Indian Institute of Technology, Kharagpur

AI60003: ARTIFICIAL INTELLIGENCE FOR ECONOMICS

Assignment 1: Term Project Report

Project Topic

Stock Price Movement Prediction using Sentiment Analysis

Abhinav Bohra, 18CS30049

November 14, 2021

1. Introduction

From the last twenty years, the application of Internet based technologies had brought a significant impact on the Indian stock market. Use of the Internet has eliminated the barriers of brokers and geographical location because now investors can buy and sell their shares by accessing the stock market status from anywhere at any time. Before investing money, it is very important for investors to predict the stock market. In today's digital world Internet based technologies such as Cloud Computing, Big Data analytics, and Sentiment analysis have changed the way we do business.

Sentiment analysis or opinion mining makes use of text mining, natural language processing (NLP), to identify and extract the subjective content by analyzing user's opinion, evaluation, sentiments, attitudes, and emotions.

Social media plays very important role in changing a company's perception towards customers. Every company is trying to create a positive image through social media, Companies have even started customer care services using their twitter handles. Even our government uses twitter for solving common man problems. In this project, I intend to explore and answer the question ***“Can social media sentiments of a company be used to predict company’s stock price movement?”***

2. Problem Statement

The objective is to **find statistical correlation between social media sentiment and stock price movement of companies.**

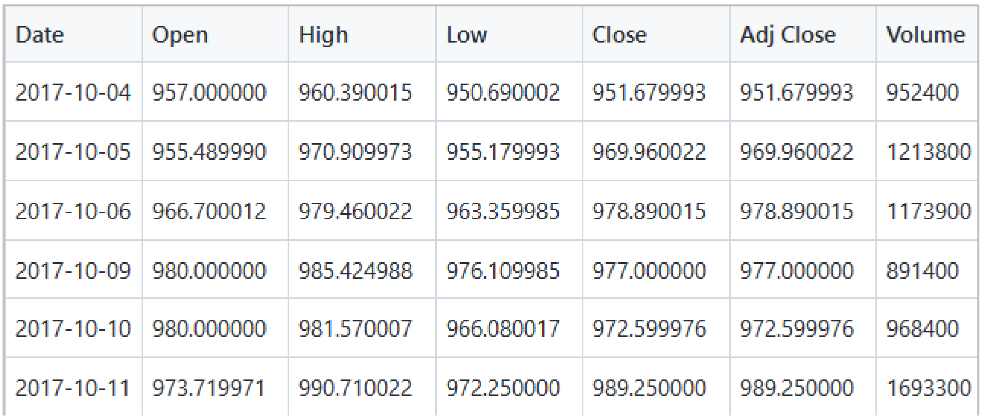
First, I predict the stock prices using the following 6 features - Open Stock price, Close Stock price, Low price, High price, Volume and Adj Close Price. Next, I performed sentiment analysis on tweets for that company posted on Twitter and use the social media sentiment score as an additional feature to predict the stock prices.

The difference in accuracy in both the models shall help us understand the correlation between social media sentiment and stock prices of companies. If the model with additional sentiment features performs better than the baseline model, then there exists a positive co-relation between social media sentiment and stock prices of a company, otherwise no such co-relation exists.

3. Dataset

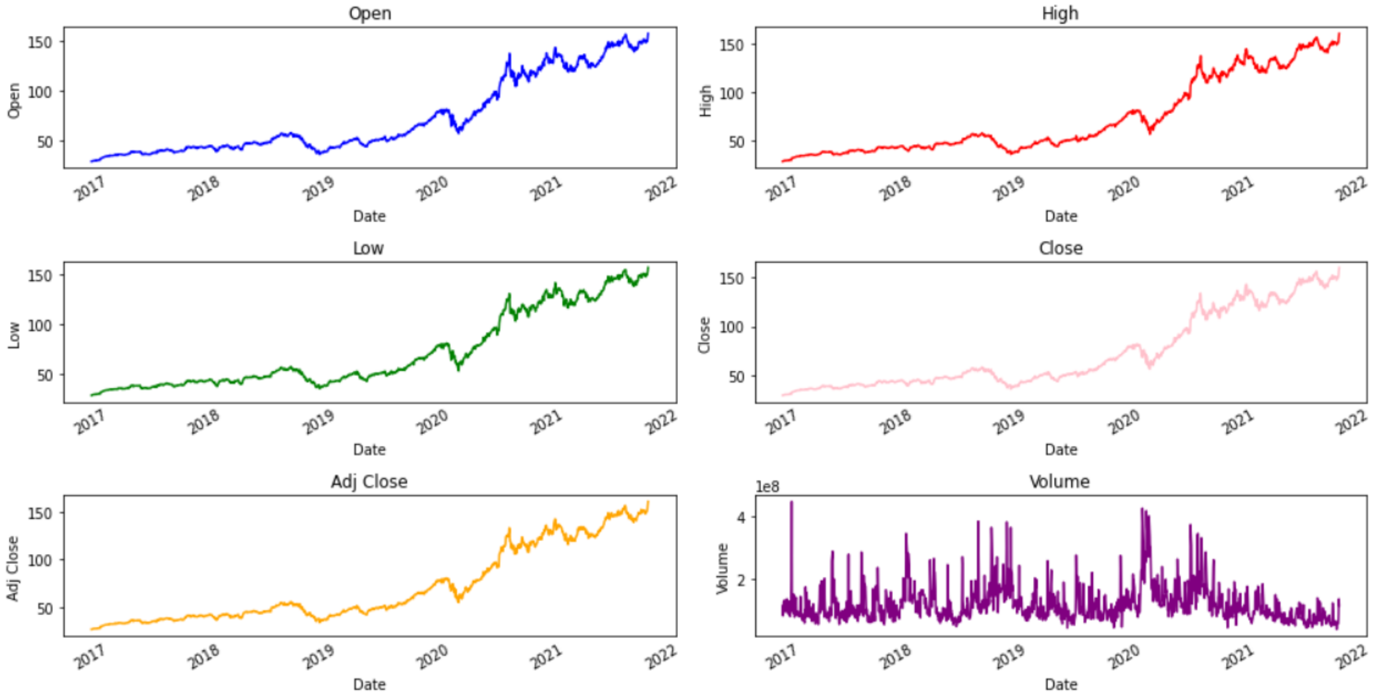
3.1 Stock Price Data

The stock market data is collected using Yahoo Finance API. I collected historical stock price data of Apple Inc for the last 4 years between the time period January 1, 2017 to November 20, 2021.



Date	Open	High	Low	Close	Adj Close	Volume
2017-10-04	957.000000	960.390015	950.690002	951.679993	951.679993	952400
2017-10-05	955.489990	970.909973	955.179993	969.960022	969.960022	1213800
2017-10-06	966.700012	979.460022	963.359985	978.890015	978.890015	1173900
2017-10-09	980.000000	985.424988	976.109985	977.000000	977.000000	891400
2017-10-10	980.000000	981.570007	966.080017	972.599976	972.599976	968400
2017-10-11	973.719971	990.710022	972.250000	989.250000	989.250000	1693300

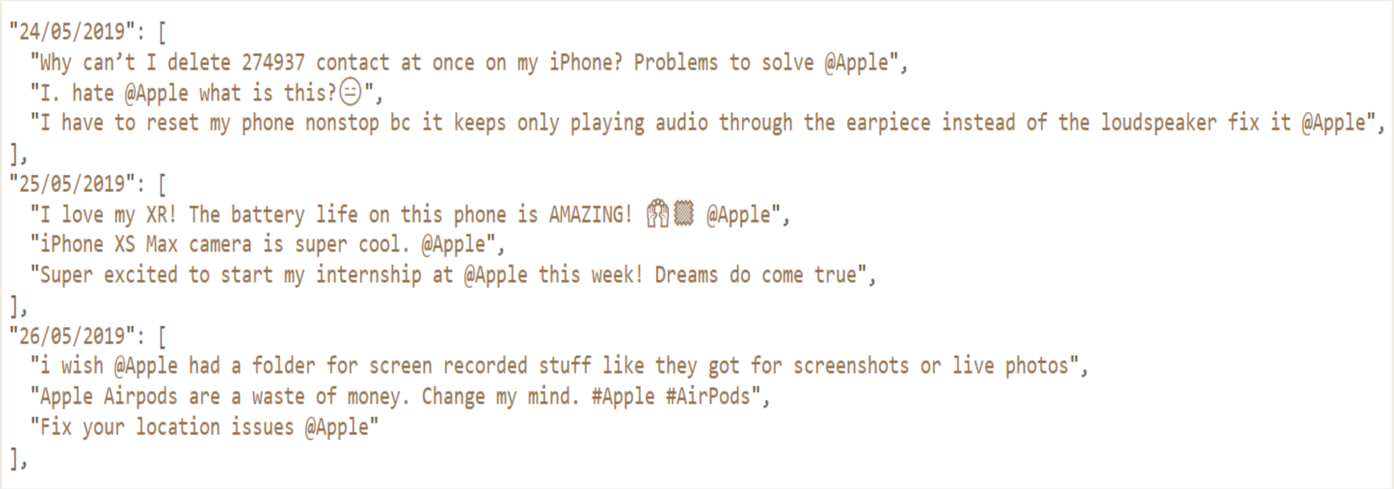
The dataset contains 6 features: [Open, High, Low, Close, Adj Close, Volume]. I've used the closing price as the ***target variable***.



3.2 Tweets Data for Sentiment Analysis

To create dataset for sentiment analysis, I scraped tweets from Twitter using *snscrape API*. I collected three tweets per day of Apple Inc. for the last 4 years between the time period Jan 1, 2017 to Nov 20, 2021.

The dataset contains **5355 tweets** in total. The rationale behind taking three tweets per day is that a single tweet cannot capture the market sentiment about the company. In general, more the number of tweets analyzed will ensure better market coverage, however due to lack of computational resources, I could take only three tweets per day.



```
"24/05/2019": [
  "Why can’t I delete 274937 contact at once on my iPhone? Problems to solve @Apple",
  "I. hate @Apple what is this?😑",
  "I have to reset my phone nonstop bc it keeps only playing audio through the earpiece instead of the loudspeaker fix it @Apple",
],
"25/05/2019": [
  "I love my XR! The battery life on this phone is AMAZING! 🙌▨ @Apple",
  "iPhone XS Max camera is super cool. @Apple",
  "Super excited to start my internship at @Apple this week! Dreams do come true",
],
"26/05/2019": [
  "i wish @Apple had a folder for screen recorded stuff like they got for screenshots or live photos",
  "Apple Airpods are a waste of money. Change my mind. #Apple #AirPods",
  "Fix your location issues @Apple"
],
```

A glimpse Twitter data for sentiment analysis

Link to codes: Google Colab Jupyter Notebook

Link to dataset: Google Drive

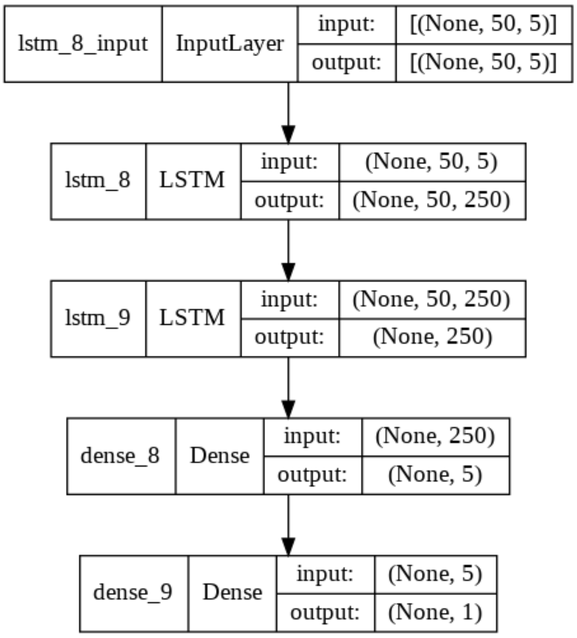
4. Experiments

4.1 Baseline Model - Multivariate Time Series Analysis with LSTM

Predicting the price of a financial asset is a complex challenge. In general, the price is determined by an endless number of variables. Economic cycles, political developments, unforeseen events, psychological factors, market sentiment, and even the weather, all these variables will more or less exert an influence on the price. In addition, many of these variables are interdependent, which makes statistical modeling even more complex.

A univariable forecast model reduces this complexity to a minimum – a single factor and ignores the other dimensions. A multivariate model can take several factors into account. For example, it can consider the relationship between the closing price and the opening price, daily highs, the price of other stocks, and so on. They offer a more detailed abstraction of reality than univariate models. Thus, I chose Multivariate model over Univariate model as a baseline model for this task. I used LSTMs as basic units of model architecture because they are known to produce robust results on time series data.

Model Architecture



lstm_8_input | InputLayer | input: [(None, 50, 5)] | output: [(None, 50, 5)]

lstm_8 | LSTM | input: (None, 50, 5) | output: (None, 50, 250)

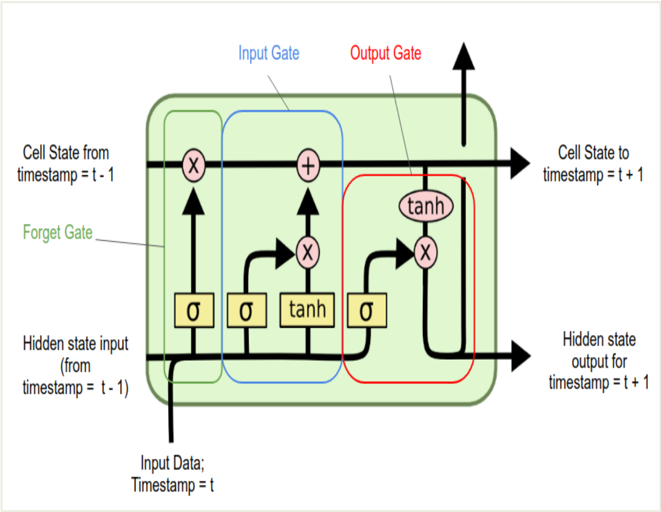
lstm_9 | LSTM | input: (None, 50, 250) | output: (None, 250)

dense_8 | Dense | input: (None, 250) | output: (None, 5)

dense_9 | Dense | input: (None, 5) | output: (None, 1)

Model Layers

-

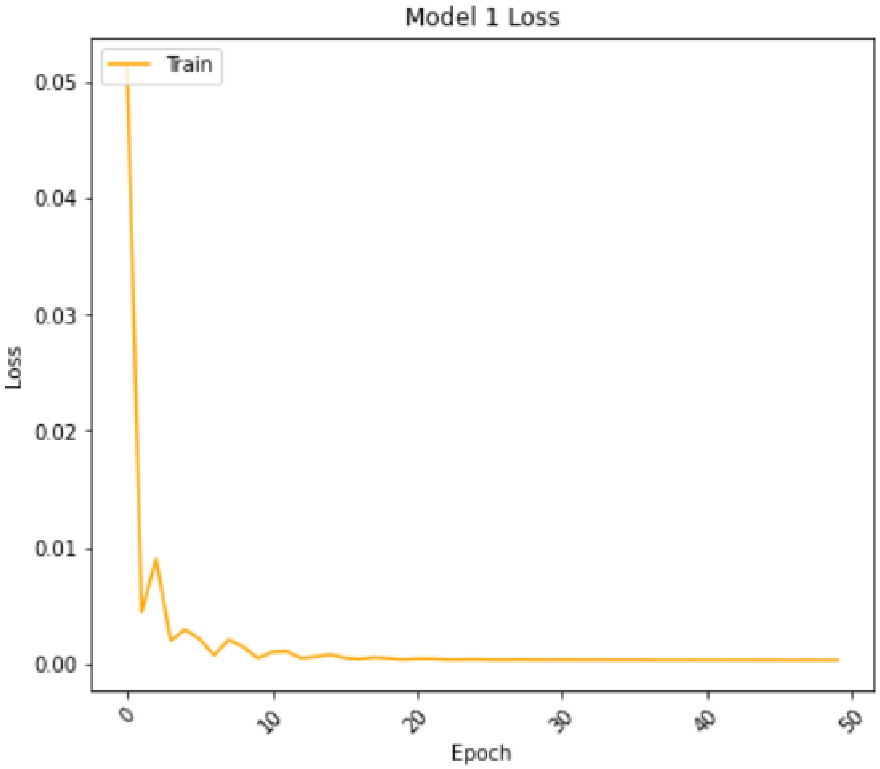

Input Gate
Output Gate
Cell State from timestamp = t - 1
Cell State to timestamp = t + 1
Forget Gate
X
+
tanh
X
X
σ
σ
tanh
σ
Hidden state input (from timestamp = t - 1)
Hidden state output for timestamp = t + 1
Input Data; Timestamp = t

LSTM UNIT

Model Details

- **Loss Function**: Mean Squared Error (MSE)
- **Activation Function:** tanh
- **Epochs**: 50 (With Early Stopping)
- **Optimizer:** Adam
- **Learning Rate**: 0.001
- **Trainable parameters:** 758,261
- The model was evaluated on **Median Absolute Error (MAE)**

Loss Curve

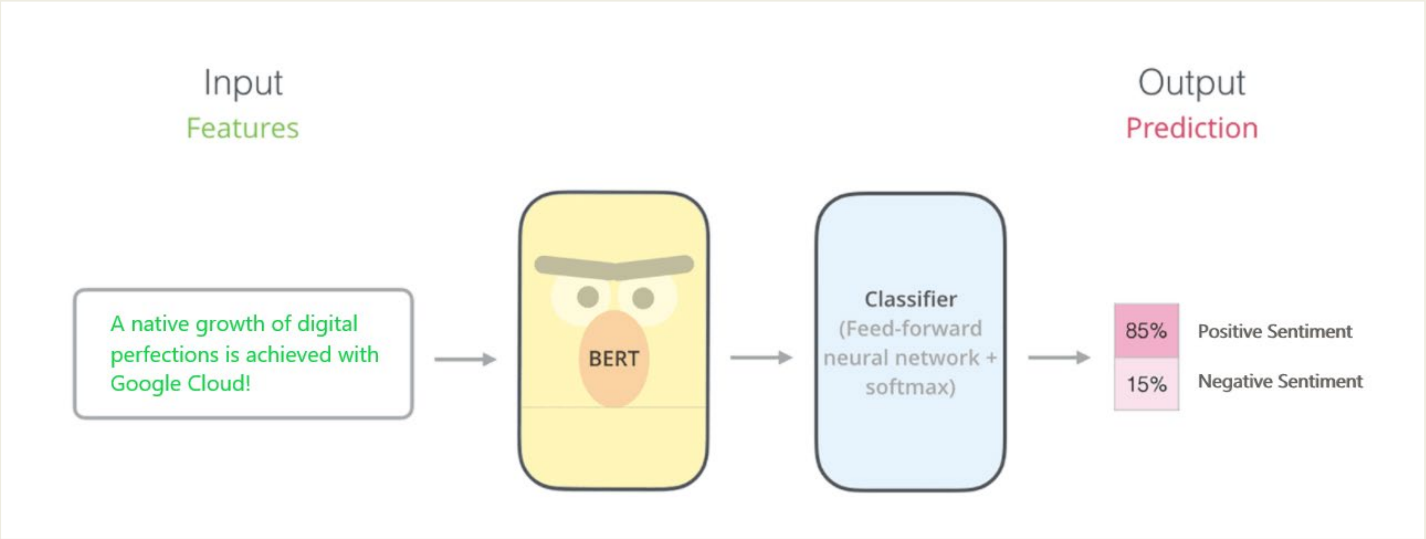


4.2 Joint LSTM & Sentiment Analysis Model

Architecture: The LSTM model is kept as base without any changes for the purpose of *fair comparison*. Along with the previous features, 3 additional features obtained from sentiment analysis of tweets were added.

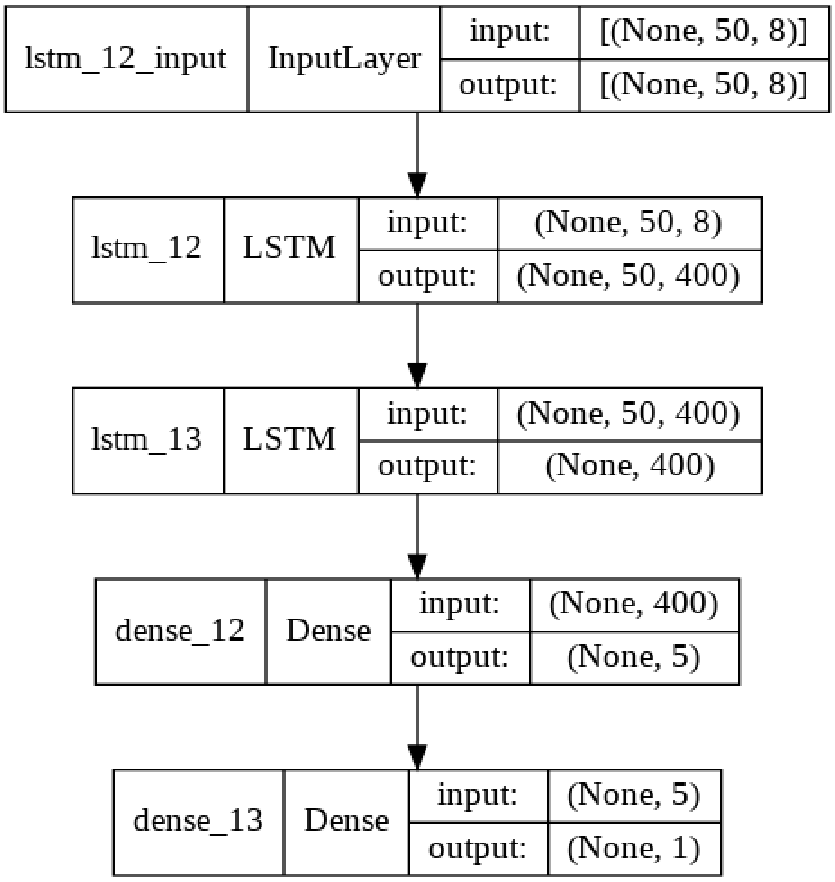
Sentiment Analysis

I leveraged the pretrained BERT (Bidirectional Encoder Representations from Transformers) model to encode the tweets. The BERT embeddings of tweets were fed to sentiment analyzer. I used roBERTa-base model trained on ~58M tweets and finetuned for sentiment analysis with the TweetEval benchmark as sentiment analyzer. The output of the model are 3 scores denoting degree of positive, negative, neutral sentiment expressed by the tweet.


Input
Features
A native growth of digital perfections is achieved with Google Cloud!
BERT
Classifier
(Feed-forward neural network + softmax)
Output
Prediction
85%
Positive Sentiment
15%
Negative Sentiment

The dataset contains 3 tweets per day. I calculated average sentiment scores per day. The output of the above model is fed as an additional feature to LSTM model for final stock price prediction.

Model Architecture

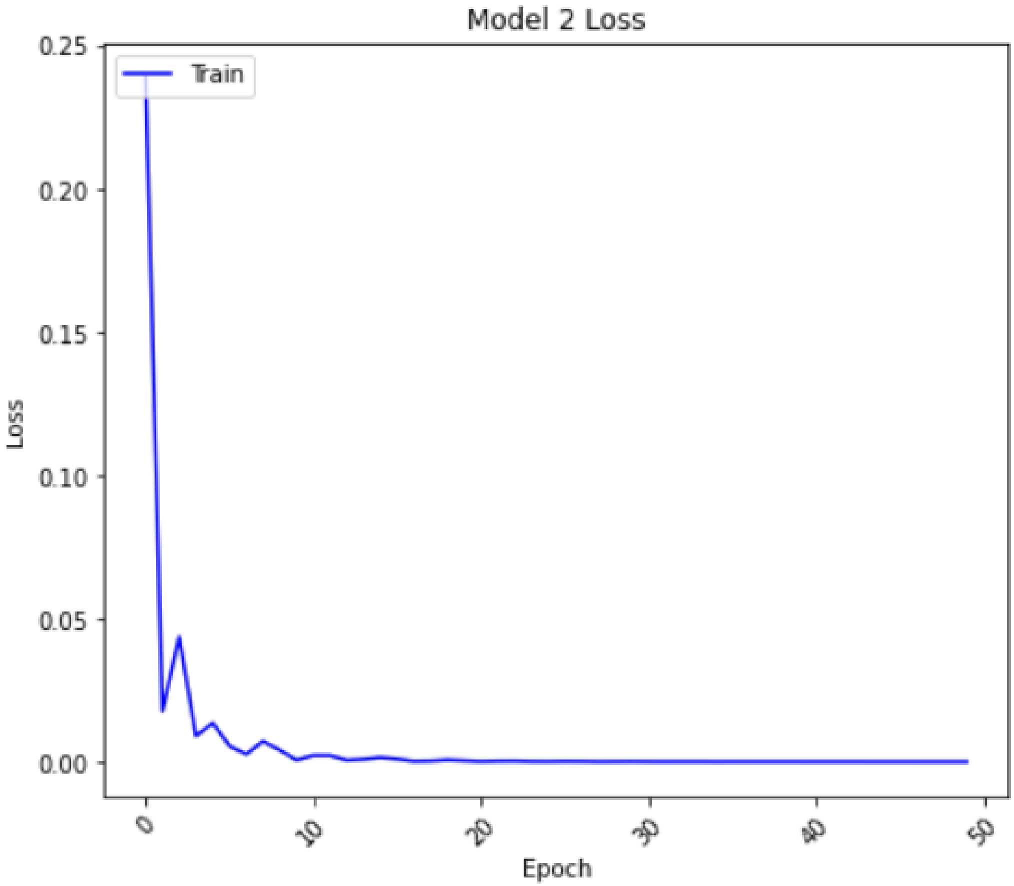


Model Details

- **Loss Function**: Mean Squared Error (MSE)
- **Activation Function:** tanh
- **Epochs**: 50 (With Early Stopping)
- **Optimizer:** Adam
- **Learning Rate**: 0.001
- **Trainable parameters**: 1,938,011
- The model was evaluated on **Median Absolute Error (MAE)**

Note: *Specifications/details of underlying LSTM model were kept same to ensure fair comparison*

Loss Curve


Model 2 Loss
Train
Loss
0.25
0.20
0.15
0.10
0.05
0.00
Epoch
0
10
20
30
40
50

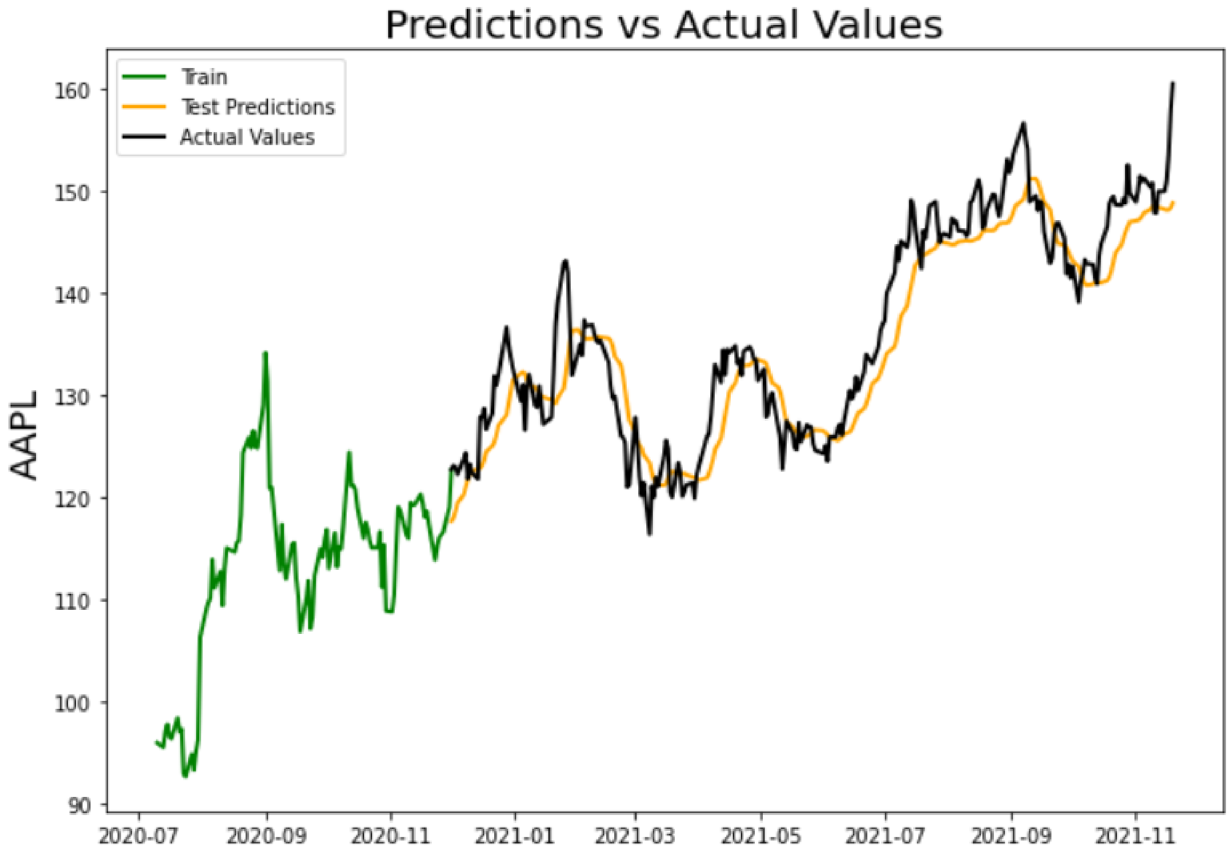
5. Results

5.1 Multivariate Time Series Analysis with LSTM (Model 1)

Median Absolute Error (MAE): 3.11

Mean Absolute Percentage Error (MAPE): 2.27 %

Median Absolute Percentage Error (MDAPE): 1.87 %

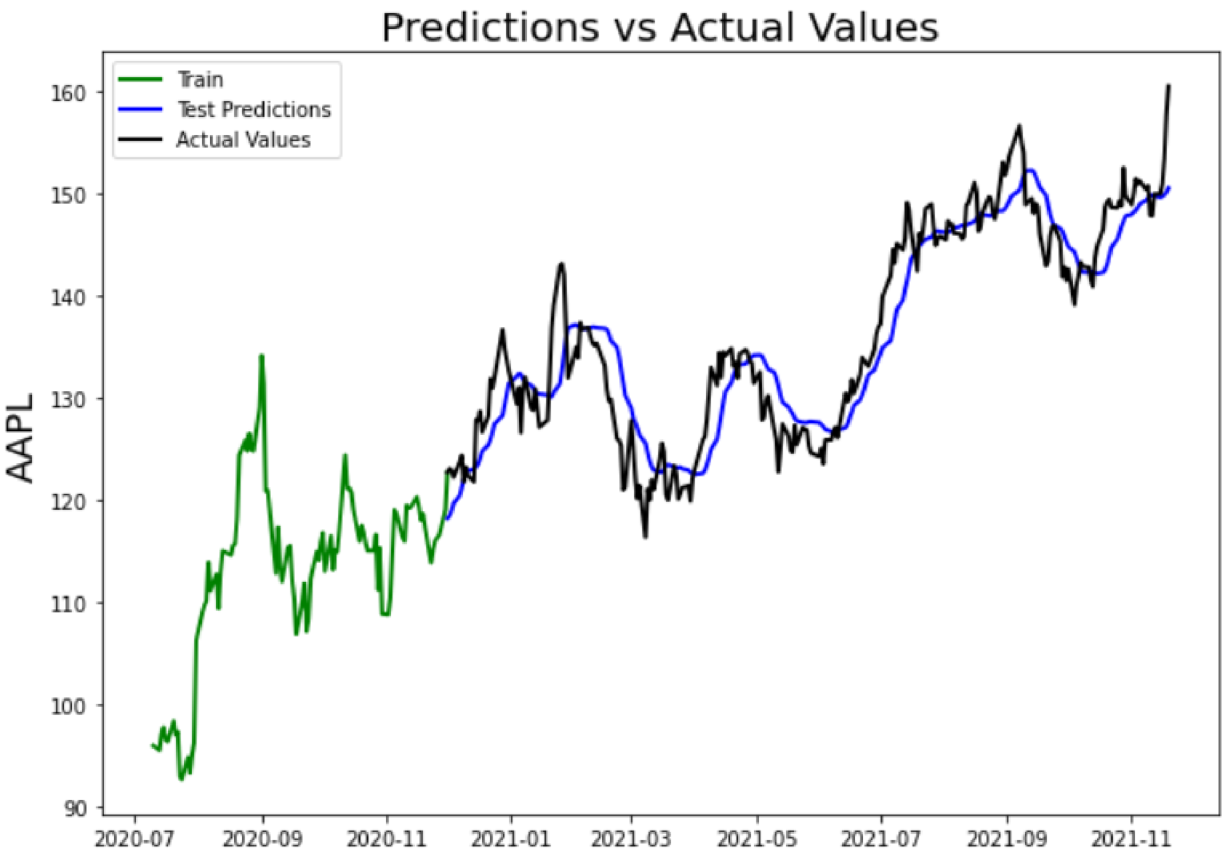


5.2 Joint LSTM & Sentiment Analysis Model (Model 2)

Median Absolute Error (MAE): 3.08

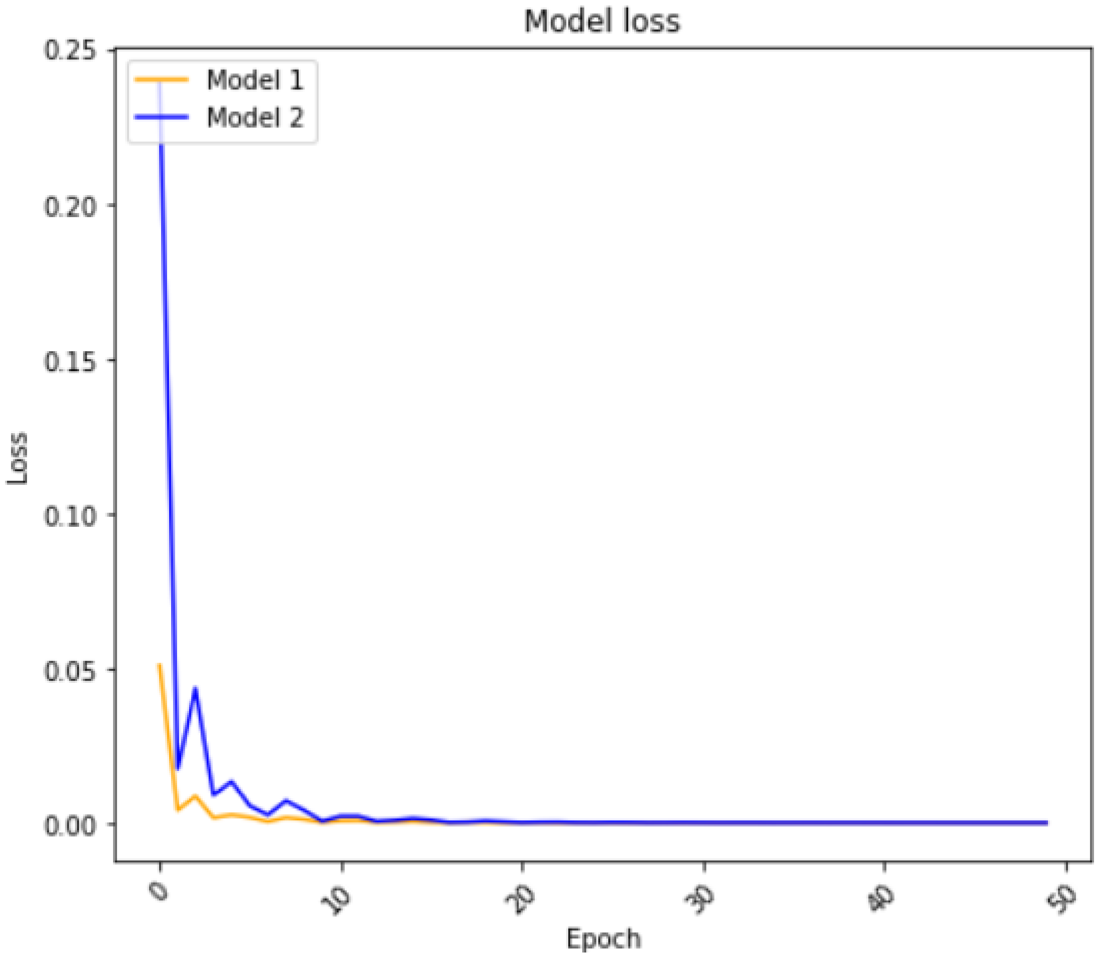
Mean Absolute Percentage Error (MAPE): 2.28 %

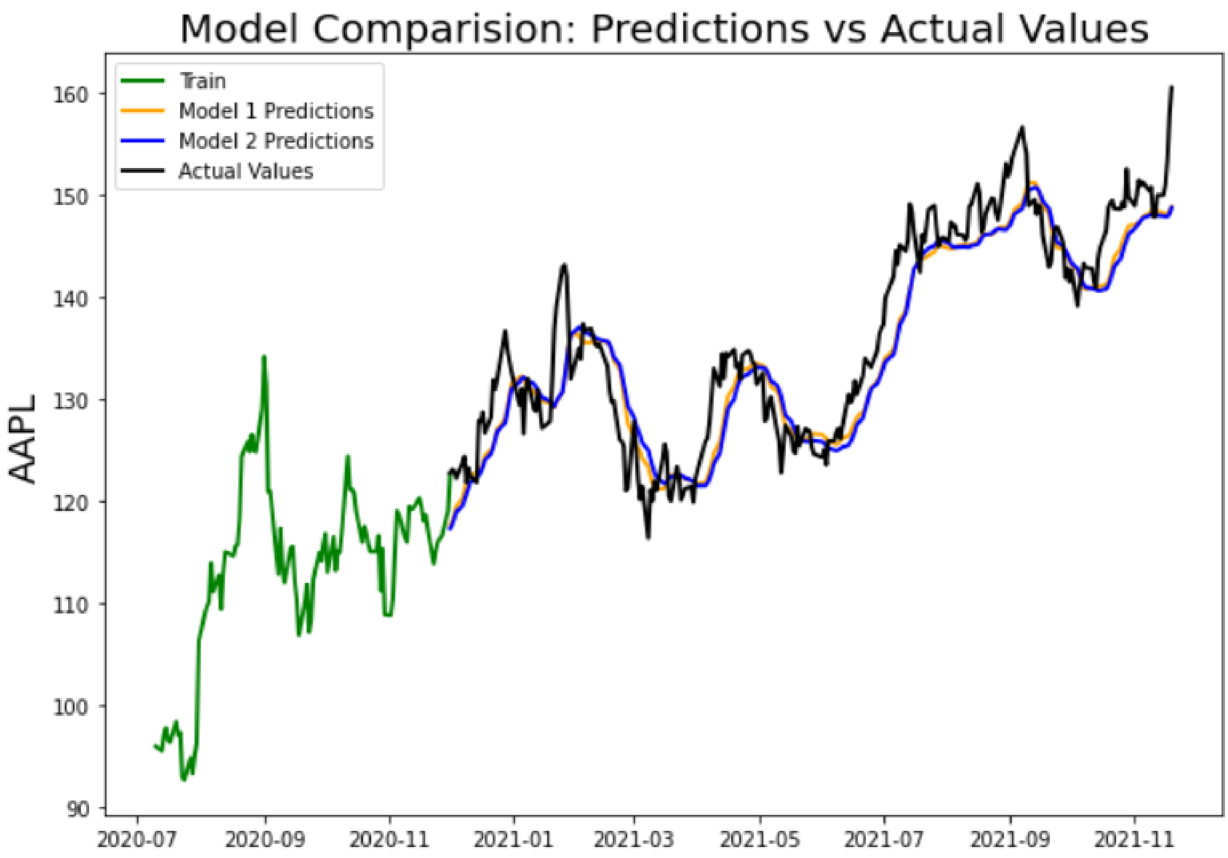
Median Absolute Percentage Error (MDAPE): 1.92 %



5.3 Performance Comparison

Metric	Model 1	Model 2
Median Absolute Error (MAE)	3.11	3.08
Mean Absolute Percentage Error (MAPE)	2.27%	2.28%
Median Absolute Percentage Error (MDAPE)	1.87%	1.92%




Model Comparision: Predictions vs Actual Values
Train
Model 1 Predictions
Model 2 Predictions
Actual Values
AAPL
160
150
140
130
120
110
100
90
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11

6. Conclusion

After comparing performance of both the models, ***I conclude that there exists statistical co-relation between the popularity on the social media and the companies' stock price.*** Median Absolute Error (MAE) score obtained using sentiment analysis is 0.03 less than that of the baseline model indicating the sentiment analysis yields better results. However, the differences are not too large and both the models perform nearly the same on Apple Stock Price Dataset.

For Future work, one can improve the dataset by collecting news articles regarding different companies in order to capture the market sentiment better. Experiments on data from multiple companies can also be conducted. Also, due to lack of computational resources, I could work with a smaller number of tweets per day. In general, analysis of a greater number of tweets will ensure better market coverage. The co-relation can be evaluated statistically using various correlation coefficients expressed by the tweet.

Finally, I conclude that the price is determined by an endless number of variables. Economic cycles, political developments, unforeseen events, psychological factors, market sentiment, and even the weather, all these variables will more or less exert an influence on the price. In addition, many of these variables are interdependent, which makes statistical modeling even more complex. While sentiment is an important indicator, developing models that take into account multiple factors shall help in increasing the robustness of the predictive models.

Link to codes: Google Colab Jupyter Notebook

Link to dataset: Google Drive