**Title**
Incorporation of Potential Sentiment Analysis Variable from Social Media in Stock Price Prediction



**Author**
MIAO, YU

**Publication Date**
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

# Incorporation of Potential Sentiment Analysis Variable from Social Media in Stock Price Prediction

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Applied Statistics

by

Yu Miao

2022

ABSTRACT OF THE THESIS

Incorporation of Potential Sentiment
Analysis Variable from Social Media
in Stock Price Prediction

by

Yu Miao
Master of Science in Applied Statistics
University of California, Los Angeles, 2022
Professor Yingnian Wu, Chair

Numerous factors impact stock prices. Some of the significant factors are not quantitative, which increases the difficulty for researchers to include them in commonly-used stock price prediction models. Among these non-quantitative factors, the influence of user-generated comments and posts on social media towards specific stocks on stock price is significant. Including these factors in stock price prediction model may improve the overall prediction accuracy. Therefore, this study introduces a flexible stock price prediction framework that includes textual data from social media. This framework can also be extended to most of the models in the stock price prediction field. The basic logic behind this framework is to convert the textual social media contents into a numerical variable - "daily sentiment score", which can be adopted in most of the prediction models. Furthermore, the framework was tested on the close price prediction for five major stocks in the US stock market: Apple, Microsoft, Tesla, Amazon, and Google. Results showed that the prediction accuracy improved for most LSTM models by including the additional sentiment variable. Future studies can be

conducted to investigate the relationship between "daily sentiment score" and daily stock price movement.

The thesis of Yu Miao is approved.

Frederic Paik Schoenberg

Nicolas Christou

Yingnian Wu, Committee Chair

University of California, Los Angeles

2022

*To my cat Didi*

*and all the other untrivials in my life*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# ACKNOWLEDGMENTS

This is my first opportunity to write an official acknowledgment for thesis. It would not have been possible without the help and support of all people and cats around me, to only some of whom it is possible to give the particular mention here.

I would like to express my deepest gratitude to everyone who accompanied, helped, and supported me during my journey. First, I am sincerely grateful to each of my committee members, Dr. Yingnian Wu, Dr. Nicolas Christou, and Dr. Frederic Paik Schoenberg, for their guidance, patience, and generous support throughout this thesis.

Further, yet importantly, I am extremely grateful for my family: Zhijie Sang, Hu Miao, Dadi (Didi) Miao, and Jing Miao for their unconditional love and endless support. I consider myself nothing without them.

# CHAPTER 1

# Introduction

Stock markets are always conceived as volatile, capricious, and unpredictable. Although numerous researchers working on stock price prediction have attempted to improve predictive accuracy, accurately predicting the stock price remains one of the most challenging tasks in financial time series forecasting due to the intrinsic complexity within the stock market [PL05]. Traditionally, methods based on statistical models (e.g., ARIMA, GARCH), machine learning (e.g., SVR), and deep learning techniques (e.g., LSTM, RNN) were extensively explored in the stock price prediction [PL05, **?**, IN20]. As these models usually lack the ability to handle text data such as news and comments on social media, which also potentially impact the stock price, an increasing number of studies has shifted to examine and include text mining methods in stock price prediction. Most studies that adopt text mining methods in stock price prediction tends to focus on the stock-related news and explored the feasibility of incorporating the stock-related news in the stock price prediction process [Fal07, FYL03, NEM10, FG18]. There has been little discussion concentrated on the impact of people's posts on social media on stock prices. In addition, while some research has incorporated text data from financial news in stock market prediction, little attention has been paid to include the text data from social media in stock market prediction.

The purpose of this paper is to introduce an innovative stock price prediction framework that enables the inclusion of unstructured text data from social media, specifically from Twitter, in the widely-used time series prediction model. The fundamental principle behind this newly-proposed framework is to convert the text data into a numerical variable - "daily

sentiment variable" through sentiment analysis that can be added in most prediction models. The proposed prediction framework offers a flexible scheme that enables the involvement of any text data from social media into the currently existing stock price prediction model. It also provides important insights for further research on whether stock price is influenced by public sentiment on social media.

Our prediction system was tested on the five major stocks in the US market: Apple, Microsoft, Amazon, Google, and Tesla. Two years (from Dec 2019 to Dec 2021) of data for these five stocks were obtained and studied in our experiment.

This paper has been organized in the following way. The second chapter gives a comprehensive overview of previous studies in this field. Chapter three describes our stock prices prediction framework and discusses the data and the methods we employed in the experiment, while chapter four presents the results from our experiment. Finally, chapter five gives a brief summary of the findings, addresses the limitations of our study, and suggests potential directions for further research.

# CHAPTER 2

# Literature Review

Traditionally, most researchers working in the stock price prediction field have attempted to apply and improve existing statistical models. The statistical approaches extensively studied in this domain can be classified into two categories: time-series and machine learning (which includes deep learning). Among the time-series models, the ARIMA model is one of the most popular ones and has been studied by numerous researchers [AAA14, STN18, MSP10]. Several studies have also employed the variants of the ARIMA model, such as the GARCH model and ARIMA-GARCH model, to predict the stock market [AC05, BR14]. Nowadays, due to the dramatic advancements in computational power, machine learning-based methodology has grown in popularity. The frequently used machine learning based models include support vector regression (SVR), Artificial Neural Network (ANN), and deep learning based models such as Long Short Term Memory [STN18, MSP10, SVG17]. To further raise the prediction accuracy in the stock market, some novel hybrid models based on these existing methods have been developed. For example, Hsu introduced a hybrid stock price prediction approach through the integration of self-organizing map neural network and genetic programming; Wang et al. presented a hybrid model based on ARIMA, ESM, and BPNN [Hsu11, WWZ12].

However, most of the models in stock price prediction have failed to take text data into account. Stocks prices are influenced by many potential factors, such as the exchange rate, the economic situation, and people's expectations [PS13]. Some of the influencing factors are not quantitative. Thus, due to the inability to process unstructured text data for most

models, most of those factors have failed to be included in the stock price prediction. In order to solve the lack of ability for utilizing text data to predict the stock price for most traditional models, several studies investigate the possibility and importance of text mining processes in stock price prediction. Schumaker et al. prove that including text data from financial news can successfully improve the prediction accuracy for stock price [SC09]. In addition, Lee et al. also highlight the importance of text data from financial event report in predicting the stock price movement [LSM14]. Due to the discovery of the significant effect of text data especially from news, an increasing amount of studies have begun to include financial news data through text mining techniques to predict the stock market. Fung et al. integrate real-time financial news into stock price prediction; Falinouss developed a classification model to predict the movement of stock price based on the financial news articles; and Nikfarja et al. developed a stock price prediction framework which is able to process textual data from the news as input [FYL03, Fal07, NEM10]. Furthermore, Hagenau et al. enhanced the existing text mining approach in financial news for stock price prediction [HLN13].

The majority of the studies that emphasize the importance of textual data in stock market prediction focus on the text data from financial news as financial news data are one of the most accessible text data related to the stock price in the past years. In current society, the emergence of online social media that enables people to share their thoughts and comments publicly empowers another type of textual data that used to be roughly inaccessible - individuals' comments about particular stocks or companies. Huang and Liu find out that people's reviews under the financial news correspond to the stock price [HL20]. Similarly, Sun et al. also prove that user-generated content from social media impact the stock price movement [SLF16].

Although social media contents seem to play an important role in stock price prediction in today's society, few studies have discussed and proposed model that includes individual-generated comments and thoughts on social media in stock price prediction. Urolagin utilized tweet contents from Twitter to predict stock price movement using Naive Bayes, and SVM

classifier [Uro17]. In addition, Jin et al. bring out a deep-learning based stock price prediction model which includes sentiment analysis for online comments [JYL20]. Though both of these models introduced by Urolagin and Jin et al. are proved to perform favorably, they are not flexible enough to accommodate other proven successful models in stock price prediction. Therefore, further studies are required to establish a flexible prediction framework that includes text data from social media and accommodate most of the proven successful models in stock price prediction.

# CHAPTER 3

# Methodology

This study proposes an innovative stock price prediction framework that processes text data from social media and includes the resultant daily sentiment variable in commonly-used prediction models. Compared to traditional time-series approaches in stock price prediction where the independent variables are primarily numerical or categorical, our prediction framework is able to process text data posted on social media and convert those text data into a numerical daily sentiment score that can be adopted in time-series prediction models.

An overall view of this prediction framework is shown in Fig 3.1. Twitter is one of the most popular social media platforms that allows people to discuss their views and opinions of everything, including stocks and companies. The contents of each tweet containing the stock keywords tends to reflect individuals' expectations towards the prospects of that stock, and research shows that stock trading responds to changes in expectations. Both of these are in turn related to stock price changes [HR12]. Therefore, in our experiment, we tested our prediction framework using the tweet contents data as the text data and close price as our numerical data. The text data from Twitter posts were cleaned and converted into daily sentiment variables using sentiment analysis. The new daily sentiment variable could then be included in the widely-used time series models in stock price prediction; AR/VAR and LSTM models were selected in our case for experiment. The primary objective of our experiment is to evaluate the feasibility and performances of this newly-proposed prediction framework on the close price prediction for the stock compared to the original model without the text mining variable.

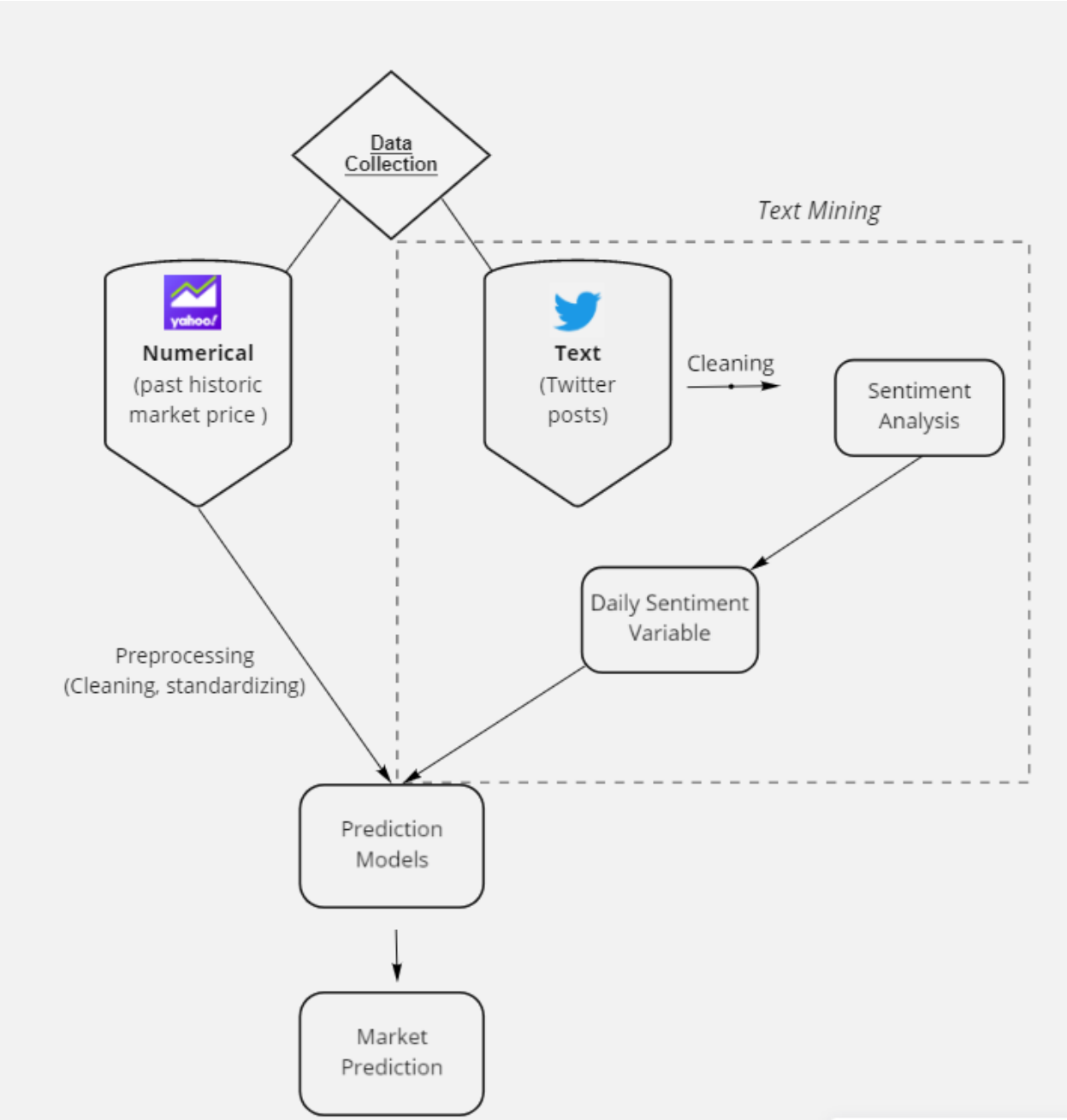



Figure 3.1: Overall View of Stock Price Prediction System

## 3.1 Data

This newly proposed stock price prediction system focused on the close price prediction. It was tested on five major stocks in the US stock market: Apple, Microsoft, Amazon, Google, and Tesla. These five stocks are included in the SP 500 index and are the top companies with huge trading volume and well-known public popularity.

The data we used in this study can be roughly classified into two categories: numerical and text data. Numerical data includes the past daily close prices for each stock. We obtained these historic close data from Yahoo Finance. For text data, we scraped all the past tweets data that was written in English and contained any of the five stock names ($APPL, $MSFT, $AMZN, $GOOGL, $TSLA) from Twitter API.

Two years of data between December 2019 to December 2021 with 506 daily close prices for each stock were used in this study. The specific amount of English tweets posted within this two year and contained each of the following stock keywords are showed in the Table 3.1. According to Table 3.1, Tesla stock is the most popular stock on twitter, where Apple stock is the least popular stock.

| | **Number of Tweets** |
|---|---|
| ***$AAPL*** | 31,909 |
| ***$AMZN*** | 604,700 |
| ***$GOOGL*** | 177,151 |
| ***$MSFT*** | 350,768 |
| ***$TSLA*** | 1,922,936 |

Table 3.1: Number of Tweets Comments for Each Stock

These text data were processed and converted into 506 daily sentiment variables for each stock using the text mining approaches described in the following sections. For every stock in our prediction system, the first 80% of the data were used as training data for ARIMA and

VAR model, while the remaining 20% became test data for ARIMA and VAR model. We trained the LSTM model on the first 70% of the whole dataset; the remaining 30% became the test data.

## 3.2 Text Mining

Text mining refers to the methods of extracting and discovering useful information of interests from textual data. In our prediction system, tweets data were analyzed using several text mining approaches and converted into numeric variable - "Daily Sentiment Variable". Figure 3.2 presents steps in this conversion process.

First, the tweets data were cleaned to remove any hyperlinks, emojis, and special characters. After that, sentiment analysis was conducted to extract the sentiment scores that underlie each tweets using the VADER model, since the main goal of the text mining process in our framework is to assess the public expectations towards the stock they posted. Based on the sentiment scores for each tweets, the daily sentiment scores for each stock were calculated by averaging the sentiment scores of all the tweets posted within that day.

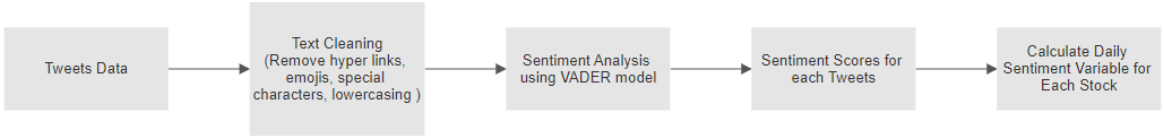



Figure 3.2: Text Mining Process in the Prediction System

### 3.2.1 VADER

The VADER model was employed in the sentiment analysis process. VADER was first introduced in 2014 as a sentiment analysis model specifically designed for social media text.

It was proved to perform outstandingly in classifying social media text into three classes (positive, negative, neutral).

The model was built up through a human-centric approach which combines qualitative analysis with empirical validation and experimental investigations [HG14]. The basic logic behind VADER is to map lexical features in the text to sentiment scores based on the built-in dictionary. By calculating the average of the sentiment scores for each lexical units within the text, it returns a sentiment score ranging from -1 to 1, with -1 referring to most negative and 1 referring to most positive. Examples of tweets and their corresponding sentiment score are showed in Table 3.2.

| **Tweets** | **Score** |
|---|---|
| Would it better to shift to $APPL as a safe heaven until he feels better? | 0.926 |
| I'll probably just keep individual stock $MSFT because of their moat. | 0.0 |
| $APPL's Q4 results - delivering a revenue miss after slow iPhone sales this quarter. | -0.5106 |
| I just lead my group to bank some great $GOOGL gains | 0.6808 |
| as long as its in $AAPL, $FB, $AMZN, $GOOGL, $TSLA, you cant lose SMH | 0.5807 |

Table 3.2: Tweets Examples and Their Calculated Sentiment Scores

## 3.3 Time Series Models

In order to test the feasibility and performance of the proposed daily sentiment variable in stock price prediction, two commonly used time series models in financial series forecasting were adopted: ARIMA (and its extension VAR),and LSTM. The original models without the daily sentiment variable and the new models with the daily sentiment variable were applied for each stock to predict their close price. The performances of the models with the daily sentiment variable and without the daily sentiment variable were compared.

### 3.3.1 ARIMA

Autoregressive Integrated Moving Average (ARIMA) was first introduced by Box and Jenkins in 1970 as a method of time-series forecasting. This model has become one of the most common and classic methods in stock price prediction. As ARIMA's name indicates, autoregression, integrated, and moving average are the three key components of the model [STN18]. (p,d,q) usually refers the order of the ARIMA model, with p represents the autoregressive terms; q indicates the integrated terms and q corresponds to the moving average terms. The general form of ARIMA (p,d,q) for $X_t$ can be expressed as the following:

$$(1 - \sum_{k=1}^{p} \alpha_k L^k)(1 - L)^d X_t = (1 + \sum_{k=1}^{q} \beta_k L^k)\epsilon_t \tag{3.1}$$

where

$$\Delta X_t = X_t - X_t - 1 = (1 - L)X_t \tag{3.2}$$

$$L^k x_t = x_{1-k} \tag{3.3}$$

$\alpha_k$ and $\beta_k$ refers to the coefficients, $\epsilon_t$ refers to the error term.

One of the most crucial procedures in building up ARIMA model is to identify the (p,d,q) order [Oza77]. Since the main objective of this experiment is to check the viability and improvement for the newly proposed system with the daily sentiment variable, we used (1,1,0) for all of our ARIMA models, which is exactly the same as first-differenced autoregressive (AR) model with order 1.

#### 3.3.1.1 VAR

The vector Autoregressive (VAR) model is another extension of the AR model. As ARIMA and AR are limited in the use of univariate cases, the development of VAR extended its

usefulness to multivariate cases. The general form of VAR with order p can be written as:

$$X_t = C_0 + B_1 X_{t-1} + B_2 X_{t-2} + ... + B_p X_{t-p} + \epsilon_t \tag{3.4}$$

where $X_t = (x_{1t}, x_{2t}, ..., x_{nt})'$ refers to an nx1 vector of time series variables, $B_n$ represents the nxn coefficient matrices $\epsilon_t$ is a n-dimensional white noise process with time-invariant positive covariance matrix [Kot].

In order to match the order of ARIMA model without the sentiment variable for comparison, first-differenced VAR with order 1 were applied to predict the daily stock price differences.

### 3.3.2 LSTM

Long Short-Term Memory (LSTM) was first introduced by Sepp Hochreiter and Jurgen Schmihuber as a novel type of recurrent neural network that is adapted to learn order dependence in sequence prediction problems efficiently [HS97]. Due to its powerful advantages in exploiting the interactions and patterns in the data, it has become one of the most widely used approaches in the domain of financial time-series forecasting.

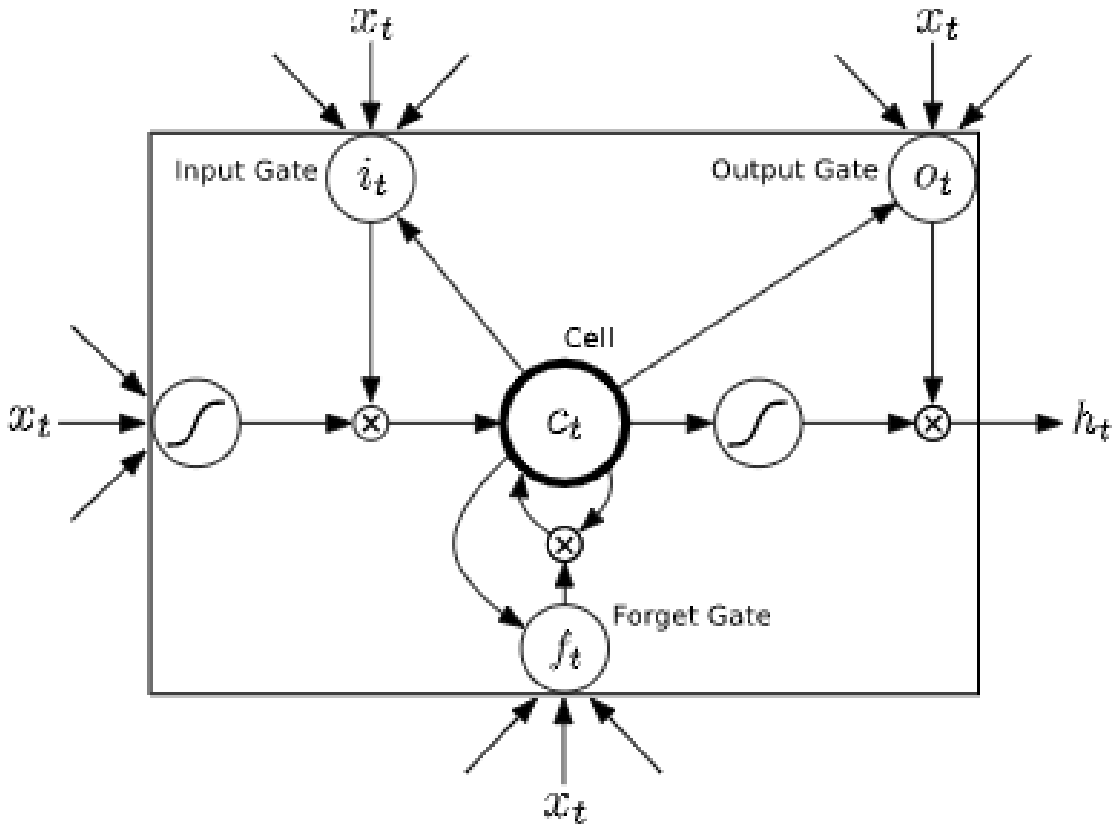



Figure 3.3: LSTM Network [Gra13]

Compared to conventional RNNs, LSTM includes the memory cell that enables it to learn longer-time dependencies. It also contains input and forget gates that learn to promote better preservation of long-term dependencies [LBH15]. Fig3.3 illustrates the typical structure for an LSTM unit. $x_t, h_t, c_t$, refer to the input data at time t, the output from the LSTM cell at time t, and the value of the memory cell at time t, respectively. At time t, the input gate (i) decides the extent of the information carried, while the forget gate (f) determines the extent of information that would drop from the previous data. The output gate (o) selects the portion of the information delivered to the next layer.

The calculation process of LSTM unit is listed below:

$$i_t = \sigma(W_{xi}x_i + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{3.5}$$

$$f_t = \sigma(W_{xf}x_i + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{3.6}$$

$$o_t = \sigma(W_{xo}x_i + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \tag{3.7}$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{3.8}$$

$$h_t = o_t tanh(c_t) \tag{3.9}$$

$\sigma$ denotes the logistic sigmoid function, and $W_{hi}, W_{xo}, W_{xf}$ represent the weight matrix for hidden-input gate, weight matrix for the input-output gate, weight matrix for the input-forget gate accordingly, etc. The bias terms in the actual calculation process (for i, f, c, and o) were omitted above for brevity [Gra13].

Sliding window refers to the approach of utilizing previous time steps to predict the next time step. Cross-validation is often used to determine the optimal window size in the prediction process. Since the main objective of this study is to test the effectiveness of including the novel daily Twitter sentiment variable, precisely the same size: 60 was used for all the LSTM models for better comparison. Additionally, all the LSTM models in our experiment have two layers with 50 neurons, one dense layer with 25 neurons and another dense layer with one neuron. Furthermore, all models were optimized through Adam and adopted mean squared error (MSE) to calculate loss.

## 3.4 Performance Evaluation

Two metrics, mean average percentage error (MAPE) and root mean squared error (RMSE), were adopted to evaluate the performances of the proposed stock price prediction system. MAPE and RMSE are widely used metrics for model evaluations. While MAPE measures the average percentage error, RMSE refers to the standard deviation of the mean prediction

errors. Since these two metrics measure the models' prediction errors, low values indicate good model performances and vice versa. RMSE and MAPE can be calculated using the following formulas:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \bar{y_t}}{y_t}\right| \tag{3.10}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}(y_t - \bar{y_t})^2}{N}} \tag{3.11}$$

Both of these two metrics were calculated for the models with the daily sentiment variable and without the daily sentiment variable on the test dataset.

Additionally, the plots of prediction values versus actual values for models with daily sentiment variable and without daily sentiment variable were also utilized to present some intuitive ideas on which models works better.

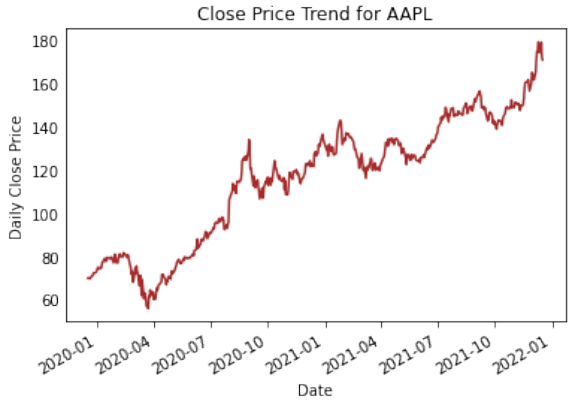
# CHAPTER 4

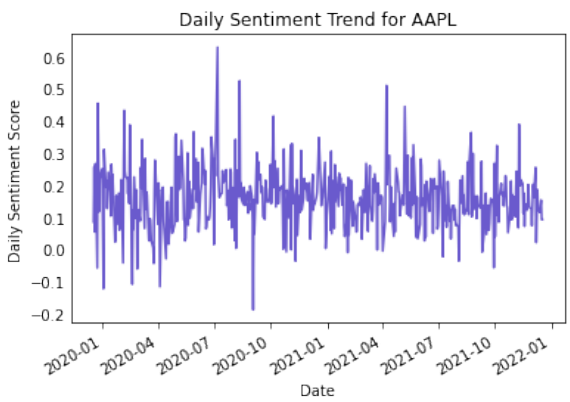
# Results and Analysis

## 4.1 Sentiment Analysis

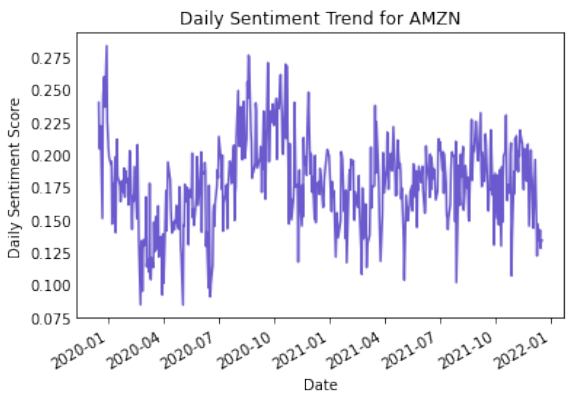
The daily sentiment variables for each stock were calculated using the approach discussed in the previous methodology section. Fig 4.1 presents the close price trend and daily sentiment trend for the five stocks from December 2019 to December 2021. In general, both of the five stocks indicate an uptrend in their close price despite the short-term price fluctuations. However, the daily sentiment trend for the five stocks seemed to fluctuate around their mean levels. And both of their mean levels are positive, demonstrating an overall positive expectation from the tweets towards the stocks.

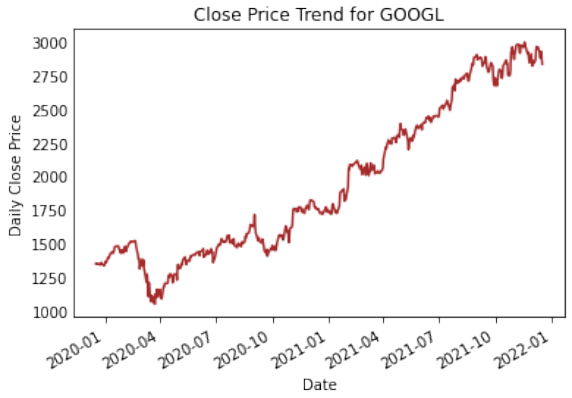
In addition, the short-term price fluctuations indirectly intercorrelate with the temporary variations in daily sentiment scores to some extent. For example, the close price for GOOGL sharply dropped from January 2020 to April 2020, while the daily sentiment score for GOOGL also demonstrated a pronounced downtrend during this period. In addition, a similar phenomenon also occurred for AMZN during the end of 2022, where both the close price and daily sentiment simultaneously fell. However, there were some circumstances where the close price trend and daily sentiment trend did not correspond with each other. Hence, the relationship between the daily Twitter sentiment and close price trends remains implicit and needs further investigation.

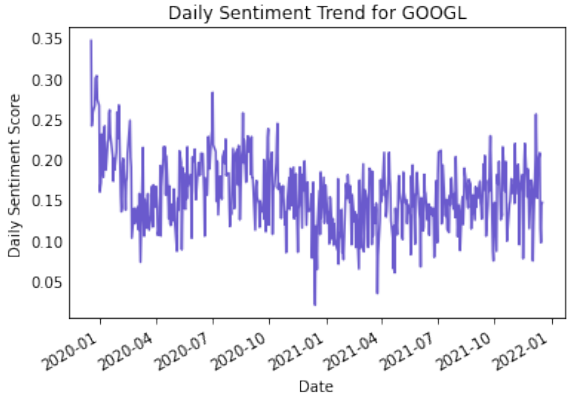

Close Price Trend for AAPL
Daily Close Price
Date

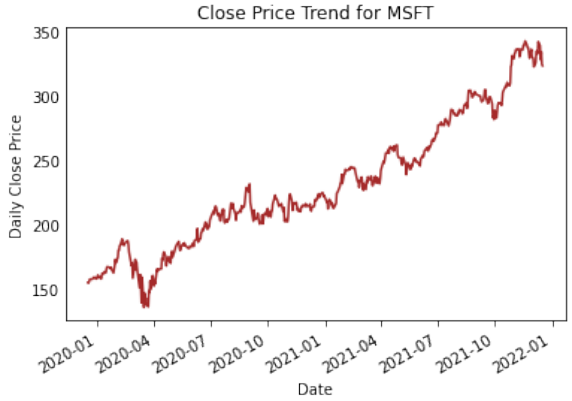

Daily Sentiment Trend for AAPL
Daily Sentiment Score
Date

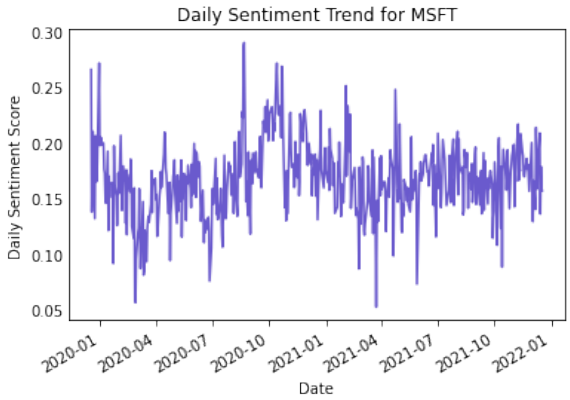

Close Price Trend for AMZN
Daily Close Price
Date


Daily Sentiment Trend for AMZN
Daily Sentiment Score
Date


Close Price Trend for GOOGL
Daily Close Price
Date


Daily Sentiment Trend for GOOGL
Daily Sentiment Score
Date


Close Price Trend for MSFT
Daily Close Price
Date


Daily Sentiment Trend for MSFT
Daily Sentiment Score
Date

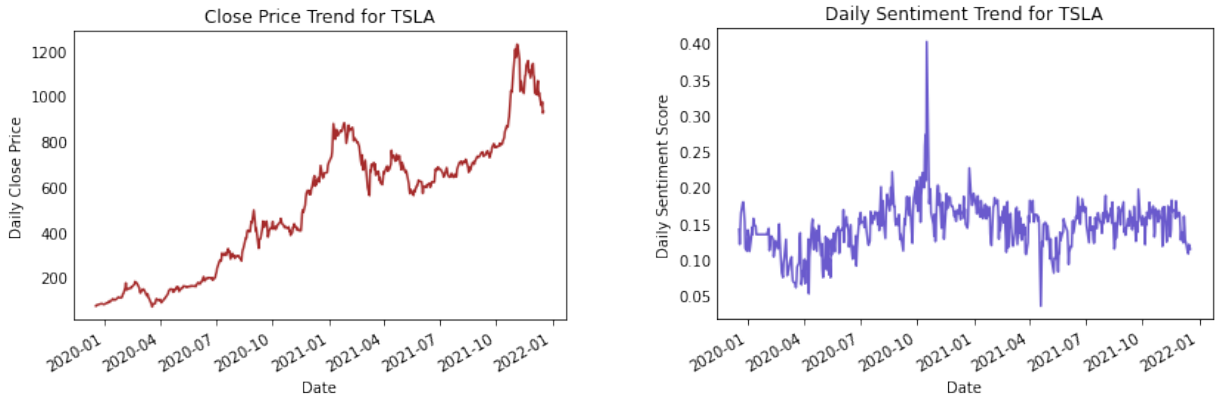



Figure 4.1: Price and Sentiment Trend for Each Stocks

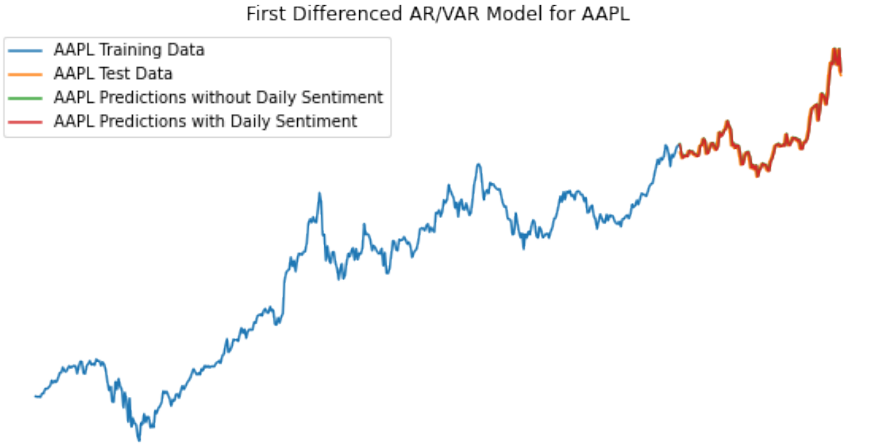
## 4.2 Model Performance

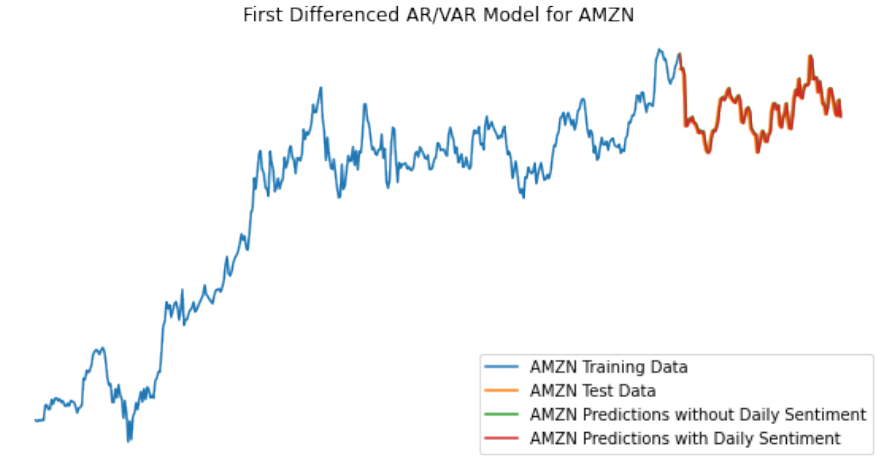
To test the effectiveness of including the daily sentiment score as an additional independent variable in close price prediction, the performance metrics (RMSE and MAPE on test data) for first-differenced ARIMA/VAR model with and without sentiment variable were calculated and presented in Table 4.1. From Table 4.1, RMSEs and MAPEs for each models with and without the additional daily sentiment score are similar.

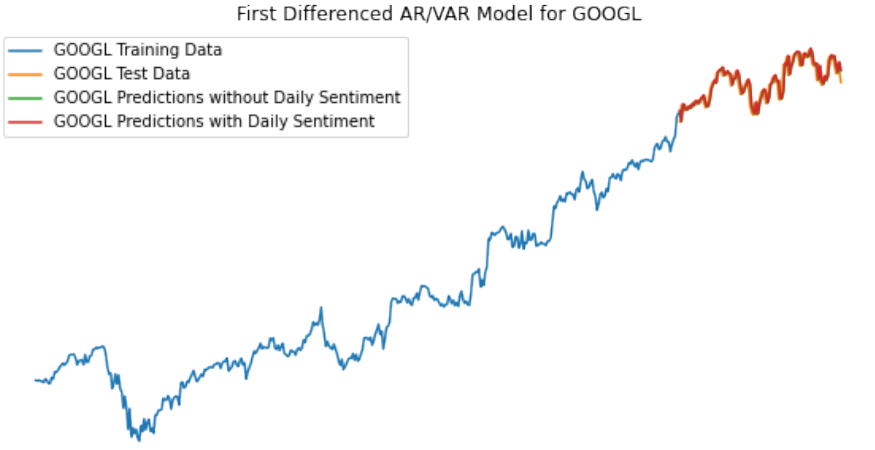
Figure 4.2 presents the visualizations for ARIMA/VAR model predictions for each stock. Since the first-differenced methods were used to predict close price for each stock, the predictions follows actual price trend closely. Additionally, the impact of including the daily sentiment score on model performances are subtle according to Figure 4.2.

| | Without Sentiment Variable | | With Sentiment Variable | |
|---|---|---|---|---|
| | **RMSE** | **MAPE** | **RMSE** | **MAPE** |
| **AAPL** | 2.24 | 0.01 | 2.24 | 0.01 |
| **AMZN** | 55.56 | 0.01 | 55.07 | 0.01 |
| **GOOGL** | 39.49 | 0.01 | 39.89 | 0.01 |
| **MSFT** | 4.02 | 0.01 | 4.02 | 0.01 |
| **TSLA** | 30.91 | 0.202 | 30.96 | 0.02 |

Table 4.1: Performances for ARIMA/VAR model


First Differenced AR/VAR Model for AAPL
AAPL Training Data
AAPL Test Data
AAPL Predictions without Daily Sentiment
AAPL Predictions with Daily Sentiment


First Differenced AR/VAR Model for AMZN
AMZN Training Data
AMZN Test Data
AMZN Predictions without Daily Sentiment
AMZN Predictions with Daily Sentiment


First Differenced AR/VAR Model for GOOGL
GOOGL Training Data
GOOGL Test Data
GOOGL Predictions without Daily Sentiment
GOOGL Predictions with Daily Sentiment

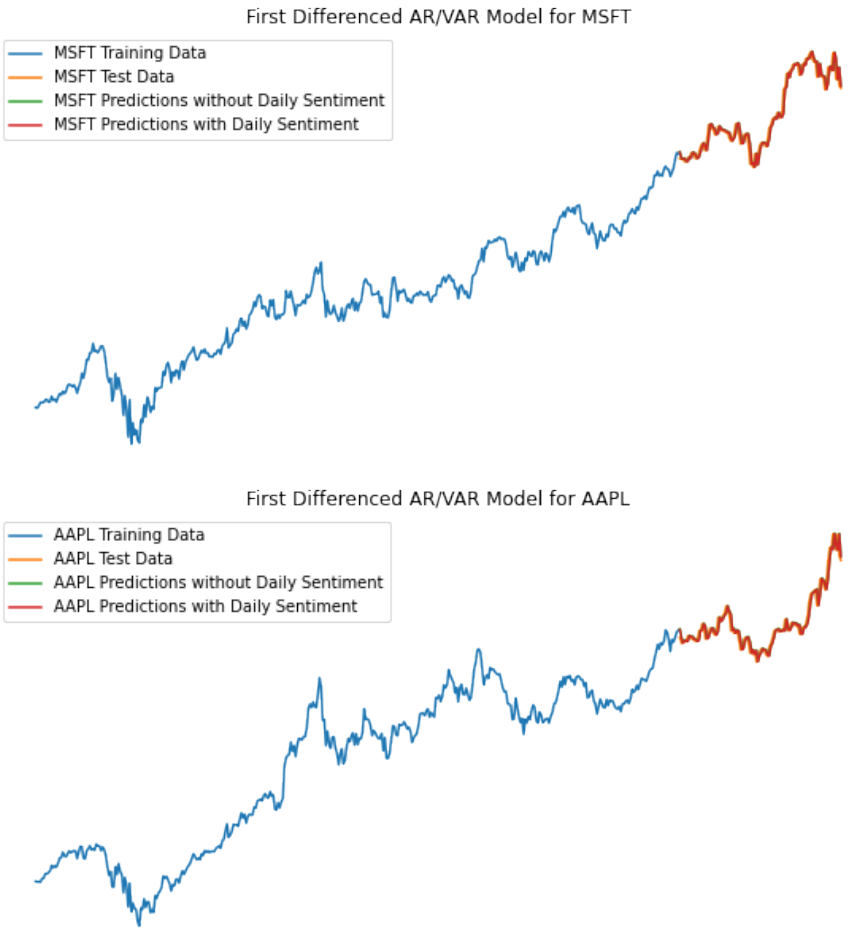
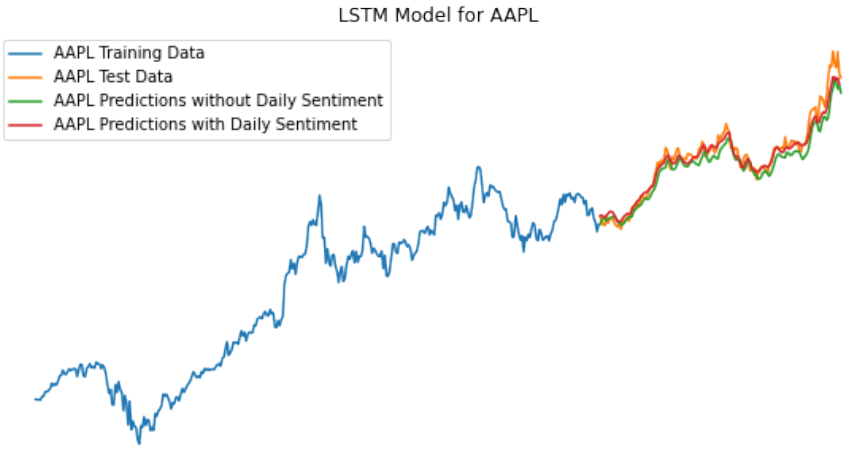



Figure 4.2: ARIMA/VAR Model Predictions

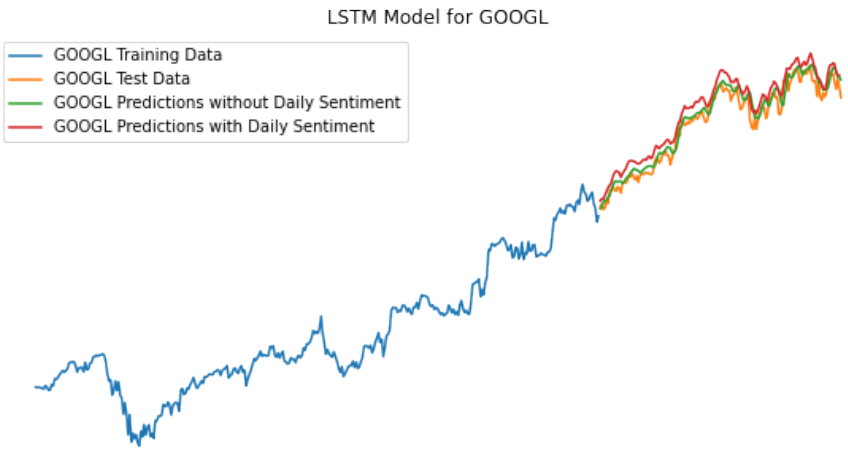
Table 4.2 presents and compares the performance metrics for LSTM model with and without the additional sentiment variable. For LSTM, three out of five stocks in our experiment experienced improved prediction performances with the sentiment variable. The only two stocks that did not experience an increased prediction power for LSTM model with the sentiment variable are GOOGL and TSLA.

| | Without Sentiment Variable | | With Sentiment Variable | |
|---|---|---|---|---|
| | **RMSE** | **MAPE** | **RMSE** | **MAPE** |
| **AAPL** | 3.99 | 0.020 | 3.07 | 0.015 |
| **AMZN** | 68.81 | 0.015 | 63.74 | 0.014 |
| **GOOGL** | 46.39 | 0.013 | 72.78 | 0.024 |
| **MSFT** | 15.30 | 0.047 | 10.08 | 0.030 |
| **TSLA** | 32.41 | 0.026 | 51.52 | 0.049 |

Table 4.2: Performances for LSTM model


LSTM Model for AAPL
AAPL Training Data
AAPL Test Data
AAPL Predictions without Daily Sentiment
AAPL Predictions with Daily Sentiment


LSTM Model for AMZN
AMZN Training Data
AMZN Test Data
AMZN Predictions without Daily Sentiment
AMZN Predictions with Daily Sentiment


LSTM Model for GOOGL
GOOGL Training Data
GOOGL Test Data
GOOGL Predictions without Daily Sentiment
GOOGL Predictions with Daily Sentiment

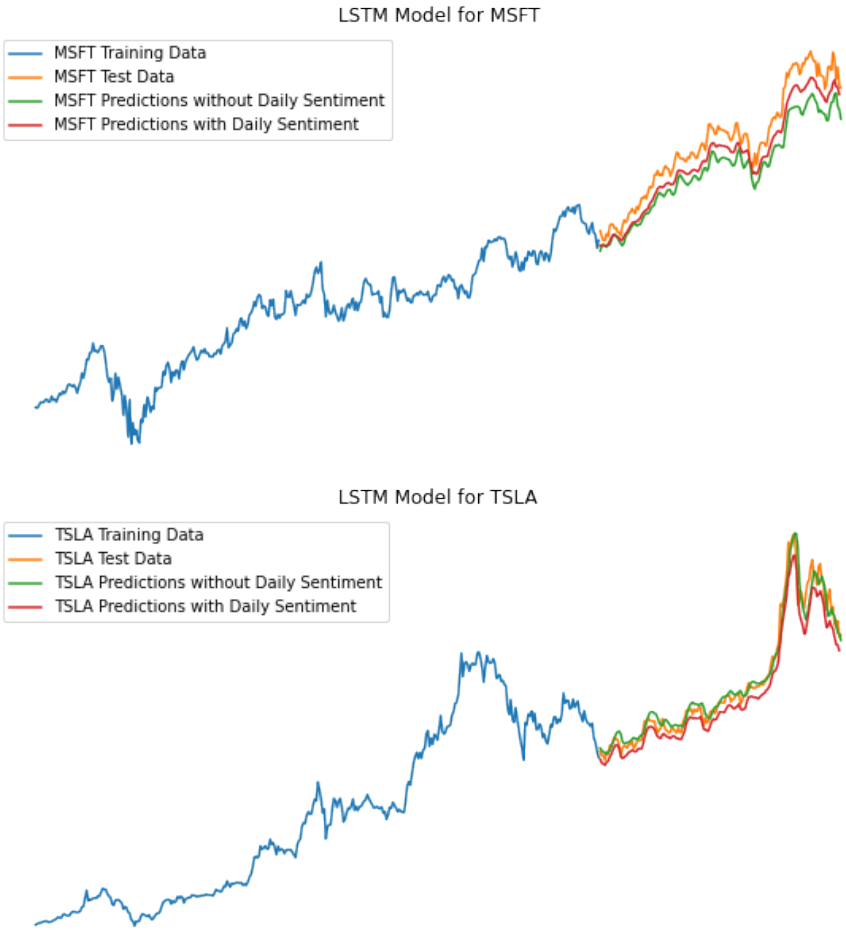



Figure 4.3: LSTM Model Predictions

Fig 4.3 shows the LSTM predictions for each stock.

# CHAPTER 5

# Discussion and Conclusion

This study proposed a novel stock price prediction framework that includes the daily Twitter sentiment variable derived from the text mining process of tweet contents. In this study, the daily sentiment trend for each stock seemingly corresponds to the close price trend for each stock to some extent. In some cases, the short-term price fluctuations match the short-term volatility in the daily sentiment score. To find out whether including the sentiment variable increase the accuracy for stock's close price prediction, the performances of first-differenced AR/VAR and LSTM models with and without this additional daily sentiment variable are compared and presented in the previous chapter. Results indicate that including the extra sentiment variable for LSTM models improved the prediction accuracy for most of the stock. However, the impact of including extra sentiment variable on first-differenced AR/VAR model were subtle.

The reasons that both AR and LSTM model fails to improve their prediction accuracy through incorporating the additional sentiment variable in some cases are complicated. Some possible explanations may be that the contents of tweets were sometimes misleading and might not completely reflect people's expectations, or our VADER model in the text mining process is not able to process each tweet 100% precisely and accurately. In addition, even if tweets can accurately present people's expectations towards the stock and our text mining process can accurately process and calculate sentiment scores for each tweet, the truth holds that people will not always take actions according to their own expectations. Furthermore, the stock price can be influenced by many other factors that were not included in this study,

such as variables from financial and macroeconomics aspects [PS13].

In general, therefore, including the daily sentiment variable may boost model performances for overall stock price prediction, even though the relationship between the daily sentiment variable and stock price remains vague and needs to be further studied. This stock price prediction framework can be applied to the scenario where market sentiment impacts the stock price dramatically and irrationally, such as the recent story of GameStop stock. Also, it can be adopted in the cases when researchers aim to account for the impact of market sentiment in stock price prediction. The stock price prediction framework that was introduced in this study is flexible and can be extended to many scenarios beyond including Twitter sentiment variable. It reveals how text data can be processed and included in the stock price prediction model in which the model is usually trained on structured numerical or categorical data. Based on the framework in this study, any textual data on the social media besides tweets can be converted into a new structured variable that can be utilized in the stock price prediction model in a similar manner.

Yet, this study was limited in several ways. First of all, the scope of our experiment was limited to the five major stocks in the US market (APPL, AMZN, GOOGL, MSFT, and TSLA) and two commonly-used time series models (ARIMA/VAR and LSTM). In addition, the current study has only examined tweets data from Twitter as text data, and the VADER model we utilized in the text mining process was specifically designed for the sentiment analysis in social media contents [HG14]. As a result, the framework introduced in this paper is only applicable for text data from social media instead of the text documents filed by the companies, such as annual information forms and prospectuses.

Also, this paper has thrown up many questions in need of further investigation. First, the causal relationship between daily Twitter sentiment score and stock price movement remains to be elucidated. Considerable more work will need to be done to determine whether the daily sentiment variable impact stock price movement significantly. More broadly, as this research focuses on retaining text data from social media in stock price prediction, further

studies, which take all text data, including financial news and word documents filed by the companies, can be undertaken.

# REFERENCES


[AAA14] Ayodele Adebiyi, Aderemi Adewumi, and Charles Ayo. "Stock price prediction using the ARIMA model." 03 2014.

[AC05] Basel MA Awartani and Valentina Corradi. "Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries." *International Journal of Forecasting*, **21**(1):167–183, 2005.

[BR14] C Narendra Babu and B Eswara Reddy. "Selected Indian stock predictions using a hybrid ARIMA-GARCH model." In *2014 International Conference on Advances in Electronics Computers and Communications*, pp. 1–6. IEEE, 2014.

[Fal07] Pegah Falinouss. "Stock trend prediction using news articles: a text mining approach.", 2007.

[FG18] Stefan Feuerriegel and Julius Gordon. "Long-term stock index forecasting based on text mining of regulatory disclosures." *Decision Support Systems*, **112**:88–97, 2018.

[FYL03] G Pui Cheong Fung, J Xu Yu, and Wai Lam. "Stock prediction: Integrating text mining approach using real-time news." In *2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.*, pp. 395–402. IEEE, 2003.

[Gra13] Alex Graves. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850*, 2013.

[HG14] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[HL20] Jia-Yen Huang and Jin-Hao Liu. "Using social media mining technology to improve stock price forecast accuracy." *Journal of Forecasting*, **39**(1):104–116, 2020.

[HLN13] Michael Hagenau, Michael Liebmann, and Dirk Neumann. "Automated news reading: Stock price prediction based on financial news using context-capturing features." *Decision Support Systems*, **55**(3):685–697, 2013.

[HR12] Michael D Hurd and Susann Rohwedder. "Stock price expectations and stock trading." Technical report, National Bureau of Economic Research, 2012.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." *Neural computation*, **9**(8):1735–1780, 1997.

[Hsu11] Chih-Ming Hsu. "A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming." *Expert Systems with Applications*, **38**(11):14026–14036, 2011.

[IN20] Mohammad Rafiqul Islam and Nguyet Nguyen. "Comparison of Financial Models for Stock Price Prediction." *Journal of Risk and Financial Management*, **13**(8):181, 2020.

[JYL20] Zhigang Jin, Yang Yang, and Yuhong Liu. "Stock closing price prediction based on sentiment analysis and LSTM." *Neural Computing and Applications*, **32**(13):9713–9729, 2020.

[Kot] Kevin Kotzé. "Vector autoregression models.".

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature*, **521**(7553):436–444, 2015.

[LSM14] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. "On the Importance of Text Analysis for Stock Price Prediction." In *LREC*, volume 2014, pp. 1170–1175, 2014.

[MSP10] Nitin Merh, Vinod P Saxena, and Kamal Raj Pardasani. "A comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting." *Business Intelligence Journal*, **3**(2):23–43, 2010.

[NEM10] Azadeh Nikfarjam, Ehsan Emadzadeh, and Saravanan Muthaiyah. "Text mining approaches for stock market prediction." In *2010 The 2nd international conference on computer and automation engineering (ICCAE)*, volume 4, pp. 256–260. IEEE, 2010.

[Oza77] T Ozaki. "On the order determination of ARIMA models." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **26**(3):290–301, 1977.

[PL05] Ping-Feng Pai and Chih-Sheng Lin. "A hybrid ARIMA and support vector machines model in stock price forecasting." *Omega*, **33**(6):497–505, 2005.

[PS13] Kanghee Park and Hyunjung Shin. "Stock price prediction based on a complex interrelation network of economic factors." *Engineering Applications of Artificial Intelligence*, **26**(5-6):1550–1561, 2013.

[SC09] Robert P. Schumaker and Hsinchun Chen. "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System." *ACM Trans. Inf. Syst.*, **27**(2), mar 2009.

[SLF16] Andrew Sun, Michael Lachanski, and Frank J Fabozzi. "Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction." *International Review of Financial Analysis*, **48**:272–281, 2016.

[STN18] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. "A comparison of ARIMA and LSTM in forecasting time series." In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1394–1401. IEEE, 2018.

[SVG17] Sreelekshmy Selvin, R Vinayakumar, EA Gopalakrishnan, Vijay Krishna Menon, and KP Soman. "Stock price prediction using LSTM, RNN and CNN-sliding window model." In *2017 international conference on advances in computing, communications and informatics (icacci)*, pp. 1643–1647. IEEE, 2017.

[Uro17] Siddhaling Urolagin. "Text mining of tweet for sentiment classification and association with stock prices." In *2017 International Conference on Computer and Applications (ICCA)*, pp. 384–388. IEEE, 2017.

[WWZ12] Ju-Jie Wang, Jian-Zhou Wang, Zhe-George Zhang, and Shu-Po Guo. "Stock index forecasting based on a hybrid model." *Omega*, **40**(6):758–766, 2012.