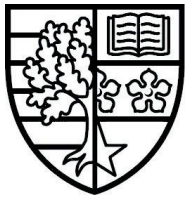
Brodie Harkins

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

This study conducts public sentiment analysis using modern techniques such as Logistic regression, Tokenization, Lemmatization and noise reduction to clean and pre-process sentiment data which will be fed through 3 different Machine Learning models that are used to forecast future stock prices. The three models that will be used are Growing Neural Gas Networks, Support Vector Machines and Long-short Term Memory models. All have their own benefits and drawbacks.

Incorporating previous studies and conducting quantitative and qualitative research to gain correct and up to date information on financial stock markets and public sentiment data. When data has been preprocessed, cleaned and prepared for input, each algorithm is fed one data set of financial information only and a data set of public sentiment and financial data. Through this, comparative results are produced showing prediction accuracy, performance and efficiency.

Throughout this paper, information regarding the computation of the predictive models has been expanded upon and explained.

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 3 different models and inputting the same data set of public sentiment and financial data into each of the 3 models. Results will contain, F1 Score (Measures how many errors a model made) and Area Under the Precision-Recall Curves (AUROC).

Graphs will be displayed showing the accuracy and performance of each model and how it performed. A comparative analysis will be conducted to show these findings and ultimately a conclusion on which model is best fitted and the effects of social media sentiment on predictive frameworks.

ACKNOWLEDGEMENTS

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

Table of Contents

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 Growing Neural Gas Network, Support Vector Machine 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 dive deep into 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, analyzing 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 Growing Neural Gas algorithms. We will compare how each algorithm works, and which is the most accurate using current day data.

When it comes to completing the task of testing and evaluating each algorithm, we will be using real life data that is publicly available to all to access. Websites like Finanzon [34], MarketStack[35] & Keggle[33] contain relevant stock market data which will produce proper and consistent results based on the algorithms we are using. Finazon & marketstack is an API which provides real-time, historical data and various other features that are beneficial for financial analysis and decision-making. Overall, a reliable choice for what we need to succeed in gathering accurate data.

1.1. **Motivation**

The motivation behind this research is to better understand and improve the accuracy of predicting stock market behavior. With its constant movements and unpredictability presents a difficult challenge to those who are new or experienced in the financial industry. 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.

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, Twitter, Instagram and Facebook, along with the availability of real-time financial data from sources such as Finazon and MarketStack, additional layers of complexity have been added to predicting stock market trends. Market sentiment, corporate performance and global events are now analyzed in conjunction with historical market data, making prediction a complex challenge. The goal of this research is to investigate how modern machine learning algorithms, such as Long-Short Term Memory (LSTM) networks, Growing Neural Gas (GNG) 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:

1. ***To Investigate and compare different machine learning algorithms****:* 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 Growing Neural Gas Network (GNG) and other machine learning techniques in predicting stock market trends. By identifying the strengths and weaknesses of each approach, the research will determine which models are most suitable for financial forecasting.
2. ***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
3. ***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.
4. ***To evaluate the predictive framework using real-world, publicly available data:*** The study will utilize real-world data from sources such as Finazon, MarketStack & 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), F1 Score (Measures how many errors a model made) and Area Under the Precision-Recall Curves (AUROC)

1.3. **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:

1. ***Comparative Analysis of Machine Learning Algorithms*:** This research conducts an analysis of different machine learning algorithms, including Long Short-Term Memory (LSTM) networks and GNG (Growing Neural Gas) algorithms. 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.
2. ***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.
3. ***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.
4. ***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.
5. ***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.

1.4. **Organization**

This thesis 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, the evolution of machine learning techniques and the influence of social media sentiment on financial markets. It will also cover the techniques used to analyze and clean social media sentiment data. This chapter will also aim to identify the gaps in existing research.

- **Chapter 3**: In this chapter, it will provide the details of the methods used in this study, including explaining the algorithms and models applied to analyze data and predict stock movements. It will also cover Methodology, Project management & Risk Analysis.

- **Chapter 4:** This chapter will look into Data cleaning and preprocessing, covering the techniques that are used when obtaining and processing data to be fed into predictive models.

- **Chapter 5:** In this chapter, it will cover the requirements needed to replicate and conduct this experiment.

2. Literature Review

In this literature review we will aim to give an overview of existing research and identify any gaps in previous research that this study will attempt to fill in. 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 [3] 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).

2.1.1. *Auto Regressive Model (AR)*

Standard AR models tend to use polynomial time. The way polynomial time algorithms work is, - for readers convivence this will be simplified to help understand the algorithm – 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. [4]

2.1.2. *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 [6].
The ARMA model uses the equation below to achieve predictions:

$$y_t = c + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

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.
- $\phi_1 y_{t-1}$: 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

1. A constant Baseline (c)
2. A weighted influence of the past value (y_t).
3. The impact of past random shocks or errors ($\phi_1 y_{t-1}$).
4. 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.

2.1.3. ***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)

2.1.4 ***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]

2.1.5 ***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. As mentioned earlier

2.1.6 ***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 modeling 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]

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

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

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

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)

2.1.7 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:

$$x_t = \mu + \omega_t + \theta_1\omega_{t-1} + \theta_2\omega_{t-2} + \cdots + \theta_q\omega_{t-q}$$

Where θ_1, , θ_2 , ..., θ_q are moving average coefficients, and ω_t 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.

2.1.8 ***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) [13] and the Partial Autocorrelation function (PACF) [14]. 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 [15] or Least Squares approach [16].

2.2. ***Machine Learning Approaches in Stock Market***

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

2.2.1. ***Growing Neural Gas (GNG)***

Growing Neural Gas networks are a type of unsupervised learning algorithm used to model and analyze complex data structures [18]. What makes a GNG network effective is where data is unknown or constantly evolving. An overview of the Growing Neural Gas network and how it is applied in stock market predictions.

Key Characteristics of GNG Networks [20]:

1. *Dynamic Structure*: Each GNG network will start with the minimum number of nodes and incrementally add new nodes based on the inputs data's complexity. This allows adaptability throughout the network to represent intricate data effectively.
2. *Topology Preservation:* By establishing connection between nodes, it helps the network hold onto the natural patterns or shapes in the data, making it easier to model and help understand complex relationships.
3. *Competitive Nodes:* Nodes will compete amongst each other to be the first new node. If won, the node is placed in with its neighbor's data also being adjusted. This process ensures that the network correctly captures the data distribution.

2.2.2. ***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.

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

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
- [W_i , b_i] - weight & biases associated with the input gate
- [h_{t-1}] - hidden state from previous step/result
- [x_t] - 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.

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 h_t 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:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Where:

- W_o ,b_o - 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.

2.2.3. *Support Vector Machine (SVM)*

Support Vector Machines operate in a way that separates data by using a line of best fit which for this model is called the Hyperplane. The Hyperplane is then used to separate data of different classes in a feature space. A feature space is where your variables will be housed and not the target variable even if this is present. An example which is inspired from *Data Origami by Cameron Davidson-Pilon [31]* shows how this feature space work. What makes SVM's great to use is their ability to recognize patterns and separate data even if it is working with small data sets. While SVM's could be used for regression tasks – this could be working to predict housing

prices to stock market predictions - it is mainly used for classification tasks - classification task could be working to sort out a group of images into their respected variable type i.e. a picture of a dog would be placed into a dog class etc. The Support Vector Machine model will be tested using datasets from Finazon and Kaggle, testing performance and accuracy.

2.3. **Influence of Social Media Sentiment on Stock Market Prediction**

Social media sentiment can heavily influence the stock market, particularly if the sentiment is negative [21]. Through networks like Reddit threads "r/WallStreetBets" have shown to have this negative effect on the stock market itself. With recent developments in Artificial Intelligence and Machine Learning algorithms, markets can calculate or at least attempt to predict market sentiment. Social Media platforms such as Instagram, Twitter (X) and Facebook can be swept in order gather information that can be analyzed at real-time to get an idea of public sentiment.

2.3.1. ***Mechanisms of Social Media Sentiment analysis***

To conduct social media sentiment analysis, tools must be used to achieve this. These tools are:

Lexicon Based Approach, Machine Learning

Through these tools, we can establish a clear public sentiment. [22] Up to 70% of decisions are made emotionally compared to 30% which are rational choices. Being able to analyze likes, comments, shares and mentions, brands can gain valuable insights into public sentiment and their emotional drivers that influence their purchase decisions. Through this, brands can improve their marketing strategies, improve customer service and make better decisions for their business.

Lexicon Based Approach

The *lexicon Based Approach* works based on using a collection or dictionary of predefined tokens/words assigned with a predefined score of +1, 0 or –1. +1 - Positive, 0 - Neutral and (-1) - Negative. During the reviewing stage of text, the text is analyzed and scores for each relevant aspect are summed separately giving the overall sentiment score. During the final stage of this approach, the polarity assignment is completed. This is where overall polarity is assigned to text based on the highest value of individual scores and the text is given an overall score, positive, neutral and negative.

Machine Learning Approach

The Machine Learning approach involves training models on labeled datasets to automatically recognize patterns, categorized text or predict outcomes. Unlike its predecessor Lexicon Based Approach, which relies heavily on predefined dictionaries, machine learning approach learns patterns from data, allowing it to generalize to new text effectively [23]. Within the Machine Learning Approach, there are two ways to go about analyzing social sentiment.

- Supervised machine learning (Relies on Labeled data)
- Lexicon-based unsupervised learning (Unlabeled Data, uses knowledge bases, lexicons etc.)

Both methods must be trained on a training set before handling actual data. When training each method, each text instance is marked with the preferred output, such as; Sentiment label, Topic category or intent label. Text samples are also logged as positive, negative or neutral. During training *Hard or Soft Classification* is used to assign labels to each piece of text such as "positive" or "negative". This would be known as *hard classification* it provides a clear and definitive sentiment prediction. *Soft Classification* provides the statistics behind each text (80% positive to 20% negative). Although this approach doesn't give you a clear-cut sentiment analysis it provides more nuanced predictions, displaying uncertainty within the model.

Sentiment analysis benefits from syntactic and Linguistic factors within the model to help with these classification issues. Syntactic factors investigate word order, part-of-speech tags and dependency relationships help the model understand how words are used in context. Linguistic factors help the model dive deeper into the context behind words and even semantic cues. Machine Learning models can recognize complex expressions, idioms and nuances such as sarcasm or irony.

An example of a Machine Learning approach is Logistic Regression (LR). Logistic regression works by multiplying an input value by a weight value which indicates the importance or influence of that feature on the prediction. This process results in a weighted sum of the input values, which is then passed through a logistic function or otherwise known as sigmoid function [24]. The Logistic function will map a weighted sum to a probability between 0 and 1. If the probability reaches past a threshold – typically around 0.5 - the instance is weighted into one class or the other. The mathematical foundation of Logistic regression uses maximum likelihood Estimation (MLE) to optimize the model's parameters [15]. MLE finds the set of parameters that maximizes the probability of the given data.

2.3.2. ***Correlation Between Social Media Sentiment and Stock Market Movements***

The influence of social media on stock market movements, websites like; Facebook, Twitter (X) and Instagram have all gained substantial attention in recent years. These social media sites all provide high amounts of real-time data on public opinion.

Social media sentiment refers to the collective attitude's, emotions and opinions expressed by users on these social media apps about various topics, including companies, economic events and even the stock market. This sentiment can provide key insights into the mood of investors and the broader public, which in turn influences the stock market prices.

What makes social media sentiment particularly valuable in financial analysis is due to how readily available it is, capturing real-time reactions to news, corporate announcements and global events. [26] This recent study showed that after using sample data from the COVID-19 pandemic, that market volatility after the pandemic is more susceptible to investor sentiments and internet sentiments have a larger impact on stock markets.

2.3.3. *Case studies and Evidence of Correlation*

Numerous studies have been released that show the effects of sentimental data from websites like Facebook and mainly Twitter. Websites like YouTube have grown from showing silly videos to becoming a hotbed for market trends and even educating online investors on improving their skills in buying and selling shares. Channels like Bullish Investor and New money, which target beginners who are looking to get into the stock market.

One tweet from Elon back in 2018 even influenced the Tesla stock:

> "Am considering taking Tesla Private at $420. Funding secured" - @elonmusk 2018 (X.com)

This tweet increased Tesla's stock by 7%, which in turn caused losses to those who shorted the stock – Shorting is the predicting of the stock price to trend downwards. Even in current times, after the 2024 Election, which saw President Trump return to the Whitehouse, Tesla's stock increased by %15 in a single day of trading [26].

As much as Facebook, Instagram and Twitter all get the spotlight for influencing the stock market, Reddit/Redditors have all shown how powerful their influence on the market can be. Back in 2021 during the time of COVD-19, GameStop seen a massive increase in their share price. Trading at $5 a share in 2020, most investors consider the stock to be dead until 2021 after new board members promised to change things going forward and increased the share price to $18 a share. Investors, hopeful that the stock price could rise, and GameStop could turn it around began investing. Within days the price rocketed to more than 2,300% higher to around $483 a share. [27]

An important case study that was conducted by Johan Bollen, Huina Mao and Xiaojun Zeng [2] revealed a significant correlation between twitter mood and movements in the Dow Jones Industrial Average. By analyzing the volume and polarity of tweets – Number of positive words and the number of negative words is counted -, the study managed to prove the shifts in public mood could anticipate stock market fluctuations with a certain degree of accuracy.

By watching these stock market prices, often a positive earnings report can correlate to an

increase in market price following the rise of positive sentiment on Twitter as the market reacts to the good news. In the past negative events can have the opposite effect following a scandal or unfavorable policy with the public can lead to a decline in share price when investors learn of the news and listen to the public response.

2.3.4. *Limitations and Challenges in Correlating Sentiment with Market Trends*

Some of the challenges when it comes to collecting public sentiment from websites like Twitter is there is a lot of noise. Noise being pointless conversations that contain various personal opinions, marketing posts and automated account (Bots) adding to the stream of information. Noise can affect the overall prediction, reducing the overall accuracy of the predictive model.

Another challenge that we must consider is how dynamic and fast these platforms move at. Platforms like Twitter have over 500 million users to this current day [28] with billions of data being put out on a constant basis it can rapidly change in response to new news or social trends. With how volatile data can be it can be difficult to maintain consistent predictive accuracy, as sentiment may shift unpredictably. There is always the possibility that public sentiment can vary on each platform.

An example of this may be users on Twitter may display a positive attitude, where attitude on Facebook may slightly more negative, resulting in sentiment that is inconsistent when cross-platform analysis is performed.

Given the above research and case studies, it shows a direct correlation between social media sentiment and stock market movements, making sentiment analysis a key tool for enhancing predictive models. Social media sentiment provides real-time insights into investor and public mood. When combined with traditional financial indicators it can improve the accuracy of stock market predictions. To help improve upon sentiment analysis, developers should focus on cross-platform data integration and improving filtration techniques to help exclude irrelevant noise.

2.4. *Gaps in Existing Research*

2.4.1. *Limited Integration of Social Media Sentiment with Traditional Financial Indicators*

One of the limitations that we can notice is the lack of integration between social sentiment and traditional financial metrics. While social media sentiment has been recognized as a potential indicator of market movements, much of the existing research tends to focus on one or the other and never really look at using a combined approach when predicting market prices.

2.4.2. ***Challenges in Accurately Filtering Noise in Social Media Data***

Social media platforms are filled with irrelevant content, spam and bot activity, which can skew sentiment analysis results. Current research often struggles with filtering out noise and accurately capturing the sentiment that genuinely influences stock prices.
Although some models attempt to account for this noise, there is no standardized method to effectively differentiate between genuine market sentiment and irrelevant chatter, especially given the volume and velocity of social media data.

2.4.3. ***Cross-Platform Sentiment Analysis is Underdeveloped***

The current research often focuses on a single platform like X (Twitter) for sentiment analysis. However, different social media platforms often have different views, behaviors which can influence sentiment differently platform to platform.
There is limited research on cross-platform analysis, where sentiment from Twitter, Facebook, Instagram and reddit is combined to form a more holistic view of market sentiment. Existing models often do not account for variations in sentiment across numerous platforms.

2.4.4. ***Insufficient Focus on Machine Learning Comparisons***

Many studies in the field of stock market predictions often forget to compare the performance of machine learning algorithms. Often, they tend to focus on one or two algorithms. Algorithms like Long-Short Term Memory (LSTM), Support Vector Machines (SVM) and Growing Neural Gas Networks (GNG). This suggests that the need to compare and study algorithms and compare their performance, accuracy, efficiency and scalability in predicting stock market movements.

2.4.5 ***Areas for future research***

Addressing these gaps in current research will help future research and developers improve the current predictive framework. Improving areas such as computational efficiency, accuracy and performance will further produce models that are extremely accurate and robust.

- Throughout this study we will investigate areas of integrating sentiment analysis and traditional financial data.
- Improving techniques to filter noise in social media sentiment analysis, especially in real-time.
- Conducting cross-platform sentiment analysis to capture a more holistic view of investors and public sentiment.
- Utilizing real-time data and advanced Natural Language Processing (NLP) techniques to improve prediction accuracy, even testing the occurrence of extreme events.

These areas listed above provide significant opportunities for research that can contribute to the development of accurate stock market predictive models.

3. **Methodology**

In this part of the study, we will cover the methodology of how this project will be carried out and cover the development lifecycle from start to finish. It will also cover areas such as researching relevant topics, data collection, analysis of data and the results produced from inputting our data into the predictive frameworks. 3 predictive frameworks have been chosen to compare; Growing Gas Neural Network, Support Vector Machine and Long-Short Term Memory.

There are two types of research, those are Qualitive and Quantitative. Quantitative research aims to collect numerical data and conducting mathematical analysis such as predicting and observing trends where Qualitive research will investigate observing trends in language, text and structure. This study will aim to utilize both due to comparing predictive frameworks with one being fed financial data and the other both financial and public sentiment and studying the effects and differences caused by public sentiment.

Research was conducted by analyzing machine learning algorithms from their traditional methods to the current most up-to-date methods used to forecast market prices. Google Scholar was used to help find recent studies that conducted similar or specific research and analysis of certain areas that this study investigates, such as certain machine learning techniques – Support Vector Machines, Growing Neural Gas Networks, Auto-regressive models – and how these techniques are utilized in the field of finance.

Quantitative research

This study will use quantitative research that will be used to obtain data on performance results from each machine learning algorithm. One algorithm will be fed a financial dataset that will contain the most recent financial data with metrics such as time series events, opening and closing share prices. Using a relevant dataset, we will produce graphs that will show the prediction of the next time series event.

GeeksforGeeks have conducted an analysis on Support Vector Machines predicting stock prices [31]. This study used data from Yahoo finance website, with the data being formatted in OHLC (Open, High, Low, Close) in a CSV file. Once data from Yahoo was obtained, it went through data processing, before being edited and ran through the predictive model.

These steps will be applied when it comes to testing, analyzing and modeling our data. However, there are shortcomings to this approach. The lack of technical indicators to help produce more accurate predictions and provide clearer information based on how strong these predictions are. The use of RSI's (Relative Strength Index) will be deployed during testing of our model. Relative Strength Indexes are used in technical analysis to analyze stock prices and

whether or not you should buy or sell stock.
Of course, when obtaining financial data this will require it to be analyzed, cleaned and prepared to be fed through predictive frameworks for forecasting. Within the data set, outliers, missing values and incorrect information are possible and will need to be solved or removed. Techniques involve deletion which will simply remove the missing field. While this is a simple but temporary solution, it can lead to incorrect and misleading predictions if the financial set consistently contains missing data. To combat this, imputation or interpolation can be used, which predict missing values using data that surrounds it.

Qualitive Research

In this study, our qualitive research amounts to the collection of social media sentiment analysis. By using public datasets from Kaggle which are clear of sensitive information, we can analyze and determine which tweet, or text reflects positively or negatively on certain stock prices.
When it comes to analyzing such data, we employ sentiment analysis techniques such as NLP which involves tokenization and Lemmatization. These techniques will shorten words to their root form while keeping the sentiment and structure of sentences. Once this process has been complete, noise reduction will take place. This process simply removes any bot, spam or irrelevant information that doesn't contribute to the overall sentiment. This process also allows for data that is relevant and will give us the best possible result in terms of accuracy and efficiency.
Once both tokenization, lemmatization and noise reduction have been complete, sentiment analysis will begin. Using machine learning techniques like Logistic Regression, Soft Classification will be used. Text will be placed into 3 different classes; Positive, Neutral and Negative. This will order text and give us a holistic view into public sentiment on certain stocks which can then be combined with financial information to be fed into our predictive frameworks.

In conclusion, using both Quantitative and Qualitive research methods allows us to gain accurate data on public sentiment and gain key insights into financial data. Using methods to clean and preprocess data will give us the best possible result in terms of efficiency, accuracy and performance on the chosen machine learning algorithms.

4. **Risk Analysis**

In this section, risk analysis will be performed, helping us identify and evaluate the risks when comparing, modelling and testing our predictive frameworks. Below are risks that could happen when we model our predictive frameworks. Each risk will have a mitigation, to show how we can avoid these risks from happening.

- *Data Quality Risks*: When collecting social media sentiment, there is a lot of noise, fake accounts, spam accounts and a lot of irrelevant posts that could easily affect sentiment analysis scores. Implementing filtering techniques and noise reduction algorithms will help mitigate the risks when it comes to sentiment analysis.

- *Model Overfitting:* Model overfitting occurs when a model learns its dataset too well. This means the dataset will then struggle to predict accurately when new datasets are introduced. A couple techniques to mitigate this is to adjust the dataset in-order to ensure that it is setup for success. Another step to take is to regularize our models [30]. This process handicaps the model's ability to memorize little details found in a training set.

- *API Dependency:* During testing stage of our models, we must be careful that our model doesn't become dependent on data from our API's and can be fed new data without becoming inaccurate when inputting new data. The API's used provide real-time financial data, downtime or limits on downloading such data could cause further delays and impacts on model output and accuracy. By storing previously fetched data locally to reduce the dependency on live API calls.

5. *Data Description and Preprocessing*

In this section we will discuss the sources utilized for this research, along with the methods used for data cleaning, transformation and feature extraction. It is crucial that our data is properly cleaned and handled with care as this can skew results when it comes to testing predictive models, ensuring reliable and accurate stock forecasting where data is noisy, unstructured and often incomplete.

5.1 *Data Sources*

When it comes to implementing the correct data to test our predictive models and practice data cleaning methods using tools like Natural Language Processing. The following data sets will be used:

Financial Data

API's such as Finazon and MarketStack will be used to gather historical stock prices, trading volumes and corporate financial performance. These datasets include daily stock prices - open, close, high, low and adjusted close prices -, trading volumes and corporate earnings reports

Social Sentiment Data

Social sentiment data will be collected from popular social media platforms such as Twitter, Instagram, and Facebook. The focus will be on extracting public sentiment related to companies, industries and economic events which could affect certain companies or market trends.

When obtaining these data sets, data scrapping tools will be used to collect posts, comments and hashtags relevant to the stocks being analyzed. Python libraries – BeautifulSoup, Scrapy and Selenium – can be used to achieve collection of social media sentiment. The use of web API's can also be utilized, API's like; Scraper API, ScrapingBee and Zyte Smart Proxy Manager can be used to handle the technical side of web scraping.

Using approaches discussed in section 2.3 we will use a Lexicon-Based approach for our text analysis which will attach a sentiment score to data giving us an idea of public sentiment and how the market may trend.

5.2 *Data Preprocessing*

During the process of data collection, we need to clean our data from irrelevant context, corrupt or missing text and prepare it for analysis and modeling.

5.2.1 ***Social Media Sentiment Data cleaning and Structure***

Missing values, duplicates, errors and irrelevant records will all need to be removed and converted into a structured format such as a CSV, JSON or XML file which will contain records, adding in labels, categories such as date, time, author and the text

included. Any units found in the data obtained will need to be normalized, converting any units, currencies or ratings. Information based on hashtags, mentions, emojis or links will need to be extracted as these can be used to capture sentiment as they are frequently used on platforms like Twitter, Facebook and Instagram.

Tokenization and Lemmatization

Text will be tokenized into individual words and lemmatization will be applied to reduce words to their root forms. An example of lemmatization would be:

"Are", "is", "being" -> "be"

This helps to retain sentimental value while also keeping the structure of sentences. Stopwords are most common throughout sentences which don't provide any information or helpful information at least. Words like "went", "where", "to", "it" etc., these words will be removed as they don't provide any relevant information that can be used to analyze sentiment.

Noise Reduction

Due to the amount of data produced from these social media platforms, we will filter out irrelevant information from automated bots and content – promotional spam – using key word filtering to remove irrelevant posts.
Sentiment scores will be added from various platforms to avoid any platform biases.

Sentiment Analysis

When performing sentiment analysis, a combination of tools will be used to analyze the data, Machine Learning techniques such as Logistic Regression will be used.
Posts will be split into 3 different types or categories: Positive, Natural and Negative.
Using Soft Classification will be employed to assign probability scores for each sentiment class.

5.2.2 *Financial Data cleaning and Preprocessing*

When it comes to cleaning financial data, it will work differently due to the type of data we are obtaining. First, we will look into the quality of our data, before cleaning. Missing values, outliers and inconsistencies are things that should be looked at and removed or fixed before implementing into our model.

When it comes to handling missing values, a few processes can be used to sort this issue:

- *Deletion*: This will remove rows or columns with missing values. If the quality of our data is not up to par, this can cause issues such as significant data loss if missing values are consistent throughout the data set.

- *Imputation*: This technique will input the correct value where the missing value is located. Using Machine learning models to predict the missing value can help create consistent a reliable data set that can imported into our model.

- *Interpolation*: This technique is used for time series data – data that is collected over time ex: Electrical activity in the brain – can be used to estimate/predict missing values based on its surrounding values.

All the above techniques will be considered and used when it comes to handling and cleaning of financial data. Another process that is important is outliers in stock prices, these outliers can skew analysis and modelling results. Methods like the Z-score method can help prevent skewed model predictions. Z score will be covered in more detail later on in this document.

Data Normalization

When it comes to normalizing our data, we can use various methods to help. Min-Max scaling, Standardization and Robust scaling can all be used to help. We will be using Min-Max to normalize/scale our data, this is essential for models like neural networks that are sensitive to input scales.

Feature Engineering

Feature engineering is the process of creating new features or transforming other features into improving our model. Features like moving averages, Relative Strength Index (RSI) and volatility measures, to capture market trends and price movements. For the experiment the use of Relative Strength Index will be used which plots a visual graph which can be used to analyze data and find patterns. Lag Features which are used for time series data, lag features – previous day's stock – are generated which feed historical context into time-series models like Long-Short Term Memory and Auto-regressive Integrated Moving Average (ARIMA)

APPENDIX

Due to the nature of this study. Project management section has been moved to the appendix. Those who seek to replicate this experiment can do so and check the requirements.

6. Project Management

In this report, the project management will cover 4 main areas. Those are System requirements, Functional and non-functional requirements. Due to the nature of this study, users will not be involved in the process of testing and evaluating a piece of software as the goal is research and test machine learning algorithms that are used to predict stock market prices and the effects of public sentiment from social media platforms such as Twitter, Instagram and Facebook. The MoSCoW method will be used to show which areas are high priority or low priority for functional and non-functional requirements.

M – Must have, S – Should have, C – Could have, W – Won't have

6.1. *System Requirements*

System Requirements will define the technical and hardware requirements that are necessary to run the predictive frameworks and related software.

6.1.1. *Hardware Requirements*

Up to date hardware is crucial to the testing of Machine Learning algorithms; it will allow us to test efficiently and get the best result possible. Hardware should at least meet the requirements of:

- High performance computer with a multicore processor (i7 or AMD Ryzen 9 or higher) allowing for efficient task handling [29]
- Minimum of 16GB RAM. For deep learning models a requirement of 32GB is recommended.
- When it comes to storage, the recommended type of storage are SSDs (Solid State Drives) over Disk Drives. This is due to SSDs having faster access speeds which results in quicker load times. The capacity of an SSD should be at least 512GB and for larger datasets 1-2TB of storage would suffice.
- NVIDIA GPU's (Graphic Processing Units) allow us to use certain capabilities such as: CUDA (Compute Unified Device Architecture) if we plan on using libraries and frameworks optimized for NVIDIA GPUs. When training models and handling large datasets, it is recommended to have a GPU which its VRAM (Video RAM) storage of 8GB – 16GB that allows the use of larger datasets and models.

6.1.2**. *Software Requirements***

- Operating System: Windows 10, macOS or Linux (Ubuntu)
- Python (Version 3.8 or Higher) with libraries like NumPy, Pandas, Scikit-Learn, TensorFlow and PyTorch
- APIs for data collection – Finanzon, MarketStack & Kaggle – to access real-time financial data and download relevant datasets.

6.1.3 ***Network Requirements***

- High speed internet connection to download datasets and updates from APIs

6.2 **Function Requirements**

Functional requirements specify the system's capabilities to meet project objectives, focusing on core functionalities. Using the MoSCoW method mentioned above, this will outline the requirements on what must, should, could and want to have when developing and testing each algorithm and displaying the results.

6.2.1 ***Data Collection***

- The system must be able to collect historical stock prices, trading volumes and social media sentiment data from datasets and from APIs (M)
- Support for real-time data streaming from external APIs (S)

6.2.2 ***Data Preprocessing***

- The system should clean, preprocess and normalize data to remove noise and outliers (M)
- Feature Engineering techniques such as moving averages and volatility measures should be supported. (M)
- Text analysis using Natural Language Processing (NLP) for sentiment analysis of social media. (M)

6.2.3 ***Predictive Modelling***

- Implement multiple machine Learning algorithms such as; Support Vector Machines (SVM), Long-Short Term Memory Networks (LSTM) and Growing Neural Gas (GNG) Networks. (M)
- The system should support model training, testing and validation using real-world financial data and public sentiment datasets (M)
- Ability to visualize model performance through charts and reports (M)

6.2.4 ***User Interface***

- A web-based dashboard for users to view to display stock predictions, sentiment analysis results and model performance metrics. (C)
- Support for downloading reports in formats like PDF and CSV files. (C)

6.3 **Non-Functional Requirements**

In this section, non-Functional requirements cover system performance and reliability.

6.3.1 ***Performance***

- Models must be able to handle large datasets from websites like Kaggle, MarketStack and Finanzon. (M)
- Response time for real-time data analysis should be under 1 second (S)
- The system should handle at least 10,000 data points per second for social media sentiment analysis (S)

6.3.2 ***Scalability***

- The system should be able to handle increasing amounts of financial and social media data. (S)
- Cloud infrastructure could be considered to help with scaling data storage and processing (C)

6.3.3 ***Security***

- All data must be safely secured behind encryption or password protected files during storage and transmission. 2 factor authentication will be considered. (M)
- Data transmission or downloads must be encrypted using HTTPS protocol to protect financial data. (M)

6.3.4 ***Reliability***

- The system must have a guarantee of 99.99% during trading hours (9:30AM – 4:00PM EST) (W)
- Backup and recovery procedures must be in place to ensure data loss does not occur. (C)

6.3.5 ***Usability***

- When displaying financial data, it is important that it is displayed correctly and easy to navigate for financial analysts and non-financial/technical users (M)

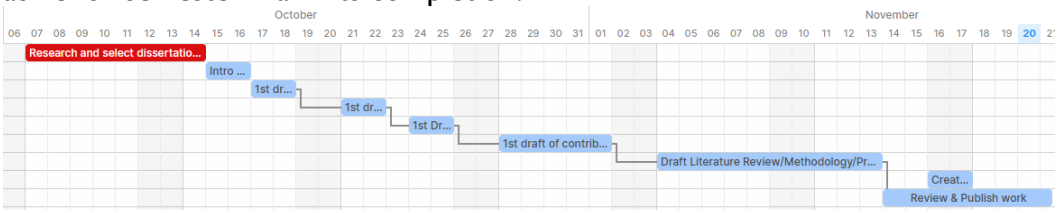
6.3.6 ***Compliance***

- Ensure compliance with data privacy regulations (e.g. GDPR) when handling datasets that are used for public sentiment (M)
- Adhere to the financial industry standards for data integrity and security (M)

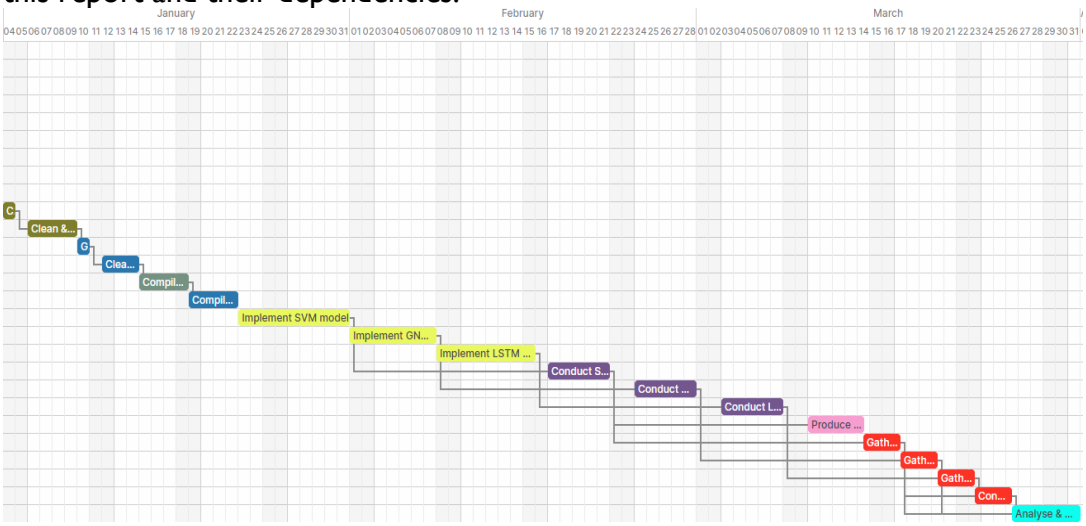
6.3.7 **Gantt Chart**

	Activity	Resource	Status	Start	End		Days	
1	Topic Selection			07-10-24	14-10-24	8.0		8.0
2	Introduction Draft			15-10-24	16-10-24	2.0		2.0
3	Motivation Outline			17-10-24	18-10-24	2.0		2.0
4	Aim Definition			21-10-24	22-10-24	2.0		2.0
5	Objective Setting			24-10-24	25-10-24	2.0		2.0
6	Contribution Analysis			28-10-24	01-11-24	5.0		5.0
7	Literature Review			04-11-24	13-11-24	10.0		10.0
8	Review Drafting			16-11-24	17-11-24	2.0		2.0
9	Final Review			14-11-24	21-11-24	7.5		7.5
10	Collect financial data			04-01-25	04-01-25	1.0		1.0
11	Clean & preprocess financial data			06-01-25	09-01-25	4.0		4.0
12	Gather public sentiment data			10-01-25	10-01-25	1.0		1.0
13	Clean & preprocess public sentiment data			12-01-25	14-01-25	3.0		3.0
14	Compile dataset containing both public sentiment & financial sentiment			15-01-25	18-01-25	4.0		4.0
15	Compile financial dataset			19-01-25	22-01-25	4.0		4.0
16	Implement Support Vector Machine model			23-01-25	31-01-25	9.0		9.0
17	Implement Growing Neural Gas Network			01-02-25	07-02-25	7.0		7.0
18	Implement Long-Short Term Memory model			08-02-25	15-02-25	8.0		8.0
19	Conduct test on SVM using dataset 1&2			17-02-25	21-02-25	5.0		5.0
20	Conduct test on GNG using dataset 1&2			24-02-25	28-02-25	5.0		5.0
21	Conduct Test on LSTM using DS 1&2			03-03-25	07-03-25	5.0		5.0
22	Produce RSI's based on test results			10-03-25	14-03-25	4.5		4.5
23	SVM Performance results			14-03-25	17-03-25	3.0		3.0
24	GNG Performance results			17-03-25	20-03-25	3.0		3.0
25	LSTM Performance results			20-03-25	23-03-25	3.0		3.0
26	Comparitive analysis			23-03-25	26-03-25	3.0		3.0
27	Conclude & Analyse findings			26-03-25	31-03-25	5.5		5.5
								118.5

Details for each task and the amount of time set aside. It also includes the time predicted to set aside for semester 2 and its completion.



Above is the plan for semester one and all the deadlines that were set to achieve completion of this report and their dependencies.



Above is the plan for carrying out the research from this dissertation and conduct, sentiment analysis, cleaning and preprocessing financial and public sentiment data into our appropriate data sets. Once this has been achieved, we will then implement our predictive models and

conduct our testing on each model, producing RSI's, gathering performance results before analyzing and concluding our findings.

In this section, the dissertation will cover PLES, this is Professional, Legal, Ethical and Social. Each section will cover the impact, contributions and address the considerations taken in order to avoid conflicts or breach of GDPR laws.

Professional

This research aims to make a significant contribution to the financial industry. By conducting an experiment that will test 3 main algorithms using 2 data sets where one will contain social media sentiment and financial sentiment. Its aim is to show the effects of combining social sentiment and financial data to make algorithms accurate and improve efficiency of stock market predictions.

1. ***Financial Institutions & Professionals***
   By creating data sets that contain information pertaining to social media sentiment and financial data, it provides the tools to investment analysts to get better, more accurate predictions of market trends.

2. ***Improving Predictive Analytics***
   By comparing 3 predictive models such as Support Vector Machines, Growing Neural Gas Networks and Long-Short Term memory networks. Using 2 different datasets to test each algorithm under different conditions

3. ***Innovating Data Utilization***
   By utilizing social media sentiment, it will use a sector of information that can heavily effect market trends. By combing this data with financial data, it can provide investment analysts a new framework to help with their predictions and decision making.

Legal

When it comes to collecting data, especially gathering public sentiment data. This dissertation will have to be careful where it collects this data ensuring it follows its ethical outline and GDPR laws.

1. ***Data Collection and Compliance***
   When collecting data on public sentiment, datasets from websites like Kaggle will be analyzed making sure there is no sensitive data contained within the datasets selected.
2. ***Data Privacy***
   Collecting data **MUST** comply with GDPR laws when gathering social media sentiment data sets. Anonymization and deletion of data will be implemented to

mitigate privacy risks.

3. ***Ethical Use of predictive Algorithms***
   The aim is to test these algorithms and maintain complete transparency when producing predictions to avoid any biased outcomes by testing the fairness and reliability of predictive model. Ensuring that the predictions produced do not affect financial decision making to avoid liability as predictions might not be 100% accurate.

Ethical

This dissertation will aim to consider the ethical implications when conducting the experiment. Ensuring that the research to be carried out is done so in a responsible manner, respecting stakeholders.

1. ***Social Media Sentiment***
   Any data collected from datasets that contain public information must be handled in a way that doesn't infringe on user rights. To avoid this, transparency is a must and disclosing data usage.
2. ***Mitigating Bias***
   Ensuring that personal and model bias is removed and reliability of each model is tested repeatedly to mitigate bias.
3. ***Stakeholders***
   Ensure that models and predictions used within this dissertation is simply for research purposes and results produced from each algorithm are not to be taken for financial advice.

Social

The aim of this dissertation is to study the effects of public sentiment on stock market predictions by combining financial data and social media sentiment data.

1. ***Public Influence***
   Study how social media sentiment creates market trends and its implications for investors.
2. ***Mitigation of Financial manipulation***
   Discuss concerns of sentiment analysis being misused to manipulate market trends and/or increase panic among investors to buy or sell.

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] Bollen J., Mao H., and Zeng X. (2011). Twitter Mood predicts the stock market. *Journal of Computational Science,* 2(1), pp.1-8. *https://www.sciencedirect.com/science/article/abs/pii/S187775031100007X*

[3] 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]

[4] 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]

[5] Lin, C-C., Jaech, A, Li, X., R.Gormley, M. and Eisner, J. (2021). Limation of Auto-Regressive Models and their Alternatives. [PDF] Available at: https://arxiv.org/pdf/2010.11939
[Accessed 3rd Nov. 2024]

[6] 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]

[7] Brownlee, J. (2023). How to Create ARIMA model for Time Series Forecasting Python. [online] Available at: ***https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/***
[Accessed 5th Nov. 2024]

[8] Jordanova, T. (2022). An Introduction to Non-Stationary Processes [Online] Available at: https://www.investopedia.com/articles/trading/07/stationary.asp#:~:text=Data%20points%20are%20often%20non,cannot%20be%20modeled%20or%20forecasted.
[Accessed 5th Nov. 2024]

[9] 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]

[10] Ounoughi, C. & Yahia, S.B (2023). Information Fusion [Online] Available at: https://www.sciencedirect.com/topics/computer-science/temporal-data
[Accessed 5th Nov. 2024]

[11] Nau, Robert (N/A). Statistical forecasting: Notes on regression and time series anaylsis [Online] Available at: https://people.duke.edu/~rnau/411arim.htm
[Accessed 5th Nov. 2024]

[12] Penn State University (N/A). Moving Average Models (MA models) [Online] Available at: https://online.stat.psu.edu/stat510/lesson/2/2.1
[Accessed 5th Nov 2024]

[13] 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]

[14] 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]

[15] 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]

[16] 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]

[17] Tran Phuoc, Pham Thi Kim Anh, Phan Huy Tam & Chien V. Nguyen (12th March 2024) Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam [Online] Available at: https://www.nature.com/articles/s41599-024-02807-x#:~:text=With%20recent%20research%20trends%2C%20a,suitability%20for%20this%20data%20type.
[Accessed 6th Nov 2024]

[18] Yongming Wu, Zijun Fu, Xiaoxuan Liu, Yuan Bing (23rd May 2023) A hybrid stock market prediction model based on GNG and Reinforcement learning [Online] Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417423009764#:~:text=This%20paper%20proposes%20a%20prediction,GNG)%20algorithm%20in%20unsupervised%20learning.
[Accessed 6th Nov 2024]

[19] Joe Alfredo Ferreria Costa, Ricardo Oliveira (September 2007) Cluster Analysis using Growing Neural Gas and Graph Partitioning [Online] Available at: https://www.researchgate.net/publication/224292862_Cluster_Analysis_using_Growing_Neural_Gas_and_Graph_Partitioning#:~:text=The%20Growing%20Neural%20Gas%20(GNG,according%20to%20the%20training%20dynamics.
[Accessed 6th Nov 2024]

[20] Jan Tunnermann (2021) Artificial Visual Attention based on a growing Neural Gas [Online] Available at: https://getwww.uni-paderborn.de/research/current/gng_attention
[Accessed 6th Nov 2024]

[21] Rohanshah (Jan 22, 2024) The influence of Social Media on Stock Market trends [Online] Available at: https://medium.com/@rohanshah0502/the-influence-of-social-media-on-stock-market-trends-feed4365f64e#:~:text=Because%20social%20media%20posts%20have,may%20disrupt%20traditional%20market%20dynamics.
[Accessed 6th Nov 2024]

[22] Ryan Pendell (Sept 6th, 2022) Customer Brand Preference and Decisions: Gallup's 70/30 Principle Available at: https://www.gallup.com/workplace/398954/customer-brand-preference-decisions-gallup-principle.aspx

[Accessed 6th Nov 2024]

[23] Mayur Wankhade, Annavarapu C.S. Rao (7th Feb 2022), A survey on sentiment analysis methods, applications and challenges Available at: https://link.springer.com/article/10.1007/s10462-022-10144-1#Fig6
[Accessed 7th Nov 2024]

[24] IBM, what is logistic Regression (LR) Available at: https://www.ibm.com/topics/logistic-regression
[Accessed 7th 2024]

[25] Yang Gao, Chengjie Zhao, Bianxia Sun & Wandi Zhao (23rd August 2022) Available at: https://jfin-swufe.springeropen.com/articles/10.1186/s40854-022-00381-2#:~:text=Specifically%2C%20Gong%20et%20al.,can%20significantly%20exacerbate%20price%20jumps.
[Accessed 7th Nov 2024]

[26] Al Root, George Glover (6th Nov 2024) Elon Musk is Going to Washington. Why Tesla Stock Is Soring Available at: https://www.barrons.com/articles/tesla-stock-musk-trump-election-rivian-1086ba9f
[Accessed 7th Nov 2024]

[27] 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]

[28] Brian Dean (17th July 2024) X(Twitter) Statistics: How many People Use X? Available at: https://backlinko.com/twitter-users
[Accessed 9th Nov 2024]

[29] GeeksForGeeks (6th Aug 2024) Hardware Requirements for Machine Learning. Available at: https://www.geeksforgeeks.org/hardware-requirements-for-machine-learning/
[Accessed 15th Nov 2024]

[30] Daliana Liu, Geoffrey Angus (31st July 2023) How to Guide: Overcoming overfitting in your ML models. Available at: https://predibase.com/blog/how-to-guide-overcoming-overfitting-in-your-ml-

models
[Accessed 16th Nov 2024]

[31] 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]

[32] Geeks For Geeks (18th June 2024) Predicting Stock price Direction using Support Vector Machine. Available at: https://www.geeksforgeeks.org/predicting-stock-price-direction-using-support-vector-machines/
[Accessed 18th Nov 2024]

[33] Kaggle available at: https://www.kaggle.com/datasets

[34] Finanzon Available at: https://finazon.io/

[35] MarketStack Available at: https://marketstack.com/