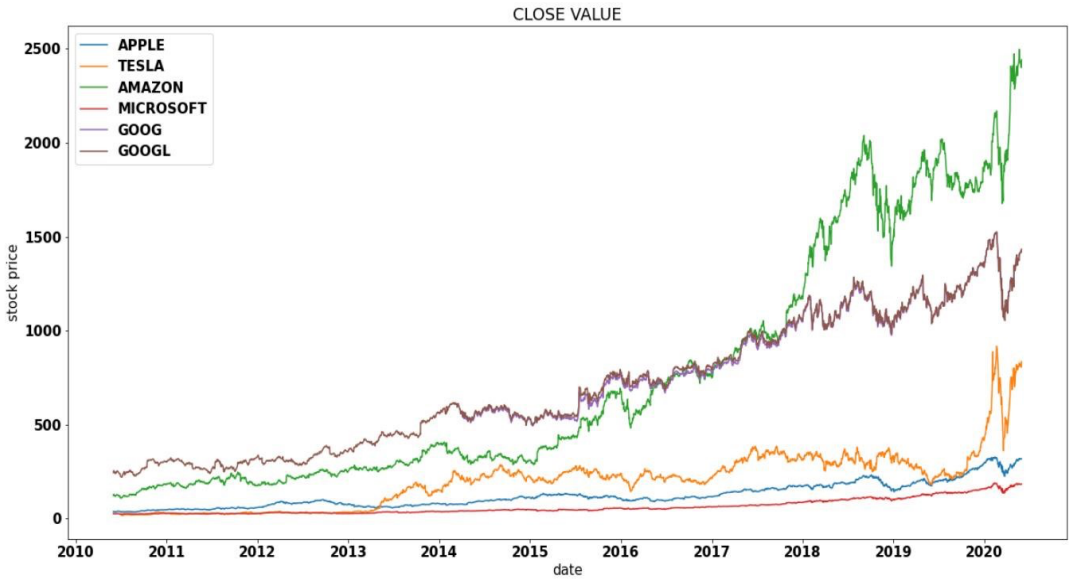
Feature	Description
tweet_id	Unique tweet id of a tweet
writer	Username of the author
post_date	The date on which the tweet was posted (in form of seconds since epoch)
body	Text of the tweet
comment_num	Number of comments
retweet_num	Number of retweets
like_num	Number of thumb-up

'CompanyValues.csv' contains the stock market data of different companies. For each day, it consists of the value of the first traded stock, the value of the last traded stock, the lowest value at which the stock was traded, and the highest value at which the stock was traded.

Table 2: Elements of stock market prices

Stock Price	Description
open_value	Price at which the first stock is traded
close_value	Price at which the last stock is traded
low_value	The lowest price at which the stock is traded
high_value	The highest price at which the stock is traded

Figure 2: Close value of each company from 2010 to 2020



Methodology

Data Preparation

We have focused on the analysis of aggregated tweets and not on the analysis of individual tweets because an ample number of tweets are posted in a single day, hence aggregating those tweets, and performing sentiment analysis gives better overall sentiment values of the general public's opinion on Twitter. First, we prepared the following types of datasets for each of the 6 companies:

(1) Same Day aggregated tweets dataset: contained aggregated value of tweets and their attributes for the current day.

(2) Previous Day aggregated tweets dataset: contained aggregated value of tweets and their attributes for the previous day.

(3) Last 3 Days aggregated tweets dataset: contained aggregated value of tweets and their attributes for the last 3 days.

For aggregation, we joined all the tweets in the desired period, removed the author from the dataset, and summed the comment numbers, retweets numbers, and like numbers. Second, we cleaned and pre-processed each of the above-prepared datasets, determined 'polarity' and 'subjectivity' values using TextBlob, and 'positivity', 'negativity', 'neutrality', and 'compound' values using Vader Sentiment Intensity Analyzer (SIA). We appended these sentiment intensity values to the original datasets. Third, we merged the sentiment values dataset and the stock market dataset to perform hypothesis testing and predictive modelling.

Distribution graphs of the variables present in the above data frames were not found to be normally distributed, which implied that the 'Spearman' correlation coefficient had to be used in place of the 'Pearson' correlation coefficient, to measure the degree of correlation among variables.

Figure 3: Methodology Flow Chart

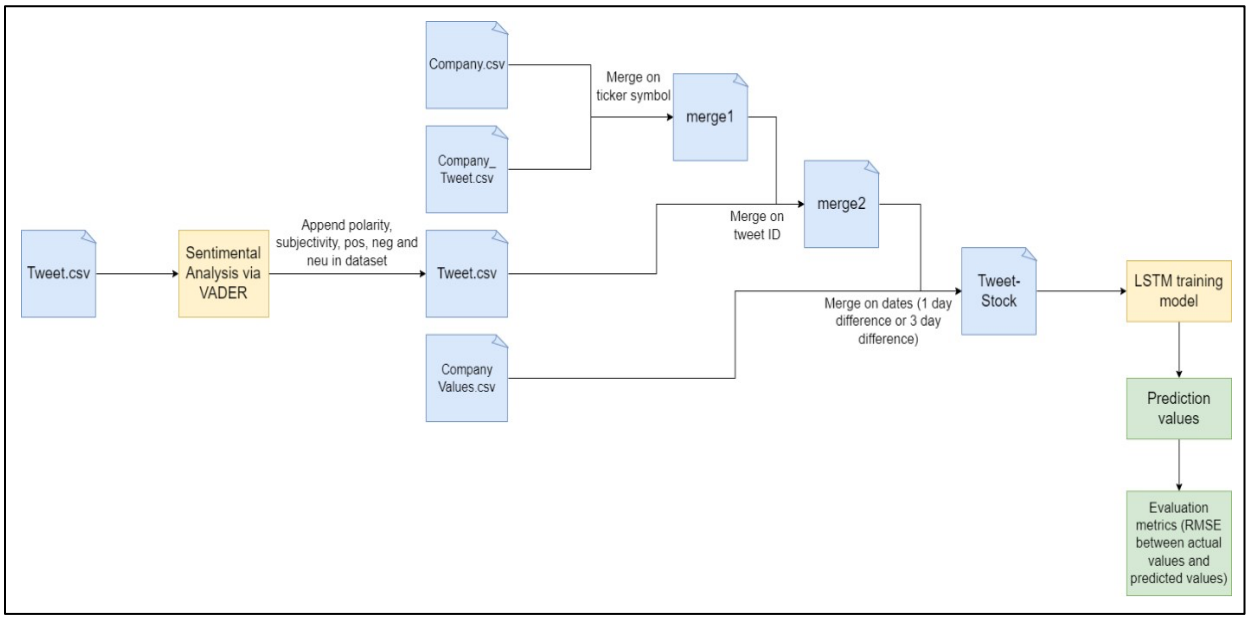
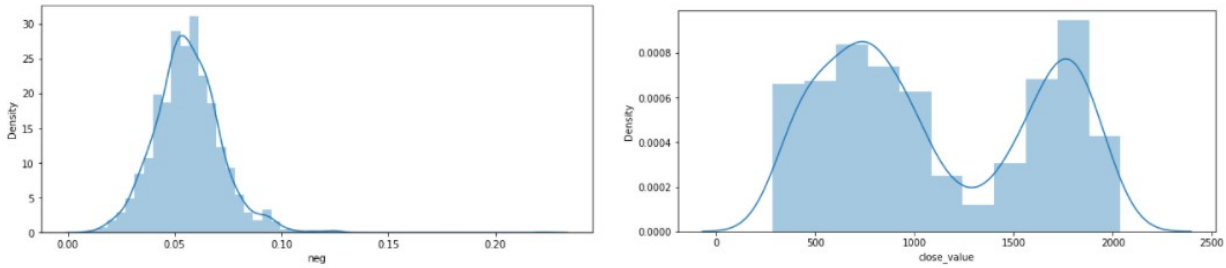


Figure 4: 'negativity' and 'close_value' for Amazon showing non-normal distribution



Our analysis suggested that 3 companies showed a strong correlation between their stock prices and Twitter sentiment values. These companies were Apple (APPL), Amazon (AMZN), and Google Inc (GOOGL). For these companies, it was noticed that 'subjectivity', 'positivity', and 'negativity' are strongly correlated with the stock prices.

We further obtained the 'p-values' for these correlation coefficients to understand the significance of correlation. The p-values were fell below the significance level (alpha threshold) of 0.05 in all the cases. Hence it was proved that the correlations were statistically significant with a corresponding confidence level of 95%.

Figure 5: Apple (AAPL), Amazon (AMZN) and Google (GOOGL) showing significant Spearman Correlation Coefficient

	Spearman Correlation Coefficient																	
	Apple						Amazon						Googl					
	same day			last 3 days			same day			last 3 days			same day			last 3 days		
Values	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity
open_value	0.51	-0.33	0.51	0.54	-0.42	0.53	0.24	0.21	0.36	0.21	0.22	0.47	0.36	-0.03	0.47	0.39	-0.07	0.6
close_value	0.51	-0.34	0.51	0.54	-0.43	0.54	0.24	0.21	0.36	0.21	0.21	0.47	0.36	-0.03	0.47	0.39	-0.08	0.6
low_value	0.5	-0.34	0.51	0.54	-0.43	0.54	0.24	0.21	0.36	0.21	0.21	0.47	0.36	-0.03	0.47	0.39	-0.08	0.6
high_value	0.51	-0.33	0.51	0.54	-0.42	0.53	0.24	0.21	0.36	0.21	0.22	0.46	0.36	-0.02	0.46	0.39	-0.07	0.59

Figure 6: Tesla (TSLA), Microsoft (MSFT) and Google (GOOG) showing insignificant Spearman Correlation Coefficient

	Spearman Correlation Coefficient																	
	Tesla						Microsoft						Goog					
	same day			last 3 days			same day			last 3 days			same day			last 3 days		
Values	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity
open_value	0.21	0.22	0.08	0.25	0.21	0.12	0.15	-0.04	0.18	0.13	-0.14	0.27	-0.07	0.11	0.07	-0.09	0.18	0.06
close_value	0.22	0.21	0.07	0.26	0.2	0.11	0.15	-0.05	0.18	0.13	-0.14	0.27	-0.07	0.11	0.07	-0.09	0.18	0.06
low_value	0.21	0.2	0.08	0.25	0.2	0.12	0.15	-0.05	0.18	0.13	-0.14	0.27	-0.07	0.1	0.07	-0.1	0.18	0.06
high_value	0.22	0.22	0.07	0.25	0.22	0.11	0.15	-0.04	0.18	0.13	-0.14	0.27	-0.07	0.11	0.07	-0.09	0.18	0.06

Model training

Since we want to predict the next day's stock price based on the previous day's stock price combined with Twitter sentiment values, our problem becomes a multivariate time series forecasting problem. In other terms, we have more than one time-dependent variable, i.e., the target variable is dependent on its past values and some other variables. This implies that we can neither use any regressor model nor can we use any univariate time series forecasting model like ARIMA (Autoregressive integrated moving average) or even extensions of ARIMA, such as SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors), which also considers seasonality in a time-series model.

Literature review suggests that "RNN is often used among lots of ANN to deal with timeseries tasks, and it is also one of the most suitable techniques for dynamic time series forecasting" (Ko & Chang, 2021, p. 5). However, using RNN introduces the risk of the model running into a vanishing gradient problem. Scholars state that "RNN will have the problem of gradient disappearance and explosion with multiple recursions. LSTM neural network has been developed to strengthen the operation of the RNN in the artificial intelligence fields" (Ko & Chang, 2021, p. 5).

For model training, the dataset is split into 70% and 30% for the training data and testing data, respectively. Initially, we train a baseline Keras LSTM model that has a visible layer with 4 inputs ('positivity', 'negativity', 'subjectivity' and previous day stock price), a hidden layer with 50 LSTM blocks, and an output layer that makes a single value prediction (next day stock price). We vary the LSTM model configurations to observe: (1) the impact of increasing LSTM blocks in a hidden layer; (2) the effects of adding a Dropout layer. Afterwards, we compare the Root Mean

Squared Error (RMSE) values obtained in all these models. The following LSTM model configurations for model training are implemented:

(1) LSTM 1: 50 LSTM blocks (hidden layer)

(2) LSTM 2: 100 LSTM blocks (hidden layer)

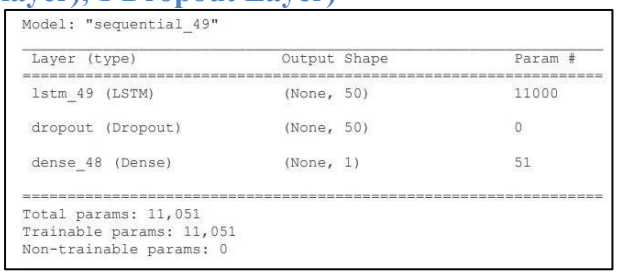
(3) LSTM 3: 200 LSTM blocks (hidden layer)

(4) LSTM 4: 50 LSTM blocks (hidden layer), 1 Dropout Layer

(5) LSTM 5: 100 LSTM blocks (hidden layer), 1 Dropout Layer

(6) LSTM 6: 200 LSTM blocks (hidden layer), 1 Dropout Layer

Figure 7: LSTM 1 (50 LSTM blocks (hidden layer)) and LSTM 4 (50 LSTM blocks (hidden layer), 1 Dropout Layer)

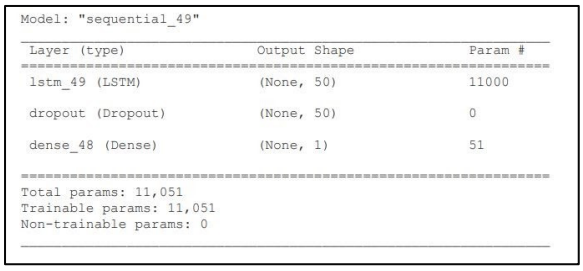


```
Model: "sequential_49"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_49 (LSTM)               (None, 50)                11000

 dropout (Dropout)            (None, 50)                0

 dense_48 (Dense)             (None, 1)                 51

=================================================================
Total params: 11,051
Trainable params: 11,051
Non-trainable params: 0
```



```
Model: "sequential_49"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_49 (LSTM)               (None, 50)                11000

 dropout (Dropout)            (None, 50)                0

 dense_48 (Dense)             (None, 1)                 51

=================================================================
Total params: 11,051
Trainable params: 11,051
Non-trainable params: 0
_________________________________________________________________
```

Results

The forecast errors are evaluated with the Root Mean Square Error (RMSE) values of the test data with the predicted output, which is calculated as shown below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

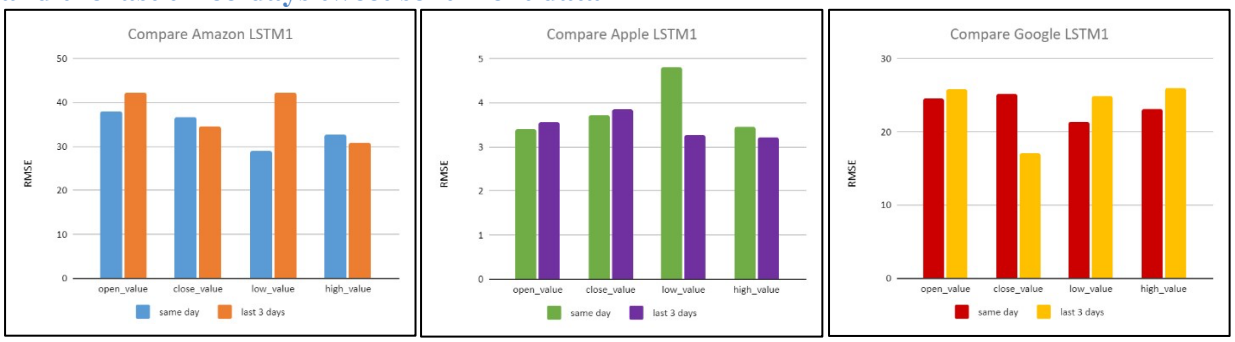
RMSE is one of the most used measures for evaluating the quality of predictions. It uses Euclidean distance to show how far predictions differ from measured true values.

Comparing the effects of sentiments of same-day tweets and the effects of sentiments of tweets accumulated over the last three days:

To address the second research question, we utilized the same-day tweets sentiment data, and last three-days tweets sentiment data to train separate stock price forecasting models.

Fig. 12 shows the comparison between the RMSE values obtained when we utilized same-day tweets sentiment values and the last three-days tweets sentiment values in the baseline LSTM model to predict the stock market prices (open value, close value, high value, low value) for Apple, Google, and Amazon.

Figure 8: RMSE comparison for baseline LSTM trained with same-day tweet sentiment data and the last three-days tweet sentiment data



Apple: The LSTM model based on the last three-day tweets sentiment values is performing better in most cases than the model for same-day tweets sentiment values. That means the sentiment values of accumulated tweets over time contribute better to stock prices prediction than sentiment values of same day's aggregated tweets. RMSE value is below 5.0 for both cases which indicates that the prediction models are a good fit for our data. The RMSE results indicate that the model has an average error of about 5 dollars (on the scale of hundred dollars) on the testing dataset.

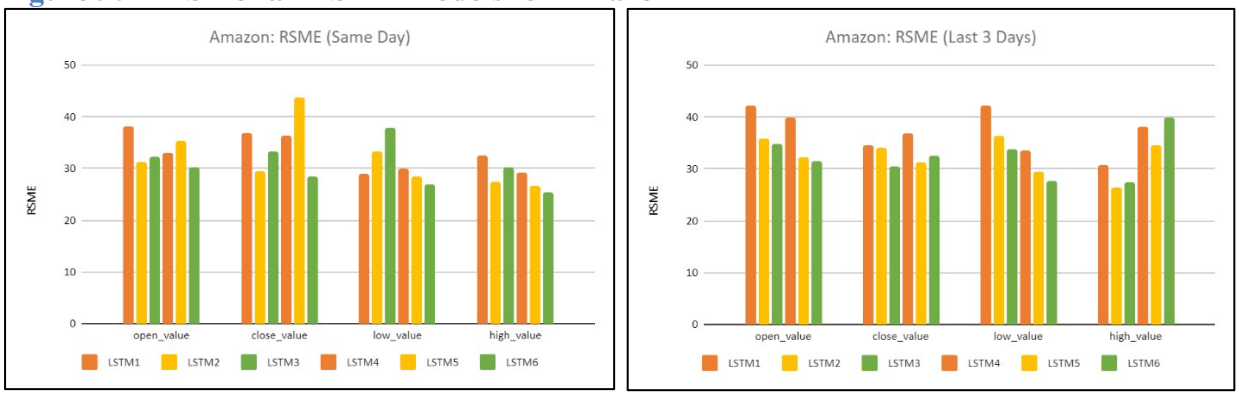
Amazon & Google: For both the companies, the baseline LSTM model showed better performance for stock prediction while utilizing same-day sentiment values than the last threeday sentiment values. The RMSE values obtained for Amazon and Google are under 40.0 and 30.0 (approx.) respectively. This indicates that, for Amazon, the model has an average error of about 40

dollars (on the scale of thousands), and for Google, the model has an average error of about 30 dollars (in the scale of thousands) on the testing datasets.

Analyzing the Effects of Varying LSTM configurations

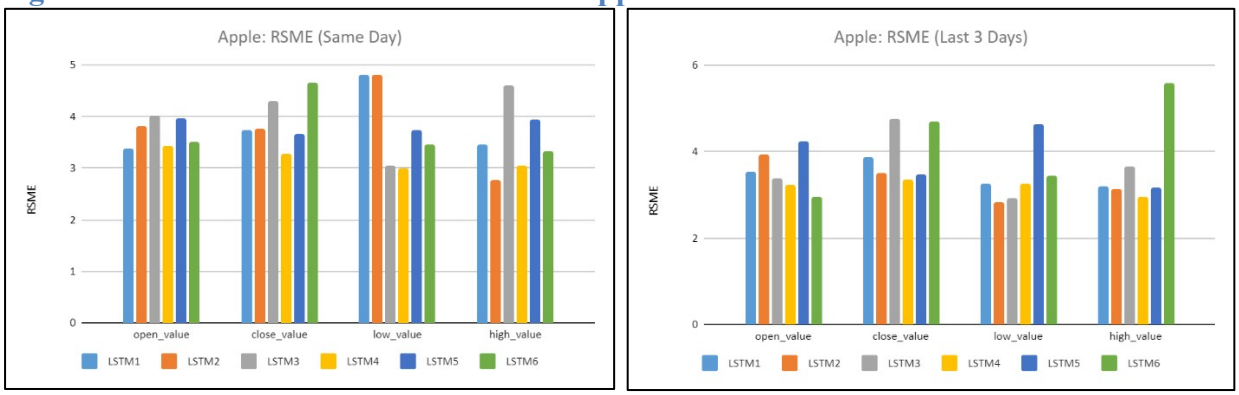
Some meaningful observations were made through experimental analysis of varying LSTM configurations, which can help us improve our model:

Figure 9: RMSE of all LSTM models for Amazon



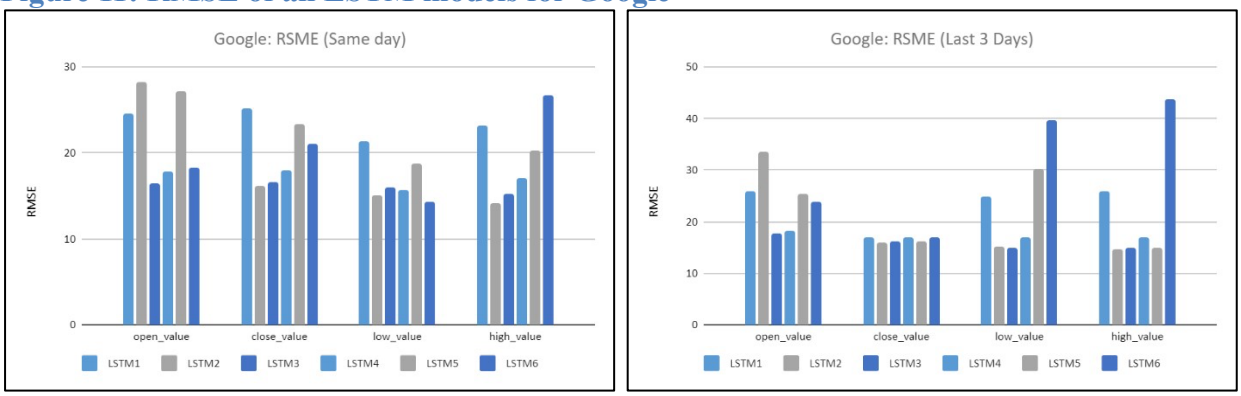
Amazon: 'LSTM 6' (200 LSTM blocks (hidden layer), 1 Dropout layer) shows better performance and lower RMSE values in most of the cases. This implies that increasing the number of LSTM blocks in the hidden layer helps in accommodating more complicated patterns whereas the Dropout layer helps in reducing the overfitting, consequently enhancing the prediction capability of the model.

Figure 10: RMSE of all LSTM models for Apple



Apple: 'LSTM 4' (50 LSTM blocks (hidden layer), 1 Dropout Layer) provides lower RMSE value and better results in most of the cases. Therefore, in this case, the Dropout Layer is playing a significant role in improving the model's predictive ability.

Figure 11: RMSE of all LSTM models for Google



Google: 'LSTM 3' (200 LSTM blocks (hidden layer)) and 'LSTM 4' (50 LSTM blocks (hidden layer), 1 Dropout Layer) provide the lower RMSE value and better results compared to other models. Both 'LSTM 3' and 'LSTM 4' have comparable performances.

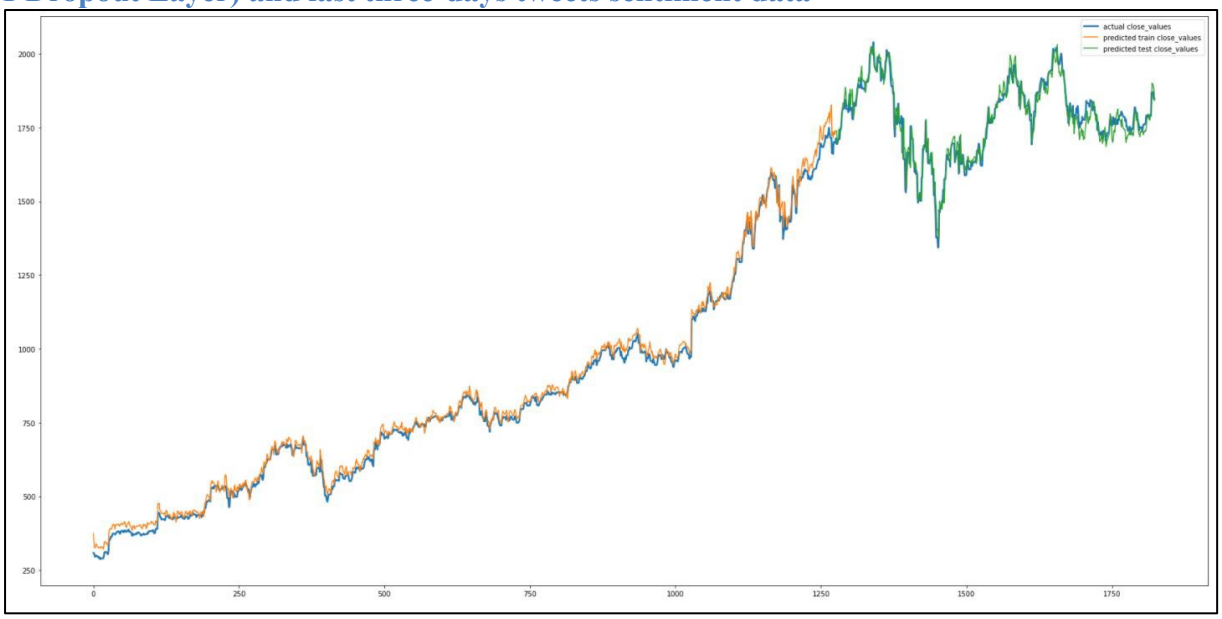
Yadav et al. (2020) state that "the performance of LSTM is highly dependent on the choice of several hyper-parameters, which needs to be chosen very carefully to achieve good results" (Yadav, Jha, & Sharan, 2020, p. 1). It can be deduced that 'LSTM 3' and 'LSTM 4' give better results in

most cases. Therefore, in our experiments, only adding more LSTM blocks in the hidden layer or just adding a dropout layer would have improved the model performance suitably.

Actual Stock Prices Vs Predicted Stock Prices

Fig. 12 shows the shows one of the stock price forecasting results achieved. In this case, the prediction result of LSTM4 is shown, when it is trained on the previous day's stock price and last three-days tweets sentiment value, to predict the next day's stock price.

Figure 12: Amazon stocks close value forecasts using LSTM4 (50 LSTM blocks (hidden layer), 1 Dropout Layer) and last three-days tweets sentiment data



Conclusion and Future Scope

After data filtering, sentimental analysis and feature selection on the entire data of all the five companies, the three companies that showed a correlation among Twitter sentiment data and Stock market data were Amazon, Apple and Google. With the other companies like Tesla and Microsoft, the correlation found was weak and insignificant. This shows that there is a correlation between Twitter data and the stock market in the case of a few companies. The LSTM model is desirable as we have a multivariate time series forecasting problem. LSTM models produced great results especially for Apple Inc with an RMSE value under 5.0 for all the possible variations.

A very desirable outcome in using the algorithm can be observed in Fig. 23, as to how negligible the difference between the actual and predicted values is which provides credibility to the values of the future prediction as portrayed in Fig. 23. The Spearman correlation test also supported the alternative hypothesis, proving how social media plays an important role in influencing stock market values. The models are trained based on features that show a strong correlation of tweet sentiment to stock values which helps in providing better accuracy and less errors. The LSTM 4 model provided very significant results due to its dropout layer which improved the performance and reduced the overfitting. However, our project has certain limitations as stock market prices are oftentimes unpredictable, as of now only Twitter data is considered for analyzing people's sentiment which may not be enough, since people who trade in stocks, also share their opinions on platforms other than that of Twitter. This might add some outliers to our model that affects its capability of analysis. Moreover, our model only relies on Tweets made in the English language since our model is trained based on data consisting of tweets only in the English language. This constrains our ability to check the impact on the stock market from tweets written in languages other than English. Furthermore, since different companies differ in how strong correlations they show, studying causality would add important puzzle pieces to studying the fewer correlations between tweets and the stock market. For future scope, there is room for more analysis and results such as "integrating Reddit posts with Twitter tweets for stock market prediction.". The paper focuses on a limited number of companies, however, there are multiple other companies on which similar analysis can be done which can be considered for future scope. There is also plenty of room in experimenting more with the data such as increasing the time lag between the social media post data and stock market data to see whether the social media sentiments take more time to show impact on the stock market.

Works Cited

Bollen, J., Mao, H., & Zeng, X. (2011, March). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1). doi:https://doi.org/10.1016/j.jocs.2010.12.007

Dickinson, B., & Hu, W. (2015, January). Sentiment Analysis of Investor Opinions on Twitter. *Social Networking, 04*(03), 62-71. doi:10.4236/sn.2015.43008

Doğan, M., Metin, Ö., Tek, E., Yumuşak, S., & Öztoprak, K. (2020). Speculator and Influencer Evaluation in Stock Market by Using Social Media. *2020 IEEE International Conference on Big Data (Big Data)*, (pp. 4559-4566). doi:10.1109/BigData50022.2020.9378170

Hu, D., Jones, C., Zhang, V., & Zhang, X. (2021). The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.3807655

Ko, C.-R., & Chang, H.-T. (2021, March 11). LSTM-based sentiment analysis for stock price forecast. *PeerJ Computer Science*. doi:10.7717/peerj-cs.408

Kordonis, J., Symeonidis, S., & Arampatzis, A. (2016). Stock Price Forecasting via Sentiment Analysis on Twitter. *Proceedings of the 20th Pan-Hellenic Conference on Informatics.*

Mao, H., Counts, S., & Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *Papers 1112.1051, arXiv.org*. Retrieved from https://arxiv.org/abs/1112.1051v1

Mao, Y., Wei, W., Wang, B., & Benyuan, L. (2012). Correlating S&P 500 Stocks with Twitter Data. *HotSocial '12: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research.* Association for Computing Machinery. doi:10.1145/2392622.2392634

Nisar, T. M., & Yeung, M. (2018, June). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science, 4*(2), 101-119. doi:https://doi.org/10.1016/j.jfds.2017.11.002

Pagolu, V. S., Challa, K., Panda, G., & Majhi, B. (2016). Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, (pp. 1345-1350). doi:DOI:10.1109/SCOPES.2016.7955659

Ruiz, E., Hristidis, V., Castillo, C., & Gionis, A. (2012). Correlating Financial Time Series with MicroBlogging Activity. *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012.* Seattle, WA, USA. doi:10.1145/2124295.2124358

Selvin, S., Ravi, V., Gopalakrishnan, E., & Menon, V. K. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).* doi:0.1109/ICACCI.2017.8126078

Shah, D., Isah, H., & Zulkernine, F. (2019, May 27). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studies, 7*(2). doi:https://doi.org/10.3390/ijfs7020026

Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2019). A Comparison of ARIMA and LSTM in Forecasting Time Series. *In M. A. Wani, M. Sayed-Mouchaweh, E. Lughofer, J. Gama, & M. Kantardzic (Eds.), Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018 (pp. 1394-1401). [8614252] Institute of Electrical and Electronics Engineers.* doi:https://doi.org/10.1109/ICMLA.2018.00227

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance, Forthcoming, 62*(3). doi:http://dx.doi.org/10.2139/ssrn.685145

Xia, Y., Liu, Y., & Chen, Z. (2013). Support Vector Regression for prediction of stock trend. *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII).* doi:10.1109/ICIII.2013.6703098

Yadav, A., Jha, C., & Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. *Procedia Computer Science, 167*, 2091-2100. doi:https://doi.org/10.1016/j.procs.2020.03.257

Appendix A

Column definitions of all the datasets used

Table A1: Column definitions for top companies' dataset (Company.csv)

ticker_symbol	Unique company symbol that indicates a particular top company
company_name	Name of the top company indicated by its ticker symbol.

Table A2: Column definition for daily values of top companies' dataset (CompanyValues.csv)

ticker_symbol	Unique company symbol that indicates a particular top company

day_date	Present-day date of which there is corresponding open value, close value, high value and low value of the stock.
close_value	The attribute represents the last price at which an individual stock was traded during the regular trading day.
Volume	The attribute represents the total number of shares that are actually traded (bought and sold) during the regular trading day.
Open_value	The attribute represents the first price at which an individual stock was traded upon the opening of the regular trading day.
High_value	The attribute represents the highest price at which a stock was traded during a regular trading day.
Low_value	The attribute represents the lowest price at which a stock was traded during a regular trading day.

Table A3: Column definitions for Company Tweet dataset (Company_Tweet.csv)

tweet_id	Unique identification number of an individual tweet given by Twitter that identifies the particular tweet uniquely.
ticker_symbol	Unique company symbol that indicates a particular top company.

Table A4: Column definitions for Tweet dataset (Tweet.csv)

tweet_id	Unique identification number of an individual tweet given by Twitter that identifies the particular tweet uniquely.
writer	The attribute represents the account name of the author who tweeted the particular tweet.
Post date	This attribute represents the time at which the tweet was created in the form of seconds.
body	The attribute represents the actual text of the tweet that was tweeted by the writer.
comment_num	The attribute represents the total number of comments that were left on that particular tweet.

retweet_num	The attribute represents the total number of times a particular tweet was reposted (retweeted)
like_num	The attribute represents the total number of likes on a particular tweet.

Appendix B

Spearman Correlation Coefficients

Figure B1: Spearman Correlation Matrix for Amazon

	comment_num	retweet_num	like_num	polarity	subjectivity	neg	neu	pos	compound
volume	0.360642	0.328944	0.147097	-0.274593	-0.328254	0.380743	0.045691	-0.158885	0.031200
open_value	0.336729	-0.219515	0.434143	-0.206486	0.539574	-0.425629	-0.407147	0.543308	0.163202
close_value	0.334268	-0.219234	0.434526	-0.200720	0.541872	-0.430630	-0.406207	0.543957	0.162068
low_value	0.327258	-0.222720	0.428826	-0.196263	0.542432	-0.433097	-0.403549	0.542215	0.161114
high_value	0.341998	-0.216730	0.438726	-0.209847	0.539011	-0.423356	-0.408207	0.543734	0.163204

Appendix C

RMSE values of all the models

Figure C1: RMSE values obtained for Apple

	RMSE											
	Apple											
	LSTM1		LSTM2		LSTM3		LSTM4		LSTM5		LSTM6	
Values	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days
open_value	3.39	3.55	3.81	3.93	4.02	3.38	3.42	3.23	3.97	4.23	3.51	2.94
close_value	3.73	3.86	3.76	3.5	4.3	4.77	3.29	3.34	3.67	3.46	4.66	4.69
low_value	4.81	3.27	4.82	2.83	3.05	2.92	2.99	3.27	3.74	4.63	3.45	3.45
high_value	3.46	3.21	2.78	3.15	4.6	3.67	3.05	2.95	3.93	3.16	3.34	5.58

Figure C2: RMSE values obtained for Amazon

	RMSE											
	Amazon											
	LSTM1		LSTM2		LSTM3		LSTM4		LSTM5		LSTM6	
Values	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days
open_value	38.02	42.15	31.16	35.92	32.38	34.84	33.04	39.94	35.27	32.3	30.15	31.39
close_value	36.78	34.52	29.58	33.93	33.34	30.56	36.45	36.83	43.87	31.26	28.35	32.54
low_value	28.92	42.3	33.35	36.26	37.86	33.72	29.92	33.66	28.56	29.54	26.83	27.7
high_value	32.59	30.74	27.52	26.31	30.18	27.32	29.15	38.25	26.73	34.69	25.44	39.85

Figure C3: RMSE values obtained for Google

	RMSE											
	Googl											
	LSTM1		LSTM2		LSTM3		LSTM4		LSTM5		LSTM6	
Values	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days	same day	last 3 days
open_value	24.6	25.81	28.2	33.43	16.47	17.68	17.85	18.21	27.14	25.49	18.34	23.73
close_value	25.11	17.04	16.2	15.95	16.55	16.31	17.93	16.87	23.32	16.21	21.1	16.87
low_value	21.33	24.94	15.15	15.1	16.07	14.95	15.62	16.92	18.74	30.28	14.28	39.55
high_value	23.14	25.97	14.16	14.76	15.19	14.92	17.03	16.97	20.33	14.9	26.66	43.76