Brodie Harkins

BSc (Hons.) Computer Systems Honours Dissertation

Stock Market Predictions: The effects of External Factors

Supervised by Dr. Mehran Sharghi

Heriot-Watt University

School of Mathematical and Computer Sciences

September 2024

DECLARATION

I, Brodie Harkins, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: Brodie Harkins

Date: 14/10/2024

ABSTRACT

Within this study we analyse the correlation between public sentiment and stock prices. Looking at sentiment from X of certain stocks like Apple, Amazon & Microsoft, we can prove the effects of public sentiment either positively or negatively, by Incorporating previous studies and conducting quantitative research to gain correct and up to date information on financial stock markets and public sentiment data. When data has been pre-processed, cleaned and prepared for input, each algorithm is fed a dataset of public sentiment and financial data. Through this, comparative results are produced showing prediction performance and efficiency.

Going forward, this dissertation will aim to investigate the correlation between public sentiment and market sentiment and the impacts it causes on the predictive frameworks. It will aim to achieve this by using 2 different models and inputting the same data set of public sentiment and financial data into each model. Results will contain the graphs produced from each model and the total mean squared error (MSE) scores which will then be compared.

The results showed that the ARIMA model performed more efficiently than the LSTM model with an average Mean Squared Error score of 2.3 units when predicting our stock market data alongside the sentiment data.

ACKNOWLEDGEMENTS

I would like to thank my Family for their unwavering support throughout the years and the constant motivation they have given me throughout my time at university. I'd also like to thank my supervisor Dr. Mehran Sharghi for his support and advice throughout the creation of my dissertation.

Table Of Contents

Chapter 1: INTRODUCTION

Within this study we will look at the techniques that the stock market uses to correctly predict future prices for stocks and how outside factors like social media; Twitter (X), Instagram, Facebook, global economic concerns, national economic data and corporate financial performance. Of course, these predictions aren't always 100% correct but aim to have a high accuracy percentage with some reaching the high 90's almost 100% [1]. Forecasting techniques such as ARIMA and Long-Short Term Memory are all used to predict future stock prices. For us to understand how these algorithms operate, we first need to understand how the stock market works. Once we have covered the basics of the stock market, we will then be able to analyse the predictive models used.

The stock market is a platform where shares in publicly traded companies are bought and sold, two of the most well-known platforms are the New York Stock Exchange-NYSE and the London Stock Exchange-LSE. Each Exchange is made up of indexes, in the UK the FTSE100 is the most well-known, its constituents are the largest UK companies by market capitalization, IE Company Value. The largest stock exchange in the world, the NYSE, has 3 of the most well-known indexes, Dow Jones Industrial Average-DIJA which counts amongst its constituent's world-famous companies such as McDonalds, Walmart and JP Morgan Chase. The NSE is also home to the NASDAQ index. Apple and NVIDIA, the two largest companies in the world, are two of its most prominent constituents. Another index on the NYSE is the S & P 500, like the FTSE100 in the UK the S & P lists its constituents based on their market capitalization.

To buy stocks and shares an investment account is required. Once opened stocks, shares and other securities can be traded. Trades are done digitally, online and buyers and sellers are matched using computerized systems. There are many financial professionals working within the financial markets. Those investing with an investment company will normally be assigned an Investment Manager. The Investment Manager will carry out various tasks including research, investment advice, portfolio management and marketing their company's services.

Investment Analysts are often employed by investment banks, investment companies and other financial organizations. Their remit is around research, analysing potential investments and providing reports for their clients. They can provide a view on a share or security as to whether their clients should consider buying, referred to as bullish, or selling, known as bearish.

Forming a view as to whether a stock or security should be bought or sold can depend on many factors and there is a myriad of financial theories that professional and retail investors can use to help them decide. Perhaps the most well-known is the Dow theory, a financial theory developed by Charles Down, founder of the Dow Jones index. Dow Theory works on the principle that the stock market follows trends and if the investor can identify the trend, they can accurately predict the direction of the market. In the early years of making predictions on the stock market, Robert Rhea, an investor in the stock market in the 1930's,

took the Dow Theory and was able to turn it into a practical indicator for either buying or selling a stock.

Robert Rhea wasn't the only individual who looked to try and predict these markets. Edson Gould was the most accurate forecaster, who used charts, market psychology and including the Senti-Meter (The DJIA divided by the dividends per share of the companies). John Magee the founder and the creator of technical analysis (Also wrote a book "Technical Analysis of Stock trends - 1948"). Magee was the first to trade solely using stock price and its pattern on historical charts.

The use of AI (Artificial Intelligence) and Machine Learning algorithms (ML) such as Auto-regression, classifier and support vector machine (SVM) have brought a new understanding of how forecasting works and have allowed stockbrokers to make better decisions on whether to be bullish or bearish. This paper will cover algorithms using Machine Learning, Long-Short-Term memory (LSTM) and AutoRegressive Integrated Moving Average Model.

1.1 Motivation

With its constant movements and unpredictability, the stock market presents a difficult challenge to those who are new or experienced in the financial industry. Take, for example GameStop, in 2021 institutional investors were shocked when a short squeeze on GME stock was started by regular retail investors resulting in a 700% rise in the share price, in February of 2021 the share price fell 80% as investors enthusiasm waned[20]. Despite this difficult challenge, the financial industry operates with mechanisms that, if understood, can be leveraged to make well-informed decisions. As financial markets evolve over time, so do the tools behind making these predictions. The motivation behind this research is to better understand and improve the accuracy of predicting stock market behavior.

In the past, traditional methods of forecasting, such as those developed by Robert Rhea, Edson Gould and John Magee have laid out the foundations for modern predictive analysis. Today, advancements in artificial in AI (Artificial Intelligence) and ML (Machine Learning) provide us with opportunities to push and stretch beyond the boundaries of these advancements and improve accuracy and efficiency.

With the rise of social media platforms, such as, X, Instagram and Facebook, along with the availability of real-time financial data from sources such as yahoo finance, additional layers of complexity have been added to predicting stock market trends. Market sentiment, corporate performance and global events are now analysed in conjunction with historical market data, making accurate predictions a complex challenge. The goal of this research is to investigate how modern machine learning algorithms, such as Long-Short Term Memory (LSTM) networks, Auto-Regressive Integrated Moving Average (ARIMA) algorithms, can enhance our ability to predict market movements more effectively compared to traditional approaches.

By conducting a thorough evaluation of these machine learning techniques, this study aims to identify methods that provide the highest accuracy in predictions. Ultimately, this research could contribute to the advancement of investment strategies, offering investors a more robust approach for making financial decisions. The motivation for this work is not solely academic but also practical. Aiming to empower investors, improve financial literacy and contribute to a deeper understanding of the forces driving market changes.

1.2 Aim and Objectives

The primary aim of this research is to evaluate the accuracy and reliability of stock market predictions by leveraging advanced machine learning algorithms and integrating diverse data sources, including social media sentiment and traditional financial data. The aim is to find new or better tools that can help analysts and investors make better, more informed decisions in an environment that is increasingly complex. To achieve this aim, the study has the following objectives:

- ***Compare machine learning algorithms performance***: This objective will focus on the analyzation of Machine Learning algorithms that are used to forecast share prices. Algorithms like Long-Short Term Memory (LSTM) and AutoRegressive Integrated Moving Average (ARIMA) in predicting stock market trends. By analysing the graphs produced from each algorithm and their MSE score we can get an idea of how each algorithm performed.
- ***Analyzing the Impact of Social Media Sentiment on Stock Market Behavior***: social media platforms such as Facebook, Instagram and Twitter (X) produce huge amounts of data that can influence the stock markets. The objective aims to use sentiment analysis techniques and perform them on social media datasets with traditional financial data sets to understand the role of public sentiment in influencing market movements. This will be done by one model getting fed public sentiment data and financial data and another model will be given financial data only, the results will show the impacts of social media sentiment
- ***Develop a predictive framework that combines multiple data sources***: This research aims to develop a predictive framework that uses a diverse range of data sources, including historical stock data, corporate financial performance and social media sentiment. By combining all these different types of data, the study seeks to improve prediction accuracy and provide a more holistic view of market dynamics.
- ***To evaluate the predictive framework using real-world, publicly available data:*** The study will utilize real-world data from sources such as Kaggle to test and validate the predictive framework. This objective aims to ensure that the proposed model is applicable and relevant to actual market conditions, providing investors with a reliable tool for decision making. Evaluation will include looking at performance of the predictive model such as Mean Squared Error (MSE) scores to help analyse the effectiveness of the models.

1.4 Contributions

This thesis makes several key contributions towards the field of market prediction, specifically focusing on the application of machine learning algorithms for enhanced prediction accuracy. The primary contributions are as follows:

- ***Comparative Analysis of Machine Learning Algorithms*:** This research conducts an analysis of different machine learning algorithms, including Long Short-Term Memory (LSTM) networks and ARIMA models. By comparing these techniques, the study identifies the strengths and weaknesses of each approach in predicting stock market behavior, providing valuable insights into their practical applications.
- ***Integration of social media and Real-Time Data*:** This study will integrate social media data, such as sentiment analysis from platforms like Twitter, Facebook and Instagram with traditional financial data. This integration allows for a more in-depth analysis of market trends, highlighting the impact of social media on stock price movements and providing a more holistic approach to market prediction.
- ***Development of a Predictive Framework*:** A key contribution of this research is the development of a predictive framework that combines historical stock data, corporate financial performance and social media sentiment analysis. The aim of this framework is to enhance the accuracy of stock market predictions by leveraging multiple data sources and advanced machine learning techniques.
- ***Practical Evaluation Using Real-World Data*:** This study will utilize real-world data, publicly available data from sources such as Finazon and MarketStack to test and evaluate the performance of the machine learning algorithms. This practical evaluation ensures that the findings are relevant and applicable to real-life scenarios, providing a reliable basis for investors and financial analysts to make informed decisions.
- ***Contribution to Financial Literacy and Investment Strategies:*** By providing insights into the effectiveness of different predictive algorithms and the influence of social media on market trends, this research contributes to improving financial literacy. It also offers practical recommendations for developing more effective investment strategies, ultimately helping investors to make better-informed decisions.

Through these contributions, this thesis aims to advance the field of financial market predictions by demonstrating the potential of machine learning algorithms to improve prediction accuracy, integrate diverse data sources and offer practical tools for investors and analysts.

1.5 Organization

This dissertation is organized as:

- **Chapter 1**: Introducing the research background, motivation and objective of this study, as well as key contributions

- **Chapter 2**: This Chapter provides a full literature review, covering traditional methods of stock market prediction, which Machine Learning algorithms are going to be tested, how these algorithms work and how testing of each algorithm will be carried out. Case studies will also be used to prove the correlation between the public sentiment's influence on the stock market and related works of the ARIMA & LSTM models.

- **Chapter 3**: This chapter will look at how we have set up our environment, starting with what language we have used to code our project alongside implementing our datasets and conducting sentiment analysis and feature engineering.

- **Chapter 4:** This chapter will look at analysis of our datasets ensuring that any results produced from our tests are significant. This will also analyse the graphs of our stock prices over time, looking at when certain events happened which caused the stock price to drop or increase.

- **Chapter 5:** This section will look at the results produced from each algorithm. Graphs are produced and analysed, and the mean-squared-error is compared.

- **Chapter 6:** This chapter will conclude our findings including a general discussion on what was successful from this experiment and what future research could be done to improve on the findings.

Chapter 2. Literature Review

In this literature review we will aim to give an overview of existing research. This chapter will focus on primarily three principal areas; traditional Methods of stock prediction, the evolution of machine learning techniques and the influence of social media sentiment on financial markets and predictions.

2.1 Traditional Stock Market Prediction Approaches

This section will investigate traditional methods used for stock market predictions. This will look at how the algorithms operated and their shortcomings. The University of Denver computer science and Engineering department [2] had a look at traditional methods that were used for stock market prediction. Some of these methods were: Auto Regressive (AR), Auto-Regressive Moving Average (ARMA) and Auto Regressive Integrated Moving Average (ARIMA).

2.2 Auto Regressive Model (AR)

Standard AR models tend to use polynomial time, if a problem an algorithm is trying to solve can solve it in polynomial time it means that the number of steps taken to solve set problem grows at a reasonable rate so, n, n^2, n^3,.....n^whole number.

For example, if you have a small problem where n is 5 and the number of steps is *n^squared - (5^2 = 25)*. For an algorithm that is manageable, even if *n* becomes larger, the number of steps will grow at a reasonable rate.

Given the information above, the assumption can be made, data that is stationary is great for this model due to data not being dynamic which for working in stock markets and how it can be difficult with large quantities of data required to be processed at fast speeds, makes meeting the requirements difficult to meet given the environment that these algorithms will be included in. [3]

2.3 Auto-Regressive Moving Average (ARMA)

This model combines two simple algorithms together, Auto Regressive and Moving Average. This model begins by taking in past data like the Auto-Regressive model but will also implement past errors and account for them when constructing future predictions [4].

The ARMA model uses the equation below to achieve predictions:

$$y_t = c + \varphi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

Here's what the equation aims to achieve:

- y_t : This is the value of time series at time t, the current value we are trying to model or predict.
- c : This is the constant term. It acts like a baseline value for the time series, meaning the model will start at this value before adding in the effects of previous values.
- $\varphi_1 y_{t-1}$: This is the AutoRegressive (AR) part of the model. The term φ_1 is a coefficient that shows the influence of the previous value of the time series (y_{t-1}) on the current value (y_t). This tells us how much of the past values influence the current one. If y_{t-1} is
-
- positive, it means that if the previous value y_t was high, it will push yt higher and lower if y_{t-1} was low.
- $\theta_1 \epsilon_{t-1}$: This is the Moving Average (MA) part of the model. The term θ_1 measures the impact of the previous error term (ϵ_{t-1}) on the current value y_t. The error term will represent the previous error or "shock" like any unexpected fluctuations. If θ_1 is significant, it means that the past errors still influence the current value.
- ϵ_t : This is the current error term. It represents the random noise or unexpected changes that affect y_t at the time **t.** This part accounts for the randomness in the time series that the model cannot explain with past values or past errors.

In summary, the equation models the current value y_t as a combination of

- A constant Baseline (c)
- A weighted influence of the past value (y_t).
- The impact of past random shocks or errors ($\varphi_1 y_{t-1}$).
- The current random noise or error (ϵ_t)

This ARMA model captures patterns in a time series where both past values and past errors influence the present, making it a powerful tool for forecasting.

2.4 Auto-Regressive Integrated Average (ARIMA)

Auto Regression Integrated Moving Average (ARIMA) model uses the traditional model of the ARMA framework allowing it to accommodate non-stationary data. Non-stationary data being data that has statistical properties that change over time, such as its meaning, variance or autocorrelation. Meaning the data behaves in such a way that it is unstable or unpredictable. When processing non-Stationary data it is converted to stationary data.

Mathematically ARIMA model is expressed ARIMA (p, d, q):

p: The number of lagged observations included in the model (The order of the AutoRegressive part)

d: The number of times the data need to be differenced to achieve stationarity (Order of Integration)

q: The number of lagged forecast errors included in the model (The order of the Moving Average part)

2.4.1 Components of the ARIMA Model

The ARIMA model is composed of 3 main components: AutoRegressive (AR), Integrated(I) and Moving Average (MA). Each component plays its own part in capturing the structure of temporal data. [9]

2.4.2 Auto Regressive (AR) Component

The first component of the ARIMA model is Auto Regressive (AR). It is denoted as *AR(p),* it uses past values of the time series to predict future values. The AR component assumes a linear relationship between the current value and its *p* previous values.

2.4.3 Integrated (I) Component

The Integrated component – denoted by d – addresses the issue of non-stationarity in the time series. As mentioned above, non-stationary data often exhibit trends or changing variance over time, making direct modelling inappropriate. To convert from non-stationary data to stationary, the modelling applies Differencing, which involves the process of subtracting the previous observation from the current observation. This process will be performed d times until the series becomes stationary. The difference equation looks like this: [11]

$$d = 0: \mathbf{yt} = y_t - Yt\text{-}1$$

$$d = 1: \mathbf{yt} = \mathbf{y_t - y_{t-1}}$$

$$d = 2: \mathbf{y_t} = (\mathbf{y_t - y_{t-1}}) - (\mathbf{y_{t-1} - y_{t-2}}) = \mathbf{y_t} - 2\,\mathbf{y_{t-1}} + \mathbf{y_{t-1}}$$

For *d = 1:* yt is the difference series. As stated above further differencing can be performed to stabilize the mean and remove trends or seasonality. For *d* = 2, this can be described as the "first difference-of-the-first difference"(Nua, Robert 2023 [11], Introduction to ARIMA: nonseason models)

2.4.4 Moving Average (MA) Component

The Moving Average part of the model, denoted as MA(q), incorporates past forecast errors into the prediction of future values. The MA component models the dependency between current observation and a combination of q past error terms. Expressed as:

$$Xt = \mu + \omega t + \theta 1 \omega t - 1 + \theta\, 2\omega t - 2 + \cdots + \theta q \omega t - q$$

Where $\theta_1, \ , \theta_2 \ \ , \theta_q$ are moving average coefficients, and ωt represents the random error at time "t". The MA component accounts for the impact of shocks or unexpected fluctuations in the data, smoothing out noise to improve forecasting accuracy.

2.4.5 ARIMA Model Specification

The specification of the ARIMA Model is ARIMA (*p, d, q).* It involves selecting the appropriate values for *p, d* and *q.* This process is aided by Autocorrelation function (ACF) [7] and the Partial Autocorrelation function (PACF) [8]. By using these tools, they can help identify the degree of difference needed to achieve stationarity and the appropriate orders of the AR and MA components. Once each part of the model has been specified, the parameters are then estimated using methods such as Maximum Likelihood [9] or Least Squares approach [10].

2.5 Machine Learning Approaches in Stock Market

Machine Learning has introduced new and advanced ways of predicting stock markets and analysing their behaviour. New techniques such as regression analysis, support vector machines and Neural networks have all been applied to analyse large amounts of historical data to look for trends and identify patterns.

2.5.1 Long-Short Term Memory Networks (LSTM)

Long-short Term Memory Networks operate best at tasks that require prediction and capturing long-term dependencies. Long-term dependencies are when the output of the model relies on 3 different gates all with their own process.

2.5.2 Forget Gate

The forget gate works as it says, it will simply "forget" irrelevant data that isn't needed for the present unit in a LSTM. This is used to make sure the LSTM model performs efficiently, not holding onto data that is deemed to be less important or is no longer required for the LSTM

2.5.3 Input Gate

When it comes to adding new information into the LSTM model, the input gate does exactly that. It will add in new information, passing it through a sigmoid function and the tanh function which will take in the vector of Xt and Ht-1. Through this, it will make sure the information in this gate is relevant and is not useless. Mathematically, the input gate
i_t is calculated as follows:

$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Where:

- [σ]- Sigmoid function, squashes input values between 0 and 1
- [Wi ,bi] - weight & biases associated with the input gate
- [$ht-1$] - hidden state from previous step/result
- [xt] - current input

The sigmoid function works by ensuring that i_t outputs values between 0(reject) and 1(accept), going through the new information and selecting which parts are to be kept.

2.5.4 Output Gate

The output gate will work to develop a hidden state which is then to be used by the input gate when it receives new information. How this hidden state is generated is by determining which parts of the cell state should be output as the hidden state ht for the current time step. This effectively controls what information shall be carried forward to the next layer or time step. Mathematically, this is defined by:

$$ot=\sigma\ (Wo \cdot [ht-1, xt] + bo)$$

Where:

- W_o B_0 - weights and Biases
    - Like the input gate, sigmoid function will limit O_t between 0 (forget) and 1(keep)

Once this process has been completed, it will pass the determined value of O_t to a filter, which will generate the hidden state.

2.6 Related Works of LSTM & ARIMA Models

Some related work of others using the ARIMA model have shown to use a different variation of the ARIMA model. This study showcases the method they have followed which used an ARIMA Model of (3, 1, 1) after conducting ACF and PACF testing. It also shows the stock prices between 2000 to late 2010 [21]. From the graph they have used you can see a massive drop in price in 2008/09 which was when the housing market crashed causing a massive shift in the economic environment.

When it comes to the LSTM model which can be used in many sectors, not just finance. It has proven to be accurate in situations where lives are at stake. This article by BytePlus [22] which covers case studies about the LSTM model, shows that it achieved 87% accuracy in predicting diabetes onset and could identify high-risk patients 2-3 years before traditional screening methods.

Chapter 3: Methodology

To achieve the aims and objectives we originally set out to accomplish, we must make sure that the data we implement into each algorithm is up-to-date and contains relevant information. Ensuring we have Open, Close, High, Low and adjusted closing values will tell us how the stock has performed in the past, allowing us to better predict the future prices. When building our predictive framework, we will ensure that our data is stationary, and our P-values are below 0.05 by analysing ADF and KPSS test results. From these results, we can then conduct trend analysis. This analysis will provide us the best model to use for our predictions and prove the correlation between stock prices and public sentiment.

3.1 Environment

When conducting the testing of each Machine Learning algorithm, it is top priority that our environment is up to date using the most recent technologies to ensure that the results produced are accurate.

3.2 Jupyter Notebook

Using Jupyter Notebook which combines code, visualizations, section for text to be implemented and other media types that can be used into a single document. This will allow me to modularize each Machine Learning algorithm into their own document, each containing their own results and findings. Each document will display graphs based on the dataset given.

3.3 Datasets

When conducting analysis of each dataset, we will use the dataset from Kaggle [12] which includes:

- Open prices
- Close prices
- High Price
- Low price
- Adjusted Close

From this dataset, 3 companies will be used to analyse and test each algorithm. To ensure that statistics can't be skewed or unreliable, all 3 will be used across each Machine Learning algorithm. The 3 main stocks are Amazon, Microsoft and Apple.

3.4 Programming Language

Python will be used to conduct this analysis due to its readability, future proofing and how commonly used it is within the field of Machine Learning.

3.5 Data Preprocessing

The first step to preparing our data for inputting into our Machine Learning algorithms is to clean and obtain the correct data needed. Upon opening the dataset, it contains:

- Date
- Open
- High
- Low
- Close
- Adj Close
- Volume
- Stock Name

Using the stock name, we will not need all columns within this dataset. For this experiment Adj Close will be used to calculate the stock returns for the day.

The first step will now be to adjust the dataset and make sure each column is appropriate for use. The "Date" column contains the date itself – YYYY/mm/dd and hh/mm/s - to format this date correctly, removing the time isn't needed and merging datasets will allow us to merge on the date.

Since we will be calculating the stock returns later, removing all other columns and keeping Date, adjusted close and stock name will help keep our dataset organized and clear with what data will be used.

3.6 Sentiment Data Preprocessing

Our sentiment dataset contains 4 columns or features which are:

```
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Date          5056 non-null   object
 1   Tweet         5056 non-null   object
 2   Stock Name    5056 non-null   object
 3   Company Name  5056 non-null   object
dtypes: object(4)
```

Figure 1: Attributes of Sentiment Dataset

We will be looking into 3 main companies which are Apple, Amazon and Microsoft. When preparing our data, the stock name will be used to find all tweets regarding Apple, Amazon or Microsoft, each will be split into their own subset i.e., Microsoft, Apple & Amazon Sentiment data.

Upon initial inspection of the sentiment dataset, we can use 'Tweet' itself to analyse sentiment. To get the emotional tone behind the body of text we can use a popular tool called VADER (Valence Aware Dictionary and Sentiment Reasoner). This is a lexicon based and a rule-based sentiment analysis tool that is specifically designed for texts that contain informal language like social media posts and reviews.

Due to the length of most tweets, twitter (X) has a limit on how many characters you can have per tweet, which makes VADER a useful tool as it analyses short pieces of text such as tweets, reviews or any user-generated content that may contain emojis, slang and abbreviations.

VADER tool will apply a sentiment score on each word before going back and combining this score into an overall score for each in entire text, which consists of four components.

- Positive (pos): Proportion of text that expresses a positive sentiment
- Negative (neg): Proportion of text that expresses a negative sentiment
- Neutral (neu): Proportion of text that expresses a neutral sentiment
- Compound: The aggregated sentiment scores that ranges from –1 (Extremely Negative) to +1 (extremely positive)

The compound score is the most important and is computed between the values of –1 to +1. It summarizes the sentiment of the text.

If compound score is > 0.05: Positive sentiment

If compound score is < -0.05: Negative sentiment

If compound score is between 0.05 & -0.05: Neutral sentiment.

This approach was used by both Amazon and Microsoft sentiment. Once sentiment has been compounded into a total sentiment score, we can then merge each dataset to then compare the adjusted closing value on the day to the sentiment score on that day.

When merging both datasets we will order our Date column to preserve the temporal sequence which is crucial for LSTM models.

3.7 Stationarity Analysis

The ARIMA model or Auto-Regressive Integrated Moving Average is a statical model best used for analyzing and forecasting time series data. Some key considerations when implementing the model is to make sure our data is stationary, if not we will use the "I" (Integrated) part of the ARIMA model to achieve stationarity.

3.8 Autocorrelation

Autocorrelation is used to help tell us if the current time series data is predictable based on its previous values. Typically, if the value of the autocorrelation is in the higher quarter, then it indicates that the time series data is highly predictable based on previous values and if low or near-zero then it suggests that the data points are more independent of each other.

By using Autocorrelation, it is a key step in telling us more about the patterns or repetitive behaviours in the data.

3.9 Partial-Autocorrelation

Partial Autocorrelation function quantifies the direct relationship between a time series and its lagged values, with the effects of shorter lags removed.

3.10 Augmented Dickey-Fuller Test

To ensure that our dataset used is stationary, we can use the Augmented Dick-Fuller test. The test is used to check if a time series is stationary – that is, whether its properties like mean and variance remain constant over time. This is important because many time series models will automatically assume stationarity. [7].

- Unit Root Concept:

  A time series with a unit root has a random walk behavior, meaning that it can wander without a fixed path. ADF test helps determine if this random wandering exists.

- Testing Process: The test estimates a regression model that looks like this:

$$\Delta Y_t = \alpha Y_{t-1} + \theta_1 \Delta Y_{t-1} + \theta_2 \Delta Y_{t-2} + \cdots + \theta_k \Delta Y_{t-k} + \mu_t$$

$$or$$

$$\Delta \boldsymbol{Y_t} = \boldsymbol{\alpha Y_{t-1}} + \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\theta_k \Delta Y_{t-k}} + \boldsymbol{\mu_t}$$

Here:

- Δy_t - is the change in the series at time t.
- α - is a constant
- β_t - captures any trend in the data
- γ - is the key coefficient; testing if $\gamma = 0$ tells us if the series has a unit root (non-stationary) or not.
- $\sum_{i=1}^{\rho} \delta_i \, \Delta_{yt-i}$ - are added to correct for any correlations in the errors

Hypotheses

- **Null Hypotheses (H_0):** The time series has a unit root (non-stationary)
- **Alternative Hypothesis (H_1):** The time series is stationary (No unit root)

By doing this, we can ensure that any results obtained from the model are valid. Using non-stationary data in models can lead to false conclusions (Spurious results). ADF testing helps us know if the data needs to be transformed i.e., Differencing.

This will also improve the accuracy of our forecasting models and the validity of the conclusions drawn from them.

3.11 Kwiatkowski-Phillips-Schmidt-Shin Test

The KPSS test founded by Kwiatkowski, Phillips, Schmidt & Shin, is another method which is used to test our data to check for stationarity in time series. It works in a complementary way compared to ADF testing:

- **Null Hypothesis**:
  The KPSS test assumes that the time series is stationary (No unit root)
- **Alternative Hypothesis:**
  The alternative is that the series is non-stationary
- **Test Produce & Equation**
  The KPSS is based on linear regression. It breaks up a series into 3 main components: A deterministic trend (β_t) , a random walk (r_t) and a stationary error (ε_t) with the regression equation:

$$x_t = r_t + \beta_t + \varepsilon_t$$

A high KPSS statistic which exceeds critical values suggests that the residuals vary too much for a time series to be considered stationary, leading to the rejection of the null hypothesis of stationarity.

Chapter 4: Conducting Analysis

4.1 Trend Analysis

Upon completing data cleaning and preprocessing, we then start plotting our financial data and analysing trends that could be a direct correlation of public sentiment. It will also tell us how strong a stock is if there are large gaps between the highs and lows (High Volatility).

Figure 2: Closing Price AAPL

Analysing the graph above which displays the Apple Closing price from 2020 to 2023, we analyse that the closing price drifts upwards from the lower levels in 2020 to the higher levels in 2024. We can see that around the beginning of 2020 the closing price hits the lowest point in the chart. This was due to the pandemic in the beginning of 2020 in March that year.

After this happens to the stock and we begin to learn more about the pandemic, this causes overall sentiment at that time to be more hopeful causing the stock to rise. This is backed up by analysing the VIX (Volatility Index) index during that time. The overall market expectations of the S&P 500 [15] and during that time we know that the VIX index was at an all-time high at a score of 82.69 which would explain the sudden dip.

4.2 ADF & KPSS Analysis

Before conducting any testing of each algorithm, we must make sure our data is stationary and has a P value < 0.05. Both conditions are used when dealing with hypothesis tests and time series data.

P-Value

A P value less than 0.05 indicates that there is less than a 5% probability that the observed results occurred by chance if the null hypothesis were true. By using this threshold, we can use

it to determine whether the findings are statically significant, allowing us to reject the null hypothesis.

Although the 0.05 level is somewhat arbitrary, it has become a conventional standard in many fields to balance the risk of type 1 errors (False-Positives) against statistical power.

Stationarity

As our data is classed as time series data, stationarity means that the statistical properties – such as mean, variance and autocorrelation – remain constant over time. Many analytical methods assume stationarity because it simplifies the modelling process and ensures that relationships identified in the data are stable.

Non-stationary data can lead to misleading inferences (like spurious regression results) because changes in the data's underlying structure, due to trends, seasonality or other factors, can distort statistical tests and p values.

When our data is stationary, the assumptions behind many models (e.g., ARIMA models) hold true, which makes the model's predictions and statistical tests more reliable.

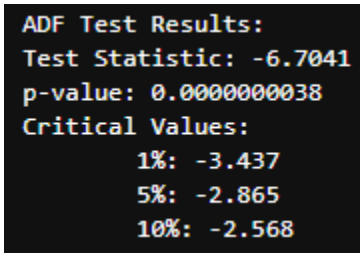


Figure 2.0: ADF Test Results

Analysing the ADF results we can say there is strong evidence against the presence of a unit root. With a test statistic of "-6.7041" well below the 1% critical value of "-3.437" and an extremely low p-value 3.8x10^-9, we can reject the null hypothesis of non-stationarity. This indicates that the time series data is stationary.

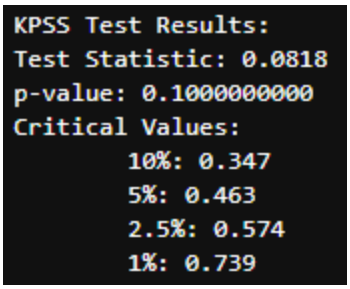


Figure 2.1: KPSS Test results

In contrast, the KPSS test results above, which assumes stationarity as its null hypothesis, yields a test statistic of 0.0818 that is far below all its critical values, with the lowest being 0.347 at the 10% level and a p-value of 0.10. Consequently, we fail to reject the null hypothesis, which further supports the conclusion that the series is stationary.

With both tests complimenting each other with their findings – ADF rejecting non-stationarity and KPSS acceptance of stationarity provide robust evidence that the statical properties such as mean and variance of the series remain constant over time. Due to the results above, we can thereby prove the enhancing the reliability of subsequent time series analyses.

4.3 Feature Engineering – Stock & Sentiment

Once sentiment analysis has been complete, we will keep our compounded value which will give us the average sentiment of each tweet. We can plot and compare to show correlations between public sentiment and stock market prices.

The first step to combining our data frames together is to merge both sentiment data and Close prices. This merging will use each data frames "Date" column, matching each day's sentiment and closing price. Using this we can then try and prove correlation between each data point.

With the use of feature engineering, we can now analyse our data frame and see that we have now successfully combined our data frames and can now begin our calculations for correlation.

4.4 Correlation (Sentiment vs Stock Price)

In most academic and applied finance, researchers often want to find out which market sentiment directly correlates to the performance of a stock price.

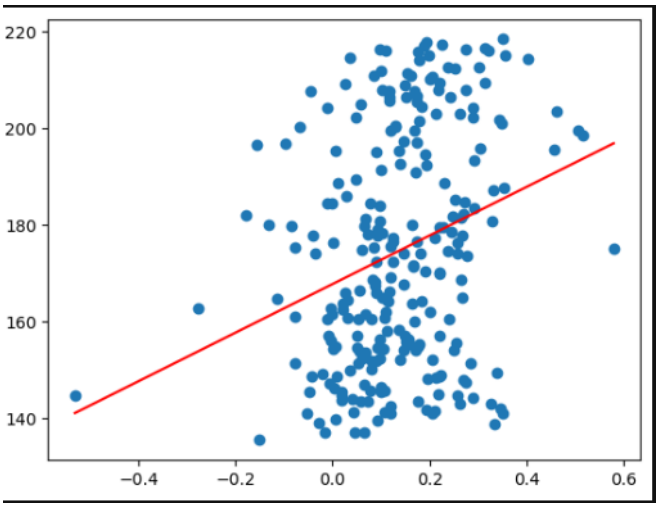


Figure 2.3: Scatter Plot (Correlation between Closing price and Sentiment Data)

By analysing the graph above we can notice a few things that can tell us about the correlation between the closing price and the sentiment data:

- **Clearly Positive Slope**

  The red regression line slops more steeply upward compared to the previous plot, indicating a stronger positive association between sentiment and stock returns. In other words, as sentiment scores increase, the corresponding returns generally appear higher.

- **Moderate-to-strong Relationship**

  Visually, the data cluster somewhat more tightly around the regression line, suggesting that sentiment could be a more reliable indicator or correlate of returns in this dataset. While there are outliers scattered further away from the center, this shows that even though we have a strong sign of correlation, it doesn't guarantee that it is perfect.

- **Potential Predictive Values**

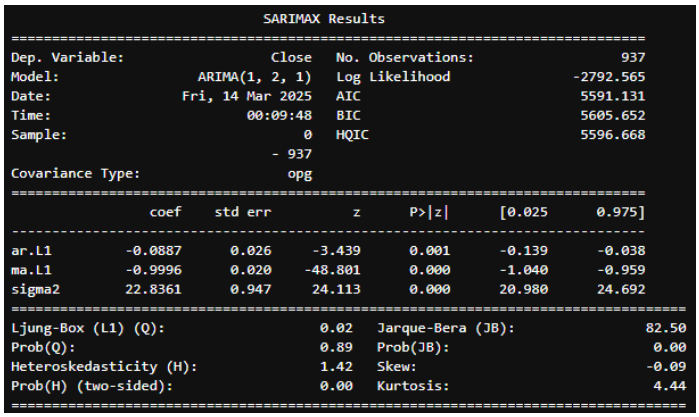
  The more pronounced slope suggests that sentiment might have stronger explanatory or predictive value for stock returns. In practice, this could motivate further analysis, such as calculating the exact correlation coefficient or incorporating sentiment as a feature in forecasting models.

- **Importance of Additional Factors**

Even though we have proven there is a strong case for correlation between sentiment data and stock performance, many other variables such as market conditions, company fundamentals or economic events can all affect stock returns. While the influence of sentiment is prevalent, it is important to highlight that other factors should be taken into account here.

4.5 ARIMA Model Implementation

Now that we have proven there is a correlation between stock prices and public sentiment, we can move onto implementing our Machine Learning models. Starting with the ARIMA model which will need to be fitted using the data we have obtained from our dataset.



SARIMAX Results

Dep. Variable:		Close	No. Observations:			937
Model:		ARIMA(1, 2, 1)	Log Likelihood			-2792.565
Date:		Fri, 14 Mar 2025	AIC			5591.131
Time:		00:09:48	BIC			5605.652
Sample:		0	HQIC			5596.668
		- 937				
Covariance Type:		opg				
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0887	0.026	-3.439	0.001	-0.139	-0.038
ma.L1	-0.9996	0.020	-48.801	0.000	-1.040	-0.959
sigma2	22.8361	0.947	24.113	0.000	20.980	24.692
Ljung-Box (L1) (Q):		0.02	Jarque-Bera (JB):			82.50
Prob(Q):		0.89	Prob(JB):			0.00
Heteroskedasticity (H):		1.42	Skew:			-0.09
Prob(H) (two-sided):		0.00	Kurtosis:			4.44

Figure 2.4: SARIMAX Results

By analysing the above results when fitting our ARIMA model, we can note a few things:

- **Model Specification**
  By Changing the number of Auto-regressive terms we further simplify our model. Before we had the ARIMA model setup as (5, 1, 0) & (3, 2, 1). Each proved to be statistically insignificant so by simplifying our model it allows us to get a more accurate and significant result.

  - **AR (1)** This is shown by including a single autoregressive term.

  - **I (2)** - the model does uses second-order differencing, this shows that we have differenced our data to achieve stationarity.
  - **MA (1)** The model will include a single moving average term which accounts for dependency between the current value and the residual value from the previous time step.
- **Model Fit**
  - **Log likelihood:** -2792.565

        - This measures how well our model fits the data. The value here is low but is lower than other analysis of other models' specifications which indicates a marginally worst fit.
    - **AIC (Akaike Information Criterion):** 5591.131
    - **BIC (Bayesian Information Criterion):** 5605.652
    - **HQIC (Hannan-Quinn Information Criterion):** 5596.668
        - The above criteria have low values which indicate a better trade-off between model fit and complexity. The values are relatively high, which could suggest the mode is not the best fit for the data or additional terms (e.g., seasonal components)
- **Coefficients**
    - **Const:** -0.0887
        - The constant time represents the mean of the time series when the MA term is 1. The p-value (0.001) is greater than 0.05 which indicates that the term is not statistically significant.
    - **Ma.L1: -**0.9996
        - The Coefficient for the moving average term is positive and statistically significant, suggesting that the MA term is adding some meaningful information
    - **Sigma2:** 22.8361
        - This is the variance of the residuals, which measures the volatility of the errors. It is statistically significant as p-value = 0.
- **Diagnostic Tests**
    - **Ljung-Box Test (Q):** 0.02 (Prob = 0.89)
        - This test will check for autocorrelation in the residuals. A high P-value suggests that there is no significant autocorrelation in the residuals, which is a good sign. It indicates that the model has captured most of the structure in the data.
    - **Jarque-Bera Test (JB)**: 82.50 (Prob = 0.00)
        - This test will check whether the residuals are normally distributed. The P-value of 0 indicates that the residuals are not normally distributed. This could be due to outliers or skewness in the data.
    - **Skewness:** -0.09
        - The residuals are slightly left-skewed, but it is closer to 0 indicating a near symmetric distribution.
    - **Kurtosis:** 4.44
        - The kurtosis is greater than 3, indicating that the residuals have heavier tails than normal distribution. This is consistent with the Jarque-Bera test result.
- **Conclusion**

- **Model Fit**: The model with its lack of autocorrelation in the residuals suggests that the model is a good fit for the data.
- **Residual Analysis:** Looking at the residuals, it shows no significant autocorrelation, but they are not normally distributed. This could affect the reliability of confidence intervals and predictions.
- **Potential Improvements:**
  - Consider removing the constant, as it is not statistically significant.
  - Explore other models, such as SARIMA or SARIMAX, if seasonality is present.
  - Investigate the non-seasonality of residuals and consider transformations (e.g., Log transformation) or robust modeling techniques.

4.6 Forecasting – Walk Forward Validation

After considering the results above, using an ARIMA model of (1, 2, 1) is best for our predictions, giving us confidence in the results produced from the model. Another factor to consider is how we will predict the time series data. Using Walk forward Validation we can improve the accuracy of our predictions.

By using Walking forward validation, we utilize a rolling window approach where the model is trained and tested on consecutive periods. This technique allows the model to be updated continuously, closely simulating real-world forecasting scenarios [15]

Some of the advantages using this technique are keeping things in order such a temporal order which is the following of one after another in time. It will help with simulating an environment where forecasting is necessary. The detection of concept drift which helps us keep the model up to date as concept drift itself isn't a bad indicator but rather a good one as it will tell us if the model isn't trained on a new undetected pattern which would lead to inaccurate predictions.

While using Walk-Forward validation is the limitation of data leakages. The main issue which can happen if a massive data leakage happens is it can cause the model to give the idea of accuracy until deployed into the real-world, leading to yield inaccurate results and leading to poor decision-making.

When implementing Walk Forward Validation, we track all observations in the history list which is seeded with the train data to which new observations are append each iteration.

Once this is complete, we can get a graph with the predicted values and the closing prices.

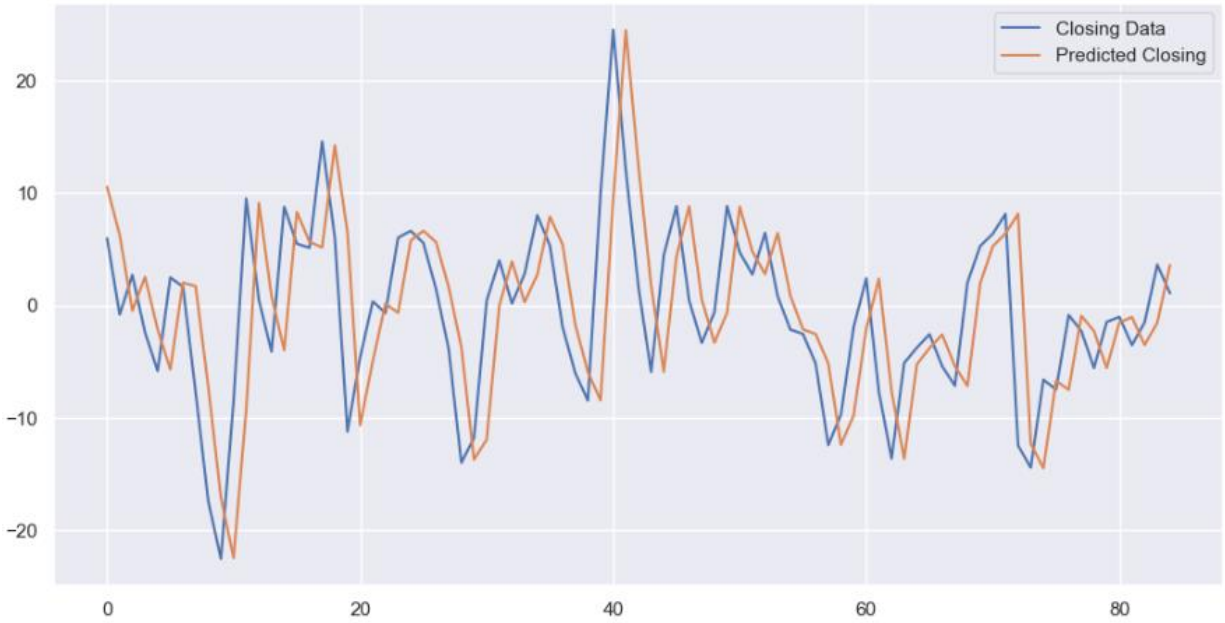


Figure 2.6: Closing Predictions

As we can see from the above the predictions are relatively accurate. This prediction scored a mean-squared error of 7.378 and the predictions follow the trend of our data. This means that our model is correctly predicting the stock price and has captured the underlying pattern within our dataset.

4.7 LSTM Model Setup (Long-Short Term Memory)

As we have shown a correlation between closing prices and public sentiment, we can then implement the LSTM. Before implementing the algorithm, we split our data into a training and testing set. This will allow us to train our algorithm on the dataset to find any patterns that may allow our model to use to increase its accuracy and understanding of the dataset – this same technique was used when setting up the ARIMA model.

We split our data by 80% with the other 20% being used for testing the model. This split will allow us to train the LSTM model with the majority of the data providing an accurate prediction.

Model: "sequential_16"

Layer (type)	Output Shape	Param #
lstm_32 (LSTM)	(None, 30, 30)	3,840
lstm_33 (LSTM)	(None, 30)	7,320
dense_32 (Dense)	(None, 25)	775
dense_33 (Dense)	(None, 1)	26

Total params: 11,961 (46.72 KB)
Trainable params: 11,961 (46.72 KB)
Non-trainable params: 0 (0.00 B)

Figure 2.7: LSTM Model Setup

The image above shows the architecture of the LSTM model we used to predict our stock prices. By using a Sequential model, it will pass the output of one layer directly to the next layer, using the architecture above we can describe each section:

1. **LSTM Layers**
    - Lstm_32: This layer has an output shape of (None, 30, 30) which will return sequences of 30-time steps with 30 features each.
    - Lstm_33: This layer has an output shape of (None, 30) meaning it returns the final output of the LSTM cell (30 features).
2. **Dense Layers**
    - Dense_32: This fully connected layer has an output shape of (None, 25).
    - Dense_33: This is the final output layer with an output shape of (None, 1) which will predict a single value.

Overall, the model is a sequential model where the LSTM layers are used to capture temporal dependencies in the stock price data, followed by dense layers to map the LSTM outputs to a final prediction. The model is lightweight reducing the chance of overfitting.

It provides efficient computational efficiency, resulting in faster training as the model has fewer parameters. This is particularly helpful when working with large datasets. We used a small dataset and having a more efficient and simpler model will allow for adaptability if a change in dataset is required by either reducing or increasing the size of the dataset.

Chapter 5: Results

Within this chapter it will cover the results of each algorithm and showcase what each graph means, analyse the direction and if there are any patterns or correlation between sentiment and Stock price. Making sure our models have captured the underlying structure of data and are accurate when being trained on our data.

5.1 Performance of ARIMA Model – Residuals

When it comes to the performance of the ARIMA model we can check to see if the residual pattern is randomly distributed without any real pattern.

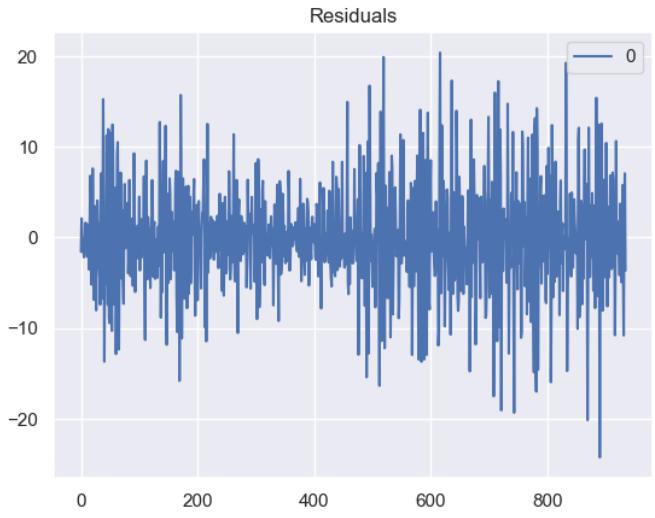


Figure 2.8: ARIMA Residuals

From the graph above we can see that the residuals are plotted randomly and often fluctuate around zero, which is a good sign. This tells us that the model has captured the underlying structure of our data. Often residuals tend to range between 20 to –20, the magnitude of the residuals provide an indication of the model's accuracy. Smaller residual values often tend to correspond to a better model fit.

The residuals also don't exhibit any discernible pattern or trends. The presence of such patterns would suggest that the model has failed to capture certain components of the data, indicating potential errors in its assumptions which lead to inaccurate or unreliable results.

5.2 Performance of ARIMA Model – MSE

By analysing each graph of predictions between stock price and public sentiment, we can prove there is some correlation between public sentiment influencing stock prices. Microsoft, Apple and Amazon are the 3 main stocks which have been used to produce each graph, first looking at Microsoft.

5.2.1 Microsoft

The result produced a mean-squared-error of 2.5

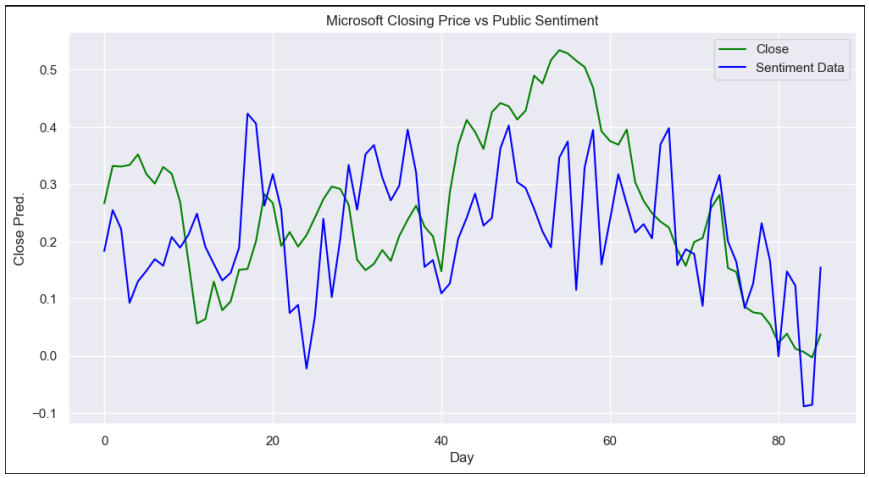


Figure 2.9: MSFT Stock & Sentiment Predictions

From the above graph we have now predicted our sentiment for Microsoft alongside its closing price.

The x-axis contains the number of days predicted and we can analyse 30, 40, 50 days up to 80 days in advance. We can clearly see that there are patterns to when sentiment rises, stock also rises at the same time. Now this shows that there is a correlation between sentiment influencing stock prices, but this isn't always 100% accurate as we can see around the 35-day mark, that the sentiment stays high where the stock price drops slightly to its 2nd lowest low.

5.2.2 Apple

After conducting analysis of the Apple stock combined with its sentiment, it produced a total mean squared error of 2.25. This indicates that the model is off around 2.25 units when predicting both stock and sentiment data.

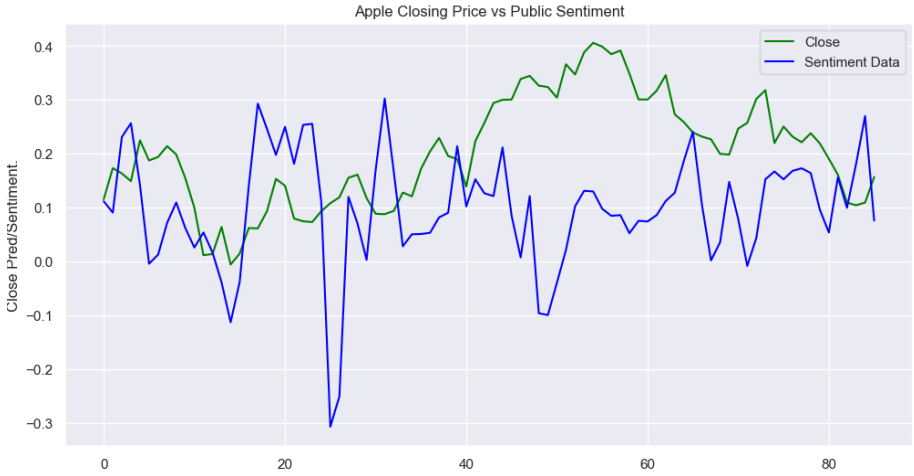


Figure 3: AAPL Stock & Sentiment Predictions

When analysing the graph above we can see that the sentiment follows the general trend of the stock price. There are some outliers as the model progresses further into the future with a slight discrepancy happening around day 25 with the sentiment dropping and rising sharply.

What can cause these sudden changes in sentiment could have several reasons:

- **Outliers & Anomalies**: Sudden changes within our data could cause these fluctuations. These outliers might be due to data entry errors, rare events or other irregularities.
- **External Factors**: Unaccounted external factors that can influence the target variable causing these sudden changes in predictions.

5.2.3 Amazon

After conducting analysis of Amazon stock closing price combined with its public sentiment, it produced a mean-squared error of 2.25 units. This indicates that our model is often off around 2.25 units when predicting price and sentiment.

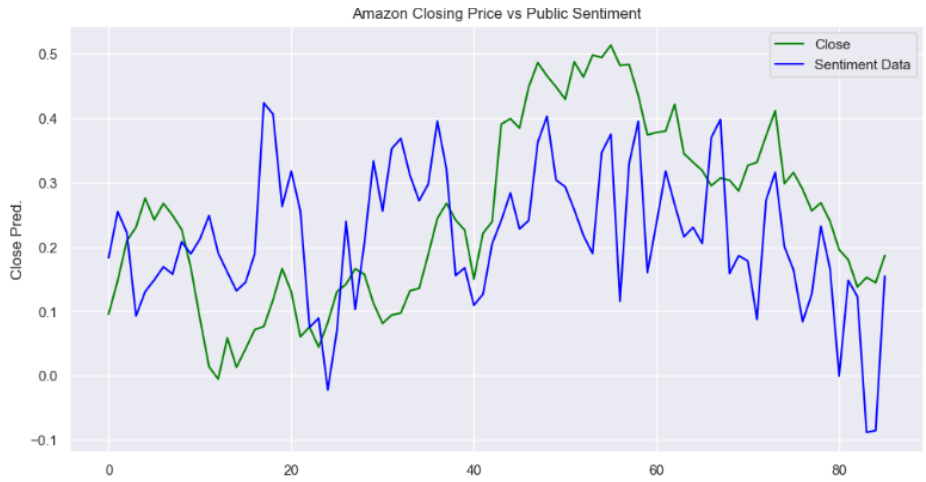


Figure 3.1: AMZN Stock & Sentiment Predictions

When analysing the graph produced above, we can see that again, public sentiment follows the trend of the stock price. We can see there are sharp increases and decreases in sentiment scores. Again, this can be due to other external factors that haven't been considered or any outliers and anomalies that have occurred before predictions such as data processing or rare events.

5.2.4 Future Improvements to the ARIMA Model

Given the results from the ARIMA model, the model works as intended and provides useful insights into the process of predicting but the model can be improved upon such as the efficiency of predicting:

- **Evolve ARIMA Model**: This model implements the X variable known as exogenous variables. This takes in external information that could influence the time series, leading to more accurate forecasts. This could be worth looking into as the ARIMA model itself only takes in a single time series value whereas ARIMAX takes in multiple variables to include external values.

- **ARIMA (p, d, q)**: When testing the ARIMA model we used p, d, q values of (1, 2,1), by adjusting the values to find the best values that will fit the best with our data. By using the ADF and PACF plots we can use different p, d, q values to test different fits for our data.

- **Apply Further Differencing**: When we difference our data, we only apply First Order Differencing. We can apply different levels of differencing to achieve proper stationarity ensuring our data is stationary but also being aware that we avoid over differencing. This could lead to a loss of information and potentially create a worse model.

5.3 Performance of LSTM Model – MSE

After conducting testing of the LSTM model, we produced 3 graphs that predict 50 days ahead. Each graph will show the predicted values of Apple, Microsoft & Amazon's closing prices alongside the predicted sentiment score of each day.

5.3.1 Apple

The first test was conducted on the Apple Stock closing price alongside its sentiment score. The model scored a mean-squared-error score of 8.1 units. This indicates that it is off on average 8.1 units when predicting. When comparing this score to the ARIMA model which scored a high of 2.5 units and a low of 2.2 units. This tells us that the ARIMA model performs slightly better compared to the LSTM model given that it has a lower MSE score.

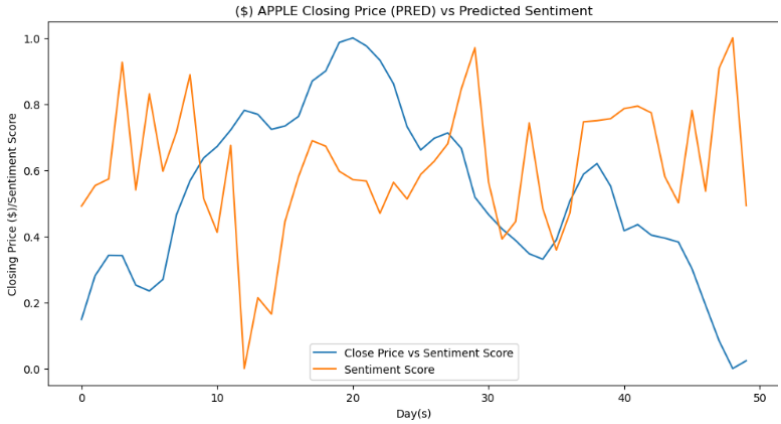


Figure 3.2: AAPL Stock & Sentiment Predictions

5.3.2 Microsoft

The second test resulted in a total mean squared error score of 13.5 units.

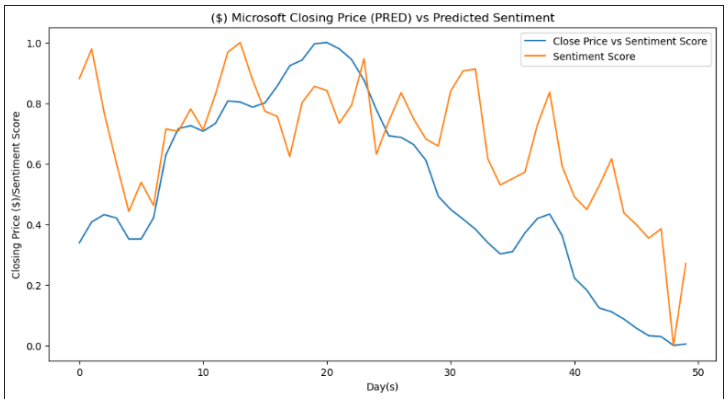


Figure 3.3: MSFT Stock & Sentiment Predictions

5.3.3 Amazon

The third test resulted in a total mean squared error score of 6.7 units.

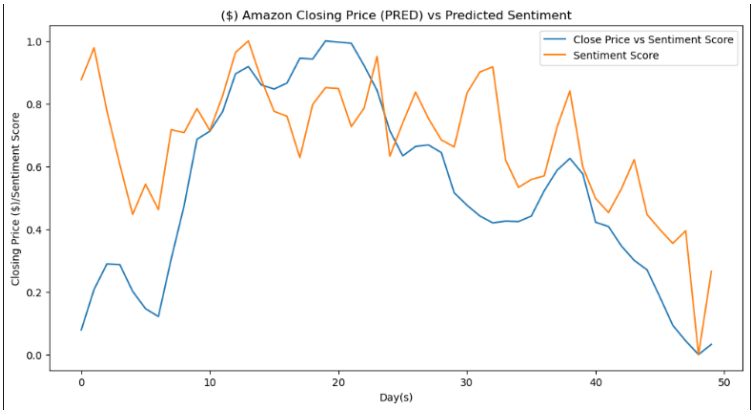


Figure 3.4: AMZN Stock & Sentiment Predictions

To get this total mean squared value we can use this formula: total_mse = ((n1 * mse1) + (n2 * mse2)) / (n1 + n2). Where n1 and n2 are the number of values within the dataset. This will give us the true mean squared value.

5.3.4 Future Improvements to the LSTM Model

Given the results above show that the ARIMA model outperforms the LSTM model with an average of 2.3 units and the LSTM's average was 9.4 units. Now, this doesn't mean that the LSTM model is inherently worst but could require more tweaking to the model itself given that we used a simplistic model.

Some improvements that can be applied to the model:

- **Early Stopping** [16]**:** This can be used to help improve results while also giving you better computing performance as it saves computing resources. Early Stopping can be used to monitor specific metrics such as the loss metric, which will help tune the algorithm, improving accuracy.
- **Utilizing Different Optimizers:** The current LSTM model uses an optimizer called 'adam' which is powerful enough to be used but there is an extension of this optimizer called 'NADAM' [17]. It uses Nesterov Momentum, which looks ahead by computing the gradient at an approximated future position based on current momentum. This can help the optimizer make more informed updates leading to faster and more stable convergence.
- **Regularizers:** This technique helps with reducing overfitting within our model [18]. This adds a penalty to the loss function within the model which stops the model from learning overly complex patterns which in turns stops overfitting.

5.4 LSTM 10, 20, 30 Days Predictions

5.4.1 Amazon 10 Days

Figure 4.0: AMZN 10 Days Prediction

5.2.2 Amazon 20 Days

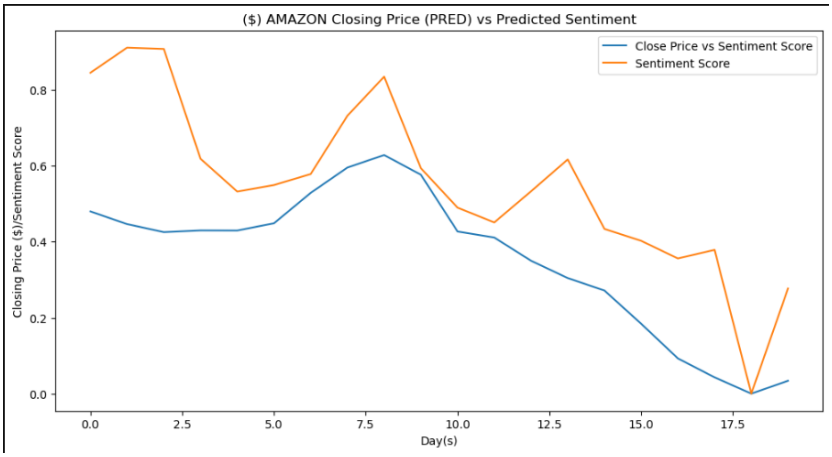


Figure 4.1: AMZN 20 Days Prediction

5.2.3 Amazon 30 Days

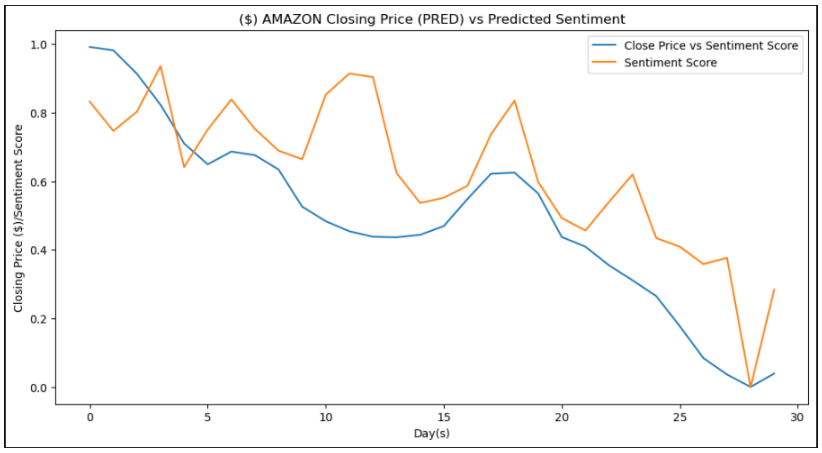


Figure 4.2: AMZN 30 Days Prediction

5.2.4 Microsoft 10 Days

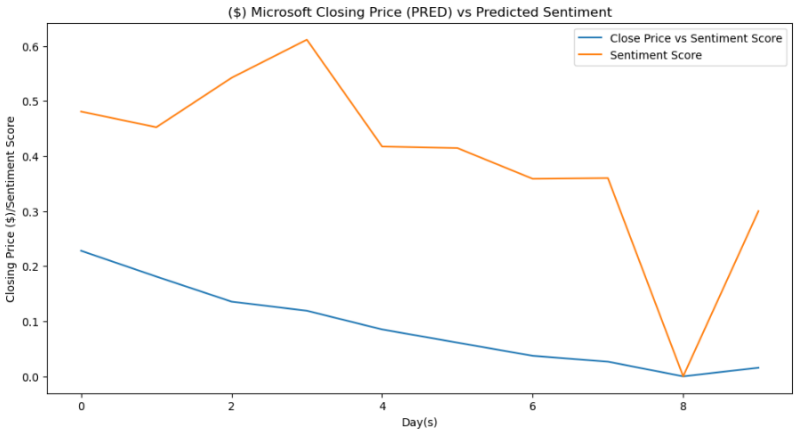


Figure 4.3: MSFT 10 Days Prediction

5.2.5 Microsoft 20 Days

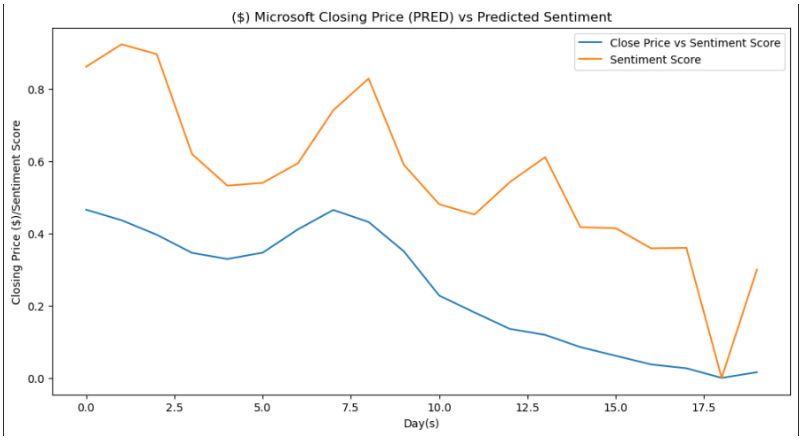


Figure 4.4: MSFT 20 Days Prediction

5.2.6 Microsoft 30 Days

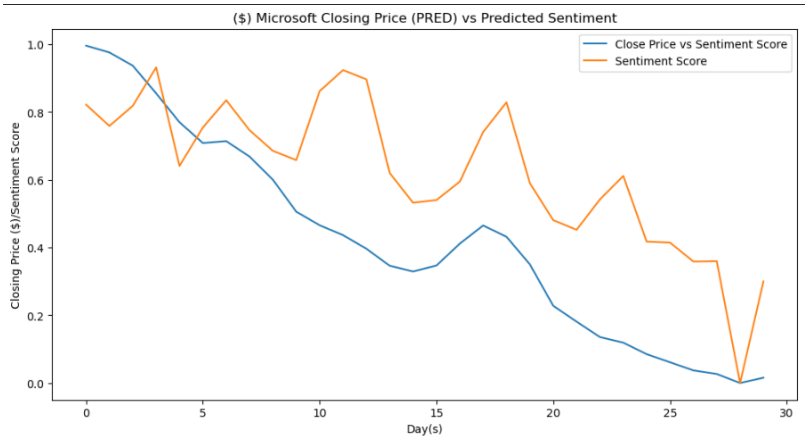


Figure 4.5: MSFT 30 Days Prediction

5.2.7 Apple 10 Days

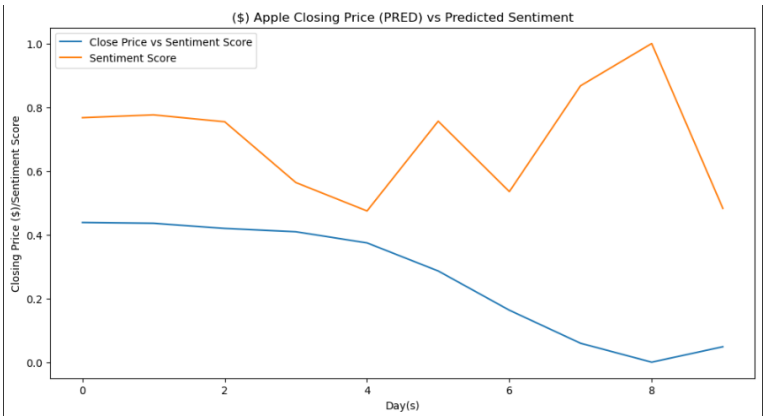


Figure 4.6: AAPL 10 Days Prediction

5.2.8 Apple 20 Days

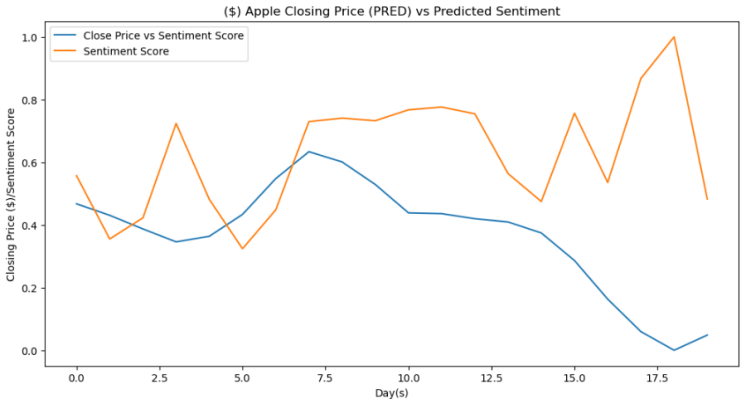


Figure 4.7: AAPL 20 Days Prediction

5.2.9 Apple 30 Days

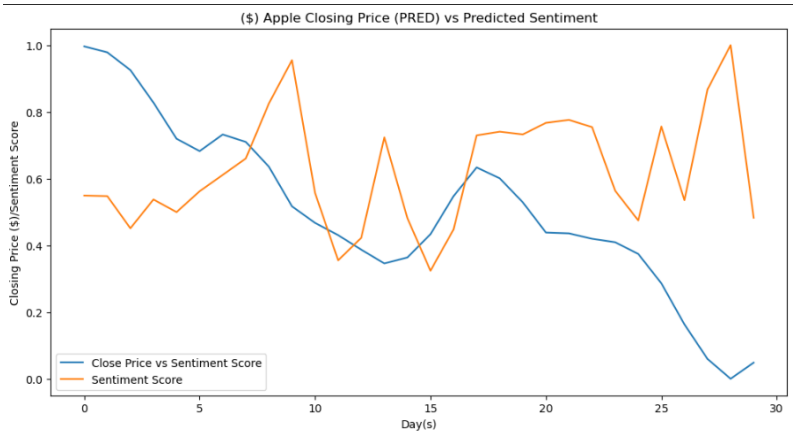


Figure 4.8: AAPL 30 Days Prediction

Analysing the above graphs, we can see that each stock and sentiment prediction follow each other, confirming the correlation between them. The only few outliers that are noticeable are Apple's 20- and 30-day predictions. Again, this can be caused by unforeseen events that can cause sentiment to drop drastically.

5.3 ARIMA 10, 20, 30 Days Predictions

5.3.1 Amazon 10 Days

Figure 4.9: AMZN 10 Days Prediction

5.3.2 Amazon 20 Days

Figure 5.0: AMZN 20 Days Prediction

5.3.3 Amazon 30 Days

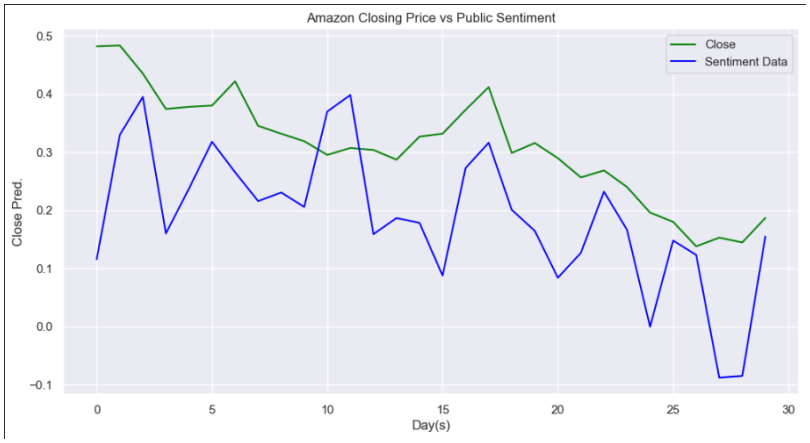


Figure 5.1: AMZN 30 Days Prediction

5.3.4 Microsoft 10 Days

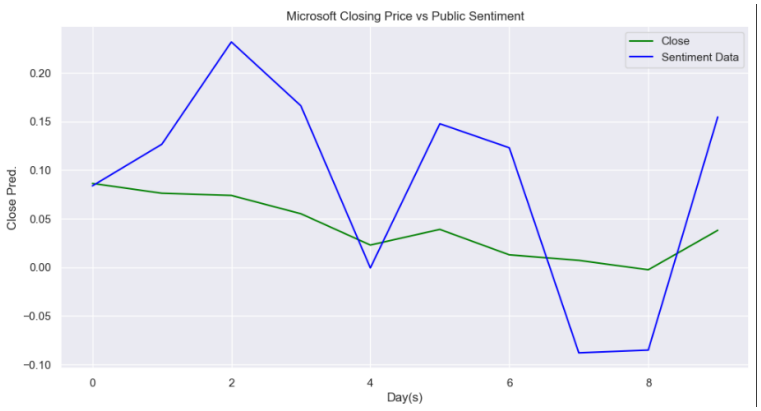


Figure 5.2: MSFT 10 Days Prediction

5.3.5 Microsoft 20 Days

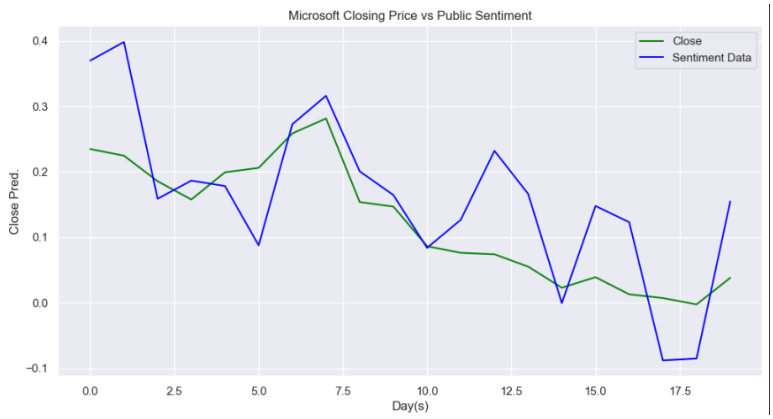


Figure 5.3: MSFT 20 Days Prediction

5.3.6 Microsoft 30 Days

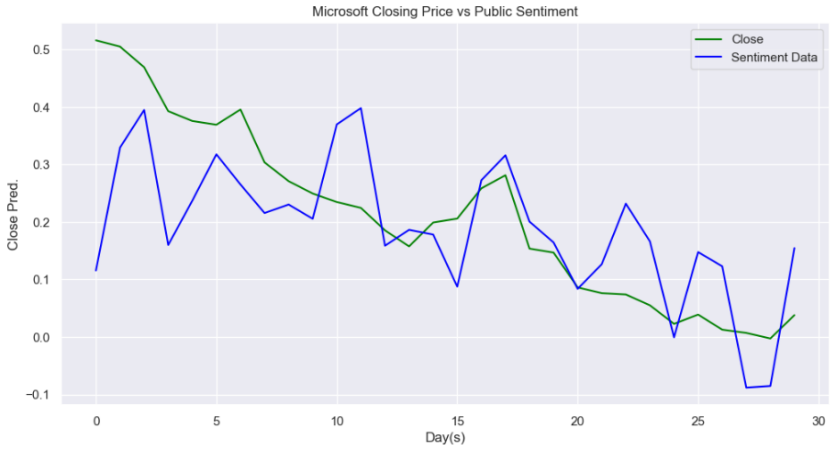


Figure 5.4: MSFT 30 Days Prediction

5.3.7 Apple 10 Days

Figure 5.5: AAPL 10 Days Prediction

5.3.8 Apple 20 Days

Figure 5.6: AAPL 20 Days Prediction

5.3.9 Apple 30 Days

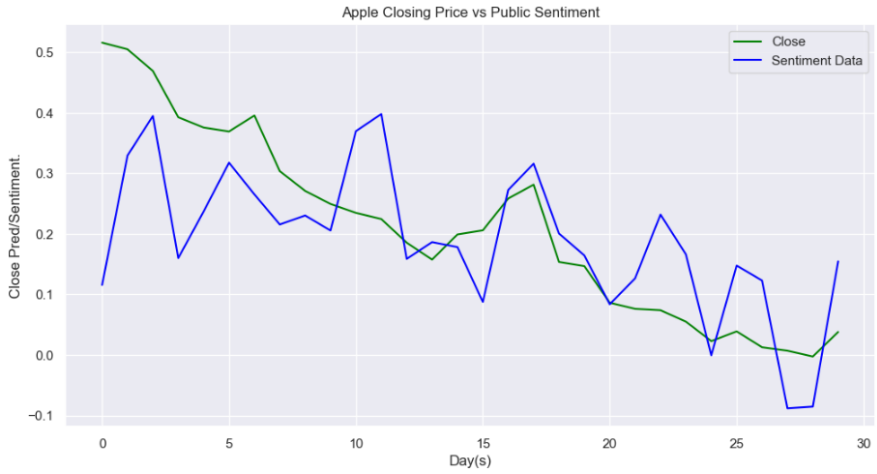


Figure 5.7: MSFT 30 Days Prediction

As we can see from the above graphs from the 10-, 20- and 30-day predictions, we can see a general trend from each graph. Public sentiment and stock price follow each other which further indicates the correlation between the effects of public sentiment on the stock price. We can also see the similarities between each stock's 10-day predictions. This could be caused by the similarities in their stock prices.

Chapter 6: Conclusion

6.1 Achievements

6.1.1 Successful Implementation of LSTM & ARIMA Models

In Chapter 4 we successfully implemented an ARIMA model which performed well given the result of its mean squared error score. When predicting the stock price alongside the sentiment score it was off on average 2.3 units when predicting stock prices of Microsoft, Amazon and Apple.

6.1.2 Comparison of MSE Scores

Also in Chapter 4, we implemented an LSTM model which performed well but underperformed compared to the ARIMA model with it being off on average by 9.4 units. It was able to predict the stock price of Microsoft, Amazon and Apple well being able to show the correlation between stock price and public sentiment by analysing the graphs produced.

6.1.3 Proof of Correlation

We were also able to show the correlation between public sentiment influencing the stock price by analysing the graphs produced by both models in Chapter 5, specifically Figures 2.9, 3.1, 3.2, 3.3 & 3.4. Before building each model, we successfully proved that the dataset we are using is stationary and the p-value is less than 0.05.

6.2 Limitations

After conducting testing of each algorithm, further testing and complexity can be added to each model. Even though the LSTM model did perform slightly worse than the ARIMA model, the addition of a larger dataset could help the models produce more accurate results. The use of a larger dataset will also help with reducing the chance of overfitting our model and improve our model to perform better on unseen data rather than just memorizing.

Another addition that can help with further research into each model is to allow for real-time sentiment from previous years. During testing we used sentiment data from 2021 up until late 2022. This can be improved by gathering data from 2020 and earlier, 2020 was also the year the COVID-19 pandemic began, and we can see from analysing the stock, its lowest price is during that time. This sentiment data was gathered from Kaggle which contained tweets about different stocks during that time, while this is great for testing it would be useful to gather sentiment data in real time from live feeds from social media.

With the difficulty of choosing the correct model, by analysing only 2 models it limits the comparison of other models that have a proven track record such as ARIMAX, SARIMA, SVG, Exponential smoothing and Neural Networks which would allow for a deeper analysis of how these models work.

6.3 Volatility of Stock Markets

Throughout this dissertation, there have been multiple case studies which have shown the effects of external factors such as political events, company decisions and public sentiment affecting the stock prices of certain shares. This shows the volatility of the stock market itself; no stock is a sure bet and those that have bet on markets often lose and some lose big. Artificial intelligence and Machine learning have been deployed into the stock market with some showing success but often struggle to predict the future as unforeseen events can affect these models greatly with these models overcompensating these events. We can see this in the graphs produced by the LSTM and ARIMA model with figure 3.2 showing a massive drop in sentiment which could be uncertainty within the model itself.

6.4 Future Work

6.4.1 Research Machine Learning Models

As mentioned in the limitations section, further research should involve investigating other models. Implementing these models should be fed through the ADF (Augmented Dickey-Fuller Test) test and the KPSS (Kwaitkowski-Phillip-Schmidt-Shin) test to ensure that any data being fed through these models and any results produced are significant.

6.4.2 Research Sentiment Analysis Techniques

It is important that when conducting sentiment analysis that the correct techniques are used to gather the most accurate scores for each type of sentiment. The use of VADER sentiment analysis in this dissertation provided accurate sentiment scores but it is still worth while researching other techniques such as training algorithms such as Naive Bayes, Support Vector Machines or Aspect Based Sentiment Analysis.

6.4.3 Refinements to LSTM & ARIMA Models

ARIMA model implementation used p, d, q values 1, 2, 1. Further testing should include different model variations to compare and find if there are other models that provide a better fit for our data or if they provide better results.

When it comes to the LSTM model, we can setup our model using the techniques mentioned in the section 5.3.4 such as Early stopping, regularizers and utilizing different optimizers that could help achieve a lower mean squared error score and even improve the computational efficiency of our model.

7. APPENDIX

7.1 References

[1] Mukherjee S., Sadhukhan B., Sarkar N., Roy D. and De S. (2021). Stock Market prediction using deep learning algorithms. *CAAI Transactions on Intelligence Technology*. Stock market prediction using deep learning algorithms - Mukherjee - 2023 - CAAI Transactions on Intelligence Technology - Wiley Online Library

[2] S. Atsalakis, G. and P. Valavanis, K. (2013). Surveying Stock Market Forecasting Techniques – Part 1: Conventional Methods. Pp49-104. Available at: https://www.researchgate.net/profile/George-Atsalakis/publication/236620807_Surveying_stock_market_forecasting_techniques_-_Part_I_Conventional_methods/links/540e0dce0cf2df04e756c884/Surveying-stock-market-forecasting-techniques-Part-I-Conventional-methods.pdf

[Accessed 2nd Nov. 2024]

[3] Roman, J. (n.d.). Autoregressive Models Overview. Published by: Penn State University [Online] Available at: https://www.e-education.psu.edu/meteo820/node/8#:~:text=An%20AR%20model%20is%20also,a%20difficult%20requirement%20to%20meet.

[Accessed 3rd Nov 2024]

[4] Mehandzhiyski, V. (2023). What is an ARMA Model? [Online] Available at: https://365datascience.com/tutorials/time-series-analysis-tutorials/arma-model/

[Accessed 3rd Nov 2024]

[5] Zahn, M. (2024). Stock Market surges on Election Day [Online] Available at: https://abcnews.go.com/Business/stock-market-surges-election-day/story?id=115512397

[Accessed 5th Nov. 2024]

[6] Nau, Robert (N/A). Statistical forecasting: Notes on regression and time series analysis [Online] Available at: https://people.duke.edu/~rnau/411arim.htm

[Accessed 5th Nov. 2024]

[7] Mohammed N. Nounou, Bhavik R.Bakshi (2000). Data Handling in Science and Technology [Online] Available at: https://www.sciencedirect.com/topics/chemistry/autocorrelation-function

[Accessed 5th Nov 2024]

[8] Penn State University (N/A). Partial Autocorrelation Function (PACF) [Online] Available at: https://online.stat.psu.edu/stat510/lesson/2/2.2

[Accessed 5th Nov 2024]

[9] Jonny Brooks-Bartlett (January 3rd, 2018) Probability concepts explained: Maximum likelihood estimation [Online] Available at: https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

[Accessed 5th Nov 2024]

[10] Rob J Hyndman and George Athanasopoulos (April 2018). Forecasting: Principles and Practice (2nd ed)

Available at: https://otexts.com/fpp2/least-squares.html

[Accessed 5th Nov 2024]

[11] Cameron Davidson-Pilon (6th June 2014) Feature Space in Machine Learning. Available at: https://dataorigami.net/2014/06/06/Feature-Space-in-Machine-Learning.html

[Accessed 18th Nov 2024]

[12] Kaggle: https://www.kaggle.com/

[13] Kaggle Dataset Available at: https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction

[14] Viren Rehal (5th August 2024) ADF Test: Augmented Dickey Fuller Equation. Available at: https://spureconomics.com/adf-test-augmented-dickey-fuller-equation/

(Accessed: 04/03/2025)

[15] Katie Kolchin, CFA, Director of Research SIFMA Insights (23rd April 2020) The VIX's Wild Ride. Available at: https://www.sifma.org/resources/research/insights/the-vixs-wild-ride/ (Accessed: 05/03/2025)

[16] Istiaq Ahmed Fahad, Understanding (28th October, 2024) Walk Forward Validation in Time Series Analysis. Available at: https://medium.com/@ahmedfahad04/understanding-walk-forward-validation-in-time-series-analysis-a-practical-guide-ea3814015abf

[17] Keras Documentation. Nadam. Available at: https://keras.io/api/optimizers/Nadam/

[Accessed 21st March 2025]

[18] Dhruv Matani (Aug 23rd, 2024) Interpreting Weight Regularization in Machine Learning. Available at: https://towardsdatascience.com/interpreting-weight-regularization-in-machine-learning-99f2677f7ef5/#:~:text=Weight%20regularization%20means%20applying%20some,to%20generalize%20to%20unseen%20inputs.

[Accessed 21st March 2025]

[19] Keras Documentation. Early Stopping. Available At: https://keras.io/api/callbacks/early_stopping/

[Accessed 21st March 2025]

[20] Annabel Smith (13th April 2021) The Reddit Revolt: GameStop and the impact of social media on institutional investors. Available at: https://www.thetradenews.com/the-reddit-revolt-gamestop-and-the-impact-of-social-media-on-institutional-investors/

[Accessed 8th Nov 2024]

[21] Hyndman, R.J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27(1), 1-22.* Available at: https://otexts.com/fpp2/arima-r.html

[Accessed 22nd March 2025]

[22] Xuan Ji. (9th March 2025). LSTM (Long Short-Term Memory) case studies: revolutionizing machine learning across industries. Available at: https://www.byteplus.com/en/topic/399645?title=lstm-long-short-term-memory-case-studies-revolutionizing-machine-learning-across-industries