IS 733 Data Mining

Term Project On

Stock Market Prediction Using Time Series Analysis.

Using Twitter Sentiments (2010-2020)

Priyadarshini Arcot

Under the Guidance of

Professor Faisal Quader

Introduction

Stock price prediction is a complex and difficult undertaking due to the vast number of factors involved, such as economic circumstances, political events, and others that may affect the stock price directly or indirectly. As a result, evaluating a single factor's influence on future pricing and trends is challenging. While more people are spending time online expressing their opinions, social media platforms' power and importance have risen to unprecedented heights in recent years. Researchers have taken an interest in Twitter, particularly those focusing on assessing public thoughts and sentiments in the stock market area, because it is a rich source of real-time information, including societal and personal opinions. According to our research, Twitter attitudes have an impact on stock price movement and forecasting for some of the largest corporations. The quantity of 'positivity,' 'negativity,' and subjectivity in tweets was also found to be substantially connected with stock price fluctuations, thereby contributing significantly to the predictive stock price model. We discovered that the deep learning model of the Long-Short Term Memory neural network (LSTM) delivers state-of-the-art results to anticipate stock prices using Twitter sentiments and historical stock prices data since stock market data is a time-series data with continuous information. According to our findings, the predictive model for Apple Inc., Amazon, and Google had a strong association with Twitter Sentiments.

Keywords: Sentiment Analysis, Twitter, LSTM, Stock Market Prediction, Apple, Google, Amazon, ARIMA.

Analysis of Twitter Sentiments' effects on the Stock Market for the business year [2010-2020]

The stock market is a highly volatile environment in which social media's influence on stock market values is regularly felt. Twitter is a real-time microblogging service that allows users to follow and comment on the thoughts and opinions of other users. The opinions of the collective populations can then be aggregated using Twitter-based models. They can be used to forecast future trends while also providing vital information about individual behavior. In order to analyze public attitudes, sentiment analysis infers the author's point of view from a piece of text and categorizes the polarity of given tweets into positive and negative tweets. Stock market forecasting using public emotion conveyed on Twitter has been a fascinating area of study. The relationship between Twitter emotions and stock price swings has been extensively researched. Previous research has demonstrated that sentiment analysis on Twitter can be used to predict stock prices. Positive news and tweets about a company on social media would, in theory, inspire individuals to invest in the firm's stock, resulting in an increase in the stock price of that company [1].

Stock prices are time-series data that may be utilized in deep learning models to derive patterns and identify trends. Perhaps these patterns are too complex for humans or other traditional computer systems to comprehend. The Recurrent Neural Network (RNN) is a type of artificial neural network that can be used to solve time series problems. After repeated recursions, the initial RNN's memory will lose its effect due to the number of complex layers. As a result, the LSTM neural network concept is introduced. It is a unique and useful RNN model that can memorize long-term or short-term values while allowing the neural network to keep only the information it requires [2].

Research Goal

The goal of this article is to anticipate the stock market prices of leading businesses such as Apple, Amazon, Google, Microsoft, and Tesla using sentimental analysis of public sentiments on Twitter. We chose Apple since it has the greatest sample size and the analysis that will be performed on the supplied data frame can be applied to other businesses. To determine the polarity (positive/negative/neutral) and subjectivity of tweets, we used emotional analysis. According to the paper's hypothesis, there is a link between the positivity or negativity of tweets and stock price movement. After successfully establishing the correlation that supports our theory, we used multiple iterations of the Long Short-Term Memory model to forecast the next day's stock market values, taking into account the positivity and negativity features of the feelings. In order to predict the next day's stock prices, the LSTM models use the 'positivity' value of the company tweet, the 'negativity' of the tweet, the open day's stock value (open value), the close day's stock value (close value), the highest stock value for the day (high value), the lowest stock value for the day (low value), and the subjectivity of the tweet. We forecasted the low value, high value, open value, and close value for the next day in different ways. The research concludes with an intriguing set of results comparing real and anticipated stock market prices, in which our LSTM model outperformed **Apple Inc.** by a factor of less than 5.0. This study has a lot of potential for businesses and stock market investors when it comes to making investing decisions.

Literature Review

Twitter is known for transmitting product quality information and reviews in real-time, allowing for input. Many studies employ statistical modeling or machine learning methods, such as Support Vector Machine (SVM), to forecast future stock fluctuations based on existing data. Computing power has increased dramatically in recent years as GPU technology has advanced [5]. With the steady expansion of the stock market over the last few decades, an increasing number of people have done stock price projection research. They attempt to evaluate and forecast stock market movements and price changes.

Bollen, Mao, and Zeng are the most well-known authors in this field. They looked into whether the aggregate public mood on Twitter is related to the Dow Jones Industrial Average Index (DJIA). For their prediction, they used a Fuzzy neural network. [6]. Ruiz et al. investigated the topic of linking Twitter microblogging activity with changes in stock prices and trading volumes using time-constrained graphs [7]. Pagolu et al. (2016) and Kordonis, Symeonidis, and Arampatzis (2016) attempted to anticipate stock market moves using Twitter sentiment analysis. Both studies

discovered a considerable link between Twitter sentiment and stock price fluctuations. Brian et al. recently used the Pearson correlation coefficient for equities to study the link of public sentiment with stock increases and losses. Each of these studies demonstrated that Twitter is a rich resource and an effective tool for doing research and making predictions. In this research, we used multiple forms of the Long Short-Term Memory model to forecast the next day's stock market values using the positivity and negative features of the sentiments from tweets.
LSTM models have the ability to store information over time.
"They have a memory capacity, in other words. Keep in mind that LSTM refers to the Long Short-Term Memory Model [S. Loukas, Jul 10, 2020]

Data Set Used For the Research

We used the datasets uploaded by Omer Metin and Mustafa Dogan to 'Kaggle' with the labels 'Tweets about the Top Companies from 2015 to 2020' and 'Values of Top NASDAQ Companies from 2010 to 2020'. The research paper 'Speculator and Influencer Evaluation in Stock Market by Using Social Media' was published in the 6th Special Session on Intelligent Data Mining track of the 2020 IEEE International Conference on Big Data. 'Company.csv,' 'Company Tweet.csv,' 'Tweet.csv,' and 'CompanyValues.csv' are the datasets. Only Apple, Amazon, Microsoft, Tesla, and Google are represented in these datasets.

'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.

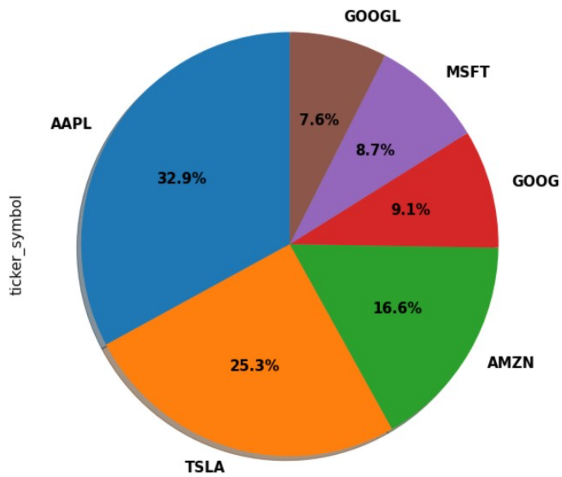
	ticker_symbol	company_name
0	AAPL	apple
1	GOOG	Google Inc
2	GOOGL	Google Inc
3	AMZN	Amazon.com
4	TSLA	Tesla Inc
5	MSFT	Microsoft

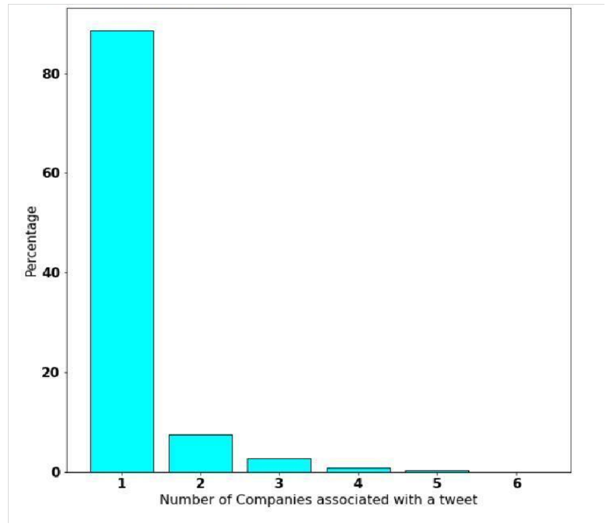
	tweet_id	ticker_symbol
0	550803612197457920	AAPL
1	550803610825928706	AAPL
2	550803225113157632	AAPL
3	550802957370159104	AAPL
4	550802855129382912	AAPL
...	...	...
4336440	1212158772015034369	TSLA
4336441	1212159099632267268	TSLA
4336442	1212159184931717120	TSLA
4336443	1212159838882533376	TSLA
4336444	1212160015332728833	TSLA

4336445 rows × 2 columns

Figure A

Figure B





'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.

Attributes associated with each tweet

Feature	Description
tweet_id	Unique tweet id of a tweet
writer	Username of the author
post_date	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

The file 'CompanyValues.csv' contains stock market data from several firms. The value of the first traded stock, the last traded stock, the lowest price at which the stock was traded, and the highest price at which the stock was exchanged are all included for each day.

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	Lowest price at which the stock is traded
high_value	Highest price at which the stock is traded

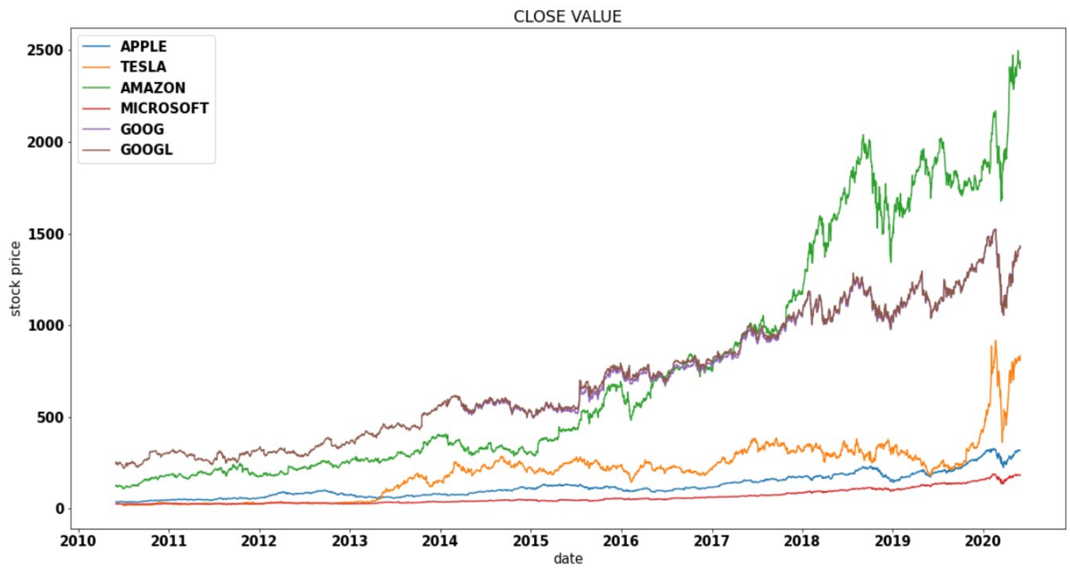
	ticker_symbol	day_date	close_value	volume	open_value	high_value	low_value
0	AAPL	2020-05-29	317.94	38399530	319.25	321.15	316.4700
1	AAPL	2020-05-28	318.25	33449100	316.77	323.44	315.6300
2	AAPL	2020-05-27	318.11	28236270	316.14	318.71	313.0900
3	AAPL	2020-05-26	316.73	31380450	323.50	324.24	316.5000
4	AAPL	2020-05-22	318.89	20450750	315.77	319.23	315.3500
...	...	...	...	...	...	...	...
17523	TSLA	2019-12-21	405.59	14785210	410.29	413.00	400.1850
17524	TSLA	2019-12-22	405.59	14785210	410.29	413.00	400.1850
17525	TSLA	2019-12-25	425.25	8054720	418.36	425.47	412.6875
17526	TSLA	2019-12-28	430.38	9956827	435.00	435.31	426.1100
17527	TSLA	2019-12-29	430.38	9956827	435.00	435.31	426.1100

17528 rows × 7 columns

<u>Different Values of Apple.</u>

The open, close, low, and high stock values of Amazon are shown in the graph below. As can be seen, there is little variation in these numbers. Comparable cases of stock valuations that are similar to each other have been observed in other companies.

The graph below depicts Apple, Tesla, Amazon, Microsoft, and Google's open values over the last decade. With the help of this graph, we can see the growth patterns and the level of increase or consistency in stock values.



‘Company.csv’, ‘Company_Tweet.csv’, and ‘Tweet.csv’ consist of data from 2015 to 2020. ‘CompanyValues.csv’ consists of data from 2010 to 2020.

Methodology Used for Obtaining Results

Preparation of Data

We focused on the examination of aggregated tweets rather than individual tweets to keep the scope of our research limited. For each of the six companies, we first prepared the following sorts of datasets:

Note: We have done an analysis of all the six companies and then obtained the desired output for the model on one company [Apple Inc] based on the hypothesis results.

- Last 3 Days' aggregated tweets dataset: contained aggregated value of tweets and their attributes for the last 3 days

We combined all of the tweets from the necessary time period, removed the author from the dataset, and totaled the number of comments, retweets, and likes. Second, we cleaned and pre-processed each of the above-mentioned datasets, using TextBlob to recover polarity and subjectivity values, and Vader Sentiment Intensity Analyzer to retrieve positivity, negativity, neutrality, and compound values (SIA). These sentiment intensity values were appended to the original datasets.

To undertake hypothesis testing and predictive modeling, we combined the sentiment values datasets with the stock market dataset.

Dropping of columns to clean the data and making the data frame:

```
[ ] # converting the 'updated_dates' column into datetime
    apple_tweets_dff['updated_dates'] = pd.to_datetime(apple_tweets_dff['updated_dates'].dt.date)
    display(apple_tweets_dff)
```

	body	comment_num	retweet_num	like_num	updated_dates
0	lx21 made $10,008 on $AAPL -Check it out! htt...	0	0	1	2015-01-01
1	Insanity of today weirdo massive selling. $aap...	0	0	0	2015-01-01
2	Swing Trading: Up To 8.91% Return In 14 Days h...	0	0	1	2015-01-01
3	Swing Trading: Up To 8.91% Return In 14 Days h...	0	0	1	2015-01-01
4	Swing Trading: Up To 8.91% Return In 14 Days h...	0	0	1	2015-01-01
...	...	...	...	...	...
1425008	Imagine calling your broker-dealer and wanting...	1	0	1	2020-01-01
1425009	$AAPL yearly~ Heck of a year.. Jan. 2, 1999~ar...	0	0	1	2020-01-01
1425010	That $SPY $SPX puuump in the last hour was the...	1	0	6	2020-01-01
1425011	I don't discriminate. I own both $aapl and $ms...	1	0	1	2020-01-01
1425012	$AAPL #patent 10,522,475 Vertical interconnect...	0	0	0	2020-01-01

1425013 rows × 5 columns

Prepared Data Set:

```
[ ] # viewing the general info of the prepared dataset
    apple_tweets_dff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1425013 entries, 0 to 1425012
Data columns (total 5 columns):
 #   Column         Non-Null Count    Dtype
---  ------         --------------    -----
 0   body           1425013 non-null  object
 1   comment_num    1425013 non-null  int64
 2   retweet_num    1425013 non-null  int64
 3   like_num       1425013 non-null  int64
 4   updated_dates  1425013 non-null  datetime64[ns]
dtypes: datetime64[ns](1), int64(3), object(1)
memory usage: 54.4+ MB
```

Example: How we Cleaned the Data for the Tweets:

```
# viewing the cleaned tweets
from tqdm import tqdm
tqdm.pandas()

display(comb_tweets_apple['clean_text'].head())
comb_tweets_apple['clean_text'] = comb_tweets_apple['clean_text'].progress_apply(lambda x: clean_tweets(x))
display(comb_tweets_apple['clean_text'].head())
```

```
0    zacks' bull of the day: apple http://seekingal...
1    free 5€ in account balance for first 100.000 m...
2    free 5€ in account balance for first 100.000 m...
3    apple: does the party end in 2015? http://seek...
4    the ever-changing world of apple http://seekin...
Name: clean_text, dtype: object
100%|██████████| 1825/1825 [04:24<00:00,  6.90it/s]
0    zacks bull of the day apple have a great weeke...
1    free in account balance for first members lnkd...
2    free in account balance for first members lnkd...
3    apple does the party end in long with successf...
4    the ever changing world of apple what dean kar...
Name: clean_text, dtype: object
```

Testing Hypotheses

The Hypothesis testing was done on the final merged dataset, which included sentiment intensity and stock market values. To determine which correlation coefficient should be utilized, we first plotted the distribution graphs of these variables. The distribution plots were discovered to be non-normally distributed, implying that the 'Spearman correlation coefficient should be employed instead of the 'Pearson correlation coefficient. Only three firms' stock prices demonstrated a high link with Twitter sentiment scores, according to our findings. Apple (APPL), Amazon (AMZN), and Google Inc. were the corporations in question (GOOGL). There was a considerable association between subjectivity, polarity, negativity, and open value, close value, low value, and high value' for these organizations.

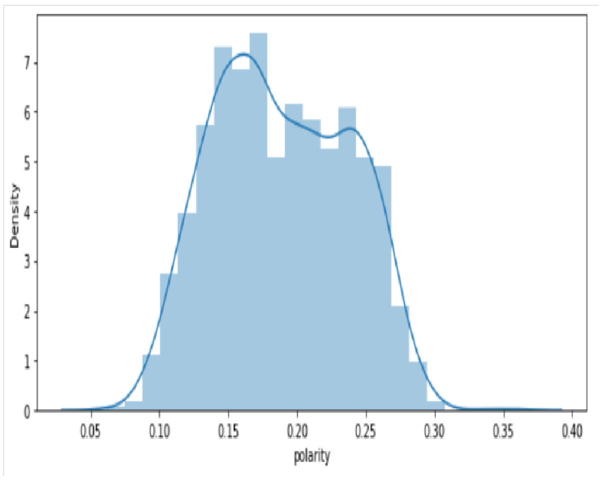
1. Filtering the dataset to get the stock market data of apple (AAPL)
2. Checking info of dataset
3. Converting the 'day_date' column to DateTime format
4. Prepare data frame for last 3 days
5. Merge the data frames
6. Plotting distribution plot of all the merged data frames

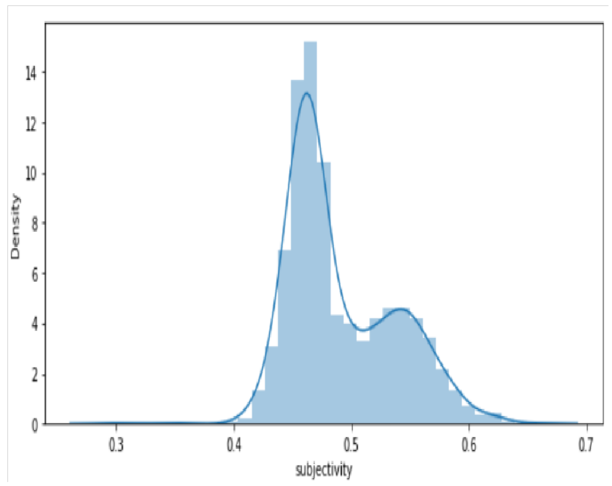
This led to the following Null and Alternative Hypothesis:

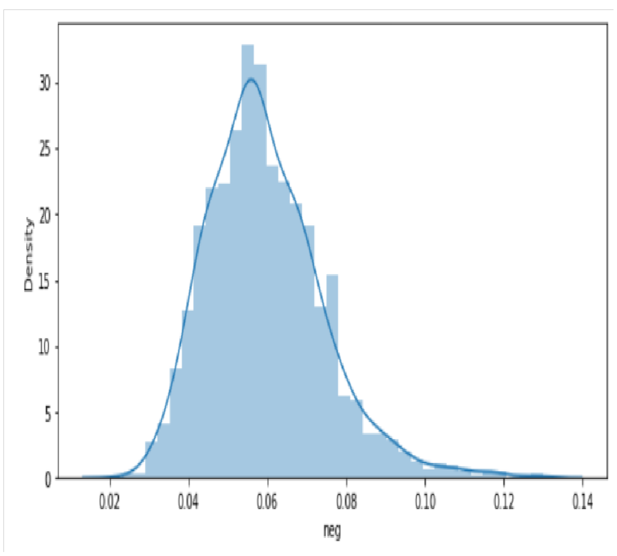
Null hypothesis: There is no relationship between the ‘subjectivity’, ‘polarity’, ‘negativity’ and ‘open_value’, ‘close_value’, ‘low_value’ and ‘high_value’.

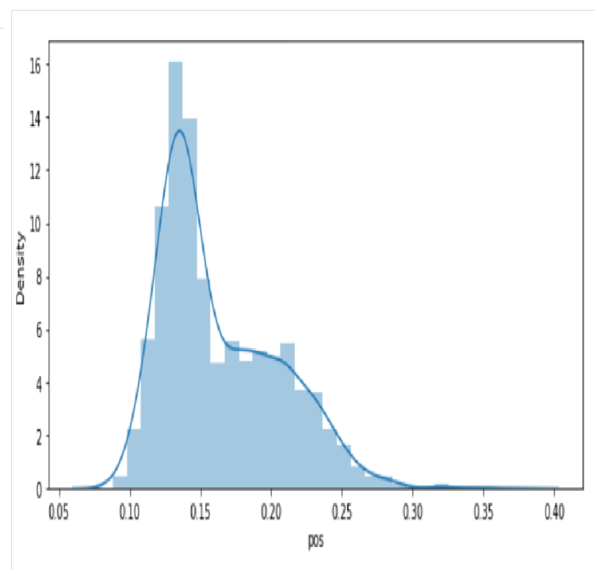
Alternative hypothesis: There's a significant correlation between the ‘subjectivity’, ‘polarity’, ‘negativity’ and ‘open_value’, ‘close_value’, ‘low_value’ and ‘high_value’

Below figures show the distribution of the graph for ‘subjectivity’, ‘polarity’, ‘negativity’. ‘positivity’.









Finding spearman correlation values and p-values and plotting correlation plot

	comment_num	retweet_num	like_num	polarity	subjectivity	neg	neu	pos	compound	volume	open_value	close_value	low_value	high_value
comment_num	1.000000	0.518826	0.863511	-0.355460	0.093500	0.196106	-0.267793	0.188960	0.088366	0.360642	0.336729	0.334268	0.327258	0.341998
retweet_num	0.518826	1.000000	0.573318	-0.066800	-0.219700	0.296700	0.085611	-0.177079	0.023606	0.328944	-0.219515	-0.219234	-0.222720	-0.216730
like_num	0.863511	0.573318	1.000000	-0.180183	0.262745	0.074592	-0.285628	0.241261	0.107217	0.147097	0.434143	0.434526	0.428826	0.438726
polarity	-0.355460	-0.066800	-0.180183	1.000000	0.296995	-0.275235	-0.085915	0.083324	-0.121522	-0.274593	-0.206486	-0.200720	-0.196263	-0.209847
subjectivity	0.093500	-0.219700	0.262745	0.296995	1.000000	-0.489511	-0.703762	0.739710	0.075982	-0.328254	0.539574	0.541872	0.542432	0.539011
neg	0.196106	0.296700	0.074592	-0.275235	-0.489511	1.000000	0.296587	-0.603099	-0.162976	0.381294	-0.424630	-0.429647	-0.432128	-0.422344
neu	-0.267793	0.085611	-0.285628	-0.085915	-0.703762	0.296587	1.000000	-0.910447	-0.068473	0.044875	-0.407072	-0.406156	-0.403474	-0.408146
pos	0.188960	-0.177079	0.241261	0.083324	0.739710	-0.603099	-0.910447	1.000000	0.187524	-0.158295	0.543669	0.544337	0.542573	0.544121
compound	0.088366	0.023606	0.107217	-0.121522	0.075982	-0.162976	-0.068473	0.187524	1.000000	0.031198	0.163202	0.162067	0.161114	0.163203
volume	0.360642	0.328944	0.147097	-0.274593	-0.328254	0.381294	0.044875	-0.158295	0.031198	1.000000	-0.351651	-0.357173	-0.366998	-0.344259
open_value	0.336729	-0.219515	0.434143	-0.206486	0.539574	-0.424630	-0.407072	0.543669	0.163202	-0.351651	1.000000	0.998533	0.999173	0.999422
close_value	0.334268	-0.219234	0.434526	-0.200720	0.541872	-0.429647	-0.406156	0.544337	0.162067	-0.357173	0.998533	1.000000	0.999372	0.999234
low_value	0.327258	-0.222720	0.428826	-0.196263	0.542432	-0.432128	-0.403474	0.542573	0.161114	-0.366998	0.999173	0.999372	1.000000	0.999074
high_value	0.341998	-0.216730	0.438726	-0.209847	0.539011	-0.422344	-0.408146	0.544121	0.163203	-0.344259	0.999422	0.999234	0.999074	1.000000

Showing significant Spearman Correlation Coefficient for Apple

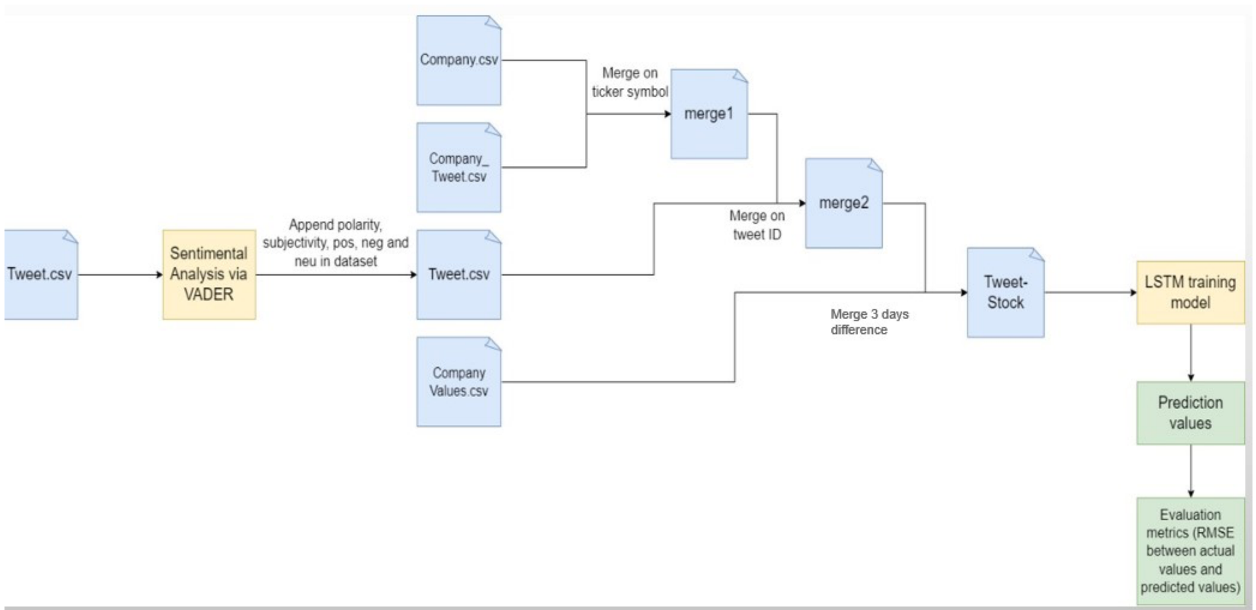
Values	Apple					
	same day			last 3 days		
	positivity	negativity	subjectivity	positivity	negativity	subjectivity
open_value	0.51	-0.33	0.51	0.54	-0.42	0.53
close_value	0.51	-0.34	0.51	0.54	-0.43	0.54
low_value	0.5	-0.34	0.51	0.54	-0.43	0.54
high_value	0.51	-0.33	0.51	0.54	-0.42	0.53

Here, we obtained the ‘p-values’ for these correlation coefficients to understand the significance of correlation. The p-values were found less than the alpha threshold of 0.05 in all the cases and hence it was proved that the correlations were statistically significant. Therefore, the Null hypothesis was rejected.

Methodology Flow Chart

Steps are included in the methodology for the stock price prediction:

1. Dataset Preparation: Last three days tweets
2. Sentiment Analysis & Feature Engineering: Positivity, negativity, neutrality, polarity, subjectivity, compound
3. Correlation: Spearman Correlation /Tweets sentiments & Stock prices
4. Stock price prediction: Multivariate time series forecasting



Model Training

Our topic became a multivariate time series forecasting problem because we wanted to anticipate the next day's stock price using the last three days' stock price combined with Twitter sentiment values. In other words, we had multiple time-dependent variables, i.e., the target variable was reliant on its previous values as well as some other variables. This meant that neither a regressor model nor a univariate time series forecasting model like ARIMA or SARIMAX could be used. ARIMA stands for the autoregressive integrated moving average. It is a statistical analysis model that uses time-series data to better understand the data set or anticipate future trends. If a statistical model predicts future values based on past values, it is called autoregressive.

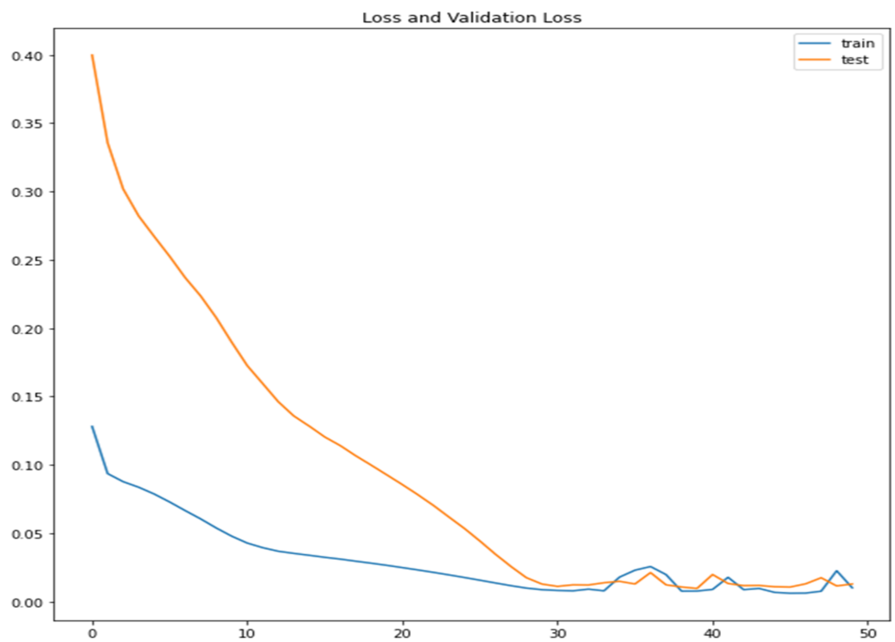
"RNN is commonly employed among many ANN to deal with time series problems, and it is also one of the most suitable techniques for dynamic time series forecasting," according to a literature study. However, RNN is susceptible to the vanishing gradient problem. With many recursions, RNN will face the problem of gradient disappearance and explosion. The LSTM neural network was created to improve the RNN's performance in artificial intelligence applications. As a result, we used Long-Short Term Memory (LSTM) models to do time-series forecasting [7]. The dataset for our model training was divided into 70 percent training data and 30% testing data, respectively. We first trained a baseline LSTM model for predicting high values with 50 neurons in the LSTM layer. Then we experimented with several LSTM model settings and compared the
RMSE values achieved in each model. For model training, we used the following configurations:

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

(2) LSTM 2: 100 neurons (hidden layer)
(3) LSTM 4: 50 neurons (hidden layer), 1 Dropout Layer
(4) LSTM 5: 100 neurons (hidden layer), 1 Dropout Layer

```
# Defining the LSTM model
# for high_value
model_3d_apple_hv = Sequential()
model_3d_apple_hv.add(LSTM(50, input_shape=(x_train_3d_hv.shape[1], x_train_3d_hv.shape[2])))
model_3d_apple_hv.add(Dense(1))
model_3d_apple_hv.compile(loss='mae', optimizer='adam')
model_3d_apple_hv.build(input_shape=(x_train_3d_hv.shape[1], x_train_3d_hv.shape[2]))

# viewing the summary of the model
print('\nsequential model for Apple High values:\n')
model_3d_apple_hv.summary()
```

```
sequential model for Apple High values:

Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 50)                11000

 dense (Dense)               (None, 1)                 51

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



The above figure shows the fitting of the model.

Results

In this study, we use Apple Inc. Data set to forecast models using LSTM were established based on different input neurons and dropout. The forecast results are evaluated with the RMSE of the test data which is calculated as shown below figure:

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

<u>Comparing the effects of sentiments of immediate tweets vs the effects of sentiments of tweets accumulated over time</u>

We aggregated the last three days' tweets including current days to yield the tweet sentiment value in order to evaluate the impact of emotional information latent in the tweets data on stock price fluctuations and to increase the accuracy of the forecast. The below figure shows the comparison between the RMSE values obtained from one of the LSTM variations which were used to predict the stock market prices(open value, close value, high value, low value) providing the previous day's tweet sentiment value and the last three day's tweet sentiment value for a company like Apple.

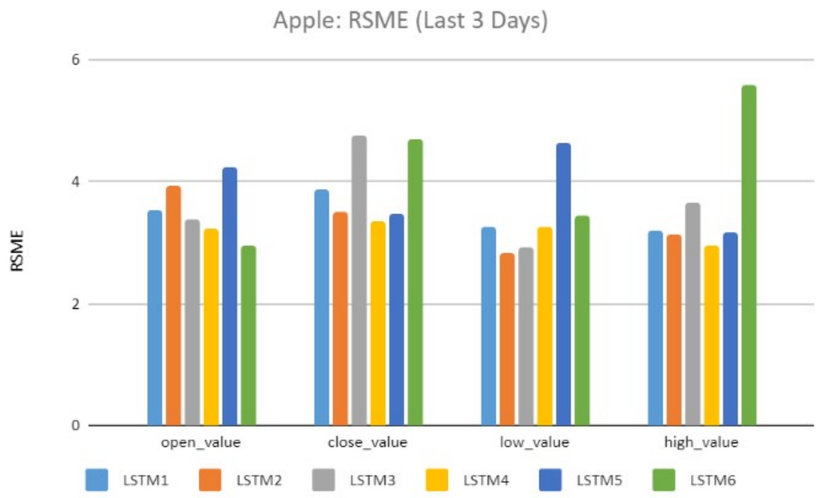
1) **Apple Inc**: In the case of Apple, the LSTM model considering the aggregated value of tweets for the last 3 days is performing very well in aggregated tweets with the RMSE value below 5.0 for both the models.

<u>Investigate how accurately we can predict the next day's stock prices of a company with the variations of LSTMs</u>

The experimental examination of LSTM variations with varying numbers of neurons and the addition of dropouts yielded some interesting results that can help us choose a good model to test how well we can predict the stock values of all firms the next day. As a result, selecting the most appropriate and effective model.

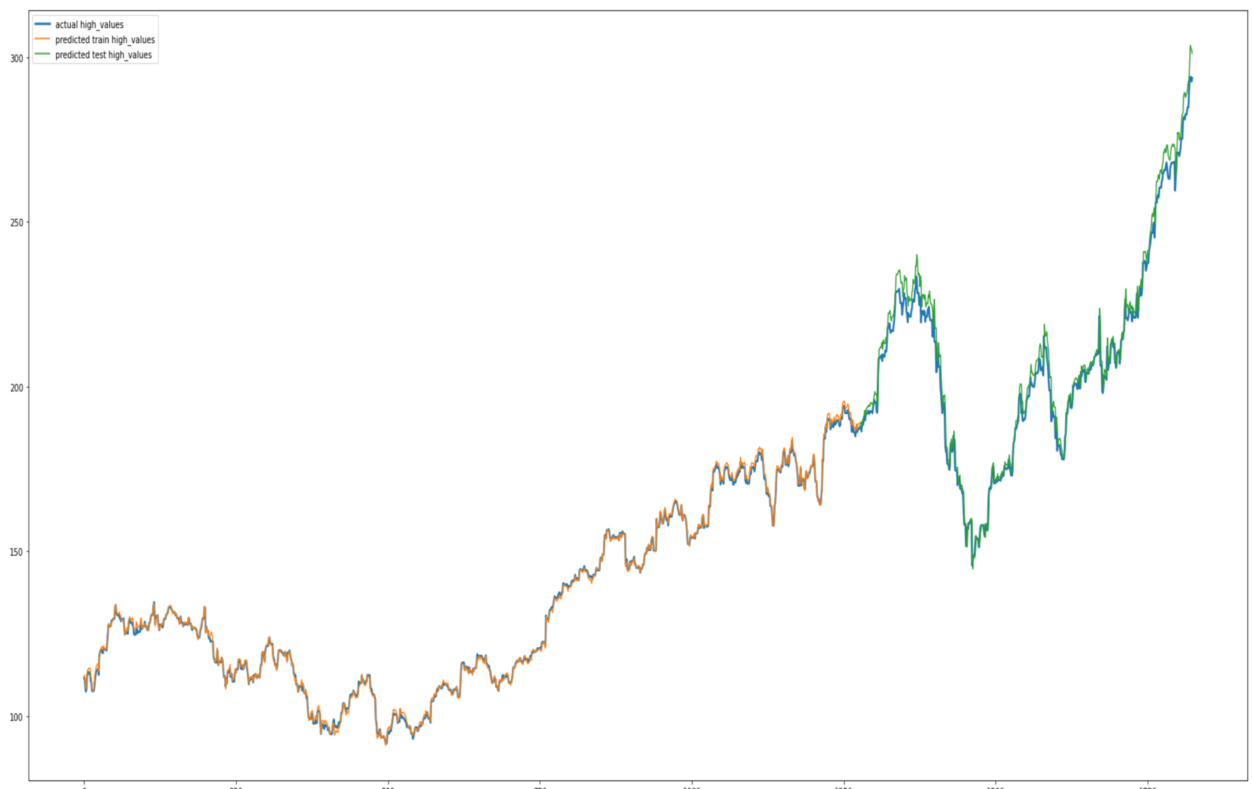
The below figure shows the RMSE values for different variations of LSTM for Apple Inc. For Apple, the overall 'LSTM 4' model having 50 input neurons and 1 Dropout Layer provides a lower RMSE value and better results compared to any other models.

As a result, the analysis of the results assists us in selecting the best performing model from a variety of LSTM models produced through trials and hyperparameter tuning, especially in the case of the majority of organizations.



ACTUAL STOCK PRICES VS PREDICTED STOCK PRICES

The below figure shows the sample reference of the predicted stock values(close value of the stock)with reference to actual stock values(close value of the stock) to the graph of predicted values and actual values.



Conclusion and Future Scope

After data cleaning, data filtering, sentimental analysis, and feature selection on the entire data of all the five companies, the three companies that showed a correlation between Twitter data and Stock market data were Amazon, Apple, and Google. With the other companies like Tesla and Microsoft, it showed that the correlation 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 time series-based sequence prediction problem in which our model predicts the stock values of the next days based on previous stock values as well as current Twitter data.

Long Short-Term Memory (LSTM) model proved to be a very efficient algorithm prediction of stock values, especially for Apple Inc with the RMSE value of 5.0.

Only Twitter data is presently used in our project to analyze people's moods, which may not be sufficient because people who trade stocks also publish their ideas on platforms other than Twitter. For a comprehensive public opinion collection, recent news can also be included. Furthermore, because various firms and companies in different industries appear to have varying levels of connection, examining causality would contribute vital jigsaw pieces to the study of tweet-stock market correlation.

Likewise, the data aggregation and processing work on it. We may also create a longer time lag between social media post data and stock market data to observe if the influence takes longer to show on the stock market. This will require more time and space for processing as the time lag increases.

References

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.

Retrieved From IEEE doi: https://doi.org/10.1145/2392622.2392634

Anita Yadav C K Jha AditiSharan (2020) Optimizing LSTM for time series prediction in the Indian stock market

Retrieved from Article: https://www.sciencedirect.com/science/article/pii/S1877050920307237#!

Padmanayana, Varsha, & Bhavya K. (2021). Stock Market Prediction Using Twitter Sentiment Analysis. International Journal of Scientific Research in Science and Technology, 265–270.

Retrieved from research gate: https://doi.org/10.32628/cseit217475

Wang, Y., Zhu, S., & Li, C. (2019). Research on multi step time series prediction based on LSTM. 2019

3rd International Conference on Electronic Information Technology and Computer Engineering (EITC).

Retrieved from IEEE: https://doi.org/10.1109/eitce47263.2019.9095044

Bollen, J., Mao, H., & Zeng, X. (2011, March). Twitter mood predicts the stock market. Journal of Computational Science, 2(1).

Retrieved from 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.

Retrieved from IEEE: doi: https://doi.org/10.4236/sn.2015.43008

Ko, C.-R., & Chang, H.-T. (2021, March 11). LSTM-based sentiment analysis for a stock price forecast. PeerJ Computer Science.

Retrieved from IEEE: doi: https://doi.org/10.7717/peerj-cs.408

Time-Series Forecasting: Predicting Stock Prices Using An LSTM Model

Retrieved From: https://towardsdatascience.com/lstm-time-series-forecasting-predicting-stock-prices-using-an-lstm-model-6223e9644a2f