National College of Ireland

Data to Dollars

MSc Research Project
MSc Artificial Intelligence for Business

Muhammad Waqas Javed
Student ID: X23168064

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Muhammad Waqas Javed		
Student ID:	X23168064		
Programme:	MSc AI for Business	**Year**	2023-2024
Module:	MSc Research Practicum/Internship part 2		
Supervisor:	Rejwanul Haque		
Submission Due Date:	12-08-2024		
Project Title:	From Data to Dollars		
Page Count:	37		
Word Count:	12000		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Muhammad Waqas Javed
Date:	11-08

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

From Data to Dollars

Muhammad Waqas Javed

X23168064

Abstract

Cryptocurrencies and Stock Markets are a hot new way of investing into and earning some big bucks. They have changed the way of how we perceive finances and how the traditional transactional or investment approaches are now replaced by thousands of stock options, and 13,000+ cryptocurrencies around the world. With a market cap of several hundred Trillion dollars, these financial instruments have attracted researchers to analyse trends, and read patterns convert that data into dollars. Capturing price values changing with time, hours after hours and weeks after weeks sometimes may not be enough, and here we also show how social sentiments affect the actual market along with the financial market trends themselves. We demonstrate the pros and cons of various Time Series models and how they perform with and without Twitter sentiments for predicting prices of a stock, generalising over capabilities for all Time Series analysis.

1. Introduction

Cryptocurrencies and traditional stock markets have both had a major impact on the financial world. With the help of blockchain, the new digital means of payments in the form of cryptocurrencies and their flagship – Bitcoin – have appeared in the world, created by the mysterious Satoshi Nakamoto. Cryptocurrencies, however, are digital currencies that are highly Liquid and are not controlled by any central authorities as unlike the fiat money. The market demand for these had increased drastically with over 13,000 cryptocurrencies currently in the market and a total market capitalization greater than $2. 5 trillion. It has drawn many participants in the form of investors, technologists, and researchers because of promises of higher returns and digital decentralised finance.

In contrast, stock market refers to a well-established financial system that involves trading in shares of public limited companies. It is mature and legal, hence competitive; prices are determined by position of companies, economic factors and the mood on the stock market. Nonetheless, both cryptocurrencies and stocks are traded in the market and depend on buyers and sellers' actions and reactions to economic news and shifts in policies.

1.1 Problem Setting

The first research question, based on the aforesaid primary issue, pertains to the identification of methods that can be used to better forecast price fluctuations in both crypto and stock markets. While analysing high-frequency data, it is often possible to employ approaches that are typical for statistical mechanics, yet conventional statistical methods seem to be useless for characterising non-linear dependencies typical for these markets. However, cryptocurrencies are particularly volatile and can be affected by changes in market sentiment and news or macros

events. Likewise, the stock market is determined by aspects including the companies' results, economic factors, global affairs, and sentiments of investors.
For research stakeholders to overcome these problems, this work reviews the use of advanced machine learning and deep learning methods including sequence processing neural networks like Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Gated Recurrent Unit (GRU). These models are appropriate for dealing with time-series data which has a temporal dimension and also can consider temporal relations that are useful for the identification of future tendencies in the prices of cryptocurrencies and stocks.

1.2 Background

The existence of cryptocurrencies and the constant existence of the stock market speak a lot about fluctuating financial structures. Cryptography is the science known for the foundation of many cryptocurrencies, which implies the technology of a decentralised ledger that cannot be successfully manipulated or changed by any third parties. The characteristics of this technology like irreversibility, consensus models have placed this technology as a fundamental pillar in the digital finance world.

In parallel, the stock market has sustained its development with such technologies and financial instruments. Reports show that application of machine learning and artificial intelligence in financial markets, in particular the cryptocurrencies and stocks has advanced considerably. These technologies can be used to build models that use past data to uncover trends and even potential future trends in price movements. Sequence driven neural networks serve well the purpose in this context because they can accept sequential data and can remember the prior state which is useful for modelling of temporal dependencies which is present in financial time series data.

1.3 Research Questions

This research is aimed to be guided through the following questions:

1. Which of the Machine Learning models can accurately analyse previous trends and or fluctuations and predict the future changes of price in either the Cryptocurrency and or the Stock market?
2. Is adding a component of public sentiments helpful for understanding the financial time series data better?

These questions are intended to determine how individual methodologies have been used to forecast prices and recognise the advantages and disadvantages of employing modern approaches such as ML and DL in both markets.

1.4. Relevance of the Study

The information gathered from this study benefits academia in addition to the communities of finance practitioners. It is evident that the accurate prediction of price changes in cryptocurrencies and stocks is one of the critical things that investors, traders, or even financial analysts consider. Such predictions can be useful in decision-making, minimising the risks involved and at the same time help to get the most out of the investments. Moreover, as this investigation is related to the development of using enhanced structures of artificial neural networks for prediction in the sphere of finance, this research enhances the body of existing knowledge. The study results may be useful in understanding similar markets in very generic terms depending on the characteristics of the advertisement and user's/investors' sensitivities to any changes in it.

1.5 Incorporating Twitter Sentiments in Predicting Prices

In recent years, social media platforms, particularly Twitter, have become significant sources of real-time data that influence market trends and investor behaviour. The high volume of user-generated content on Twitter provides valuable insights into public sentiment, which can affect the prices of cryptocurrencies and stocks. This section discusses the role of Twitter sentiments in financial prediction and the methods for incorporating this data into predictive models.

1.5.1 Relevance of Twitter Sentiments

Real-Time Data: Twitter, in recent years, as well as other social networks, can be considered as one of the important sources of real-time data affecting market and investors. Due to many accounts and the high activity on the platform, it is possible to analyse the tendency of readers and analyse their impact on the stock prices and cryptocurrencies. This section provides the reader with an overview of the position that Twitter sentiments hold when it comes to economic prediction as well as approaches that can be used to incorporate such data into the model.
Wide Reach and Influence: Quick reaction of the market is a valuable factor and twitter provides it in the form of quick reactions of the people. For instance, a change in the regulatory status that is likely to affect the market sentiment often changes the prices almost immediately
Sentiment Analysis: Twitter is one of the best tools when it comes to observing market reactions in real-time, especially when the market is very active such as cryptocurrencies and stocks. For instance, a new regulation that was recently effective in the market changes market sentiment and the price levels.

1.5.2 Methods for Incorporating Twitter Sentiments

1. **Data Collection**: Collect a dataset of tweets with specific focus on the cryptocurrencies and stocks of interest using hashtags, keywords or mentions.
2. **Pre-processing**: Pre-processing and cleaning the data which involves the removal of unnecessary information like stop words and punctuations, tokenizing the given text, lemmatization and stemming.
3. **Sentiment Analysis**: Use NLP tools that enable the categorization of different tweets as either positive, negative, or neutral. This can be done with the help of ready-made models or creating some custom ones, which will provide the sentiment scores or labels.
4. **Feature Engineering**: To the basic sentiment scores, add other pre-processed numeric data, such as past stock prices, trading volumes and simple moving averages for model construction.
5. **Model Training and Evaluation**: Use this enhanced dataset to train modalities like LSTM or hybrid modalities and assess their efficiency using quantized parameters like MAE (Mean Absolute Error) or MSE (Mean Squared Error).

1.5.3 Use Cases and Examples

- **The "Elon Musk Effect”**: Elon Musk has made an example of the impact of the message on shares, literally in social networks. For instance, a positive tweet on Dogecoin boosted its prices, thus proving that, indeed, social media sentiment can influence the market outcome.
- **Regulatory News and Market Reactions**: Due to the availability of platforms such as Twitter, news regarding regulatory action, including crackdowns on cryptocurrencies, gets disseminated fast and acts as a catalyst to market sentiment and prices.

- **Sentiment-Driven Investment Strategies**: There are investment firms that include the analysis of sentiment from social media to trade and detected shifts in sentiments to inform trading activities.

1.5.4 Challenges and Considerations

- **Data Quality and Noise**: The pool of data extracted from the Twitter environment may contain large amounts of residual and low-quality information that needs to be sorted out.
- **Sentiment Ambiguity**: Complicated language, for instance, sarcasm, might be misjudged by sentiment analysis, depending on how it was programmed.
- **Rapid Sentiment Shifts**: Thus, the sentiment on popular social networks can fluctuate often and the model needs to make continuous adjustments.
- **Market Manipulation Risks**: The ability to bring details regarding market manipulation through sprint social media-based campaigns is a real possibility, hence the need for tough model hedges.

1.6 Conclusion

Integrating Twitter sentiments as a unique feature for forming the predictive models for cryptocurrencies and stocks presents the worth and improvement on existing measures estimations. This would allow for the incorporation of market sentiment with classical financial factors in real-time if the latter is using real-time data coupled with better NLP. However, accuracy of the data used, analysis of the sentiment, as well as incidences of manipulation of the market pose a threat to accurate prediction. With the development of sentiment analysis and machine learning methods, the use of social media data in finance modelling is expected to have more developments and applications in the future.

1.7 Structure of the Thesis

The structure of our dissertation is designed to build a comprehensive understanding of the data we use, the methodology that we employ and the findings that we have to help relate how Neural Network architectures are used to predict stock prices and cryptocurrency values, for a series of time steps. Our work is organised in the following chapters:

1. **Introduction:** To discuss the background and motivation for the study. Here, we outline the problem statement, the objective of the research and the significance of doing this project on the financial markets.
2. **Literature Review:** To review and discuss the existing work in this field. We cover the various approaches, from traditional statistical approaches to modern ML techniques. Here, we highlight gaps in the existing literature.
3. **Dataset Overview:** An overview of what datasets we are using, how they will be of value and how we preprocess it to be more effective.
4. **Methodology:** This chapter details our methodologies that we employ in our research. It includes the discussion on each of the applied models and the rationale between choosing each for the configuration of our experiments.
5. **Experimentation and Results:** This chapter is the core of our research where we discuss and present our results and findings and evaluate each model in different settings.

2. Related Work

The study of cryptocurrency markets and their trends is an increasingly prominent area of interest, just like the stock market once was. This is driven by the rapidly evolving and the growing adoption of digital assets. In this chapter, we provide a comprehensive review of the existing work done as we explore the conducted methodologies, along with discussion of the theoretical ideas that have been either discussed or applied to predict values of cryptocurrencies in the future. Our review and our work dives into the complexity of cryptocurrency valuation setting, the impact of time series factors and the role of sentiments attached to a cryptocurrency.

In this chapter, we also aim to appreciate and review the growing importance of Machine Learning and Deep Learning for making financial predictions possible, focusing on sequence-aware Neural Network architectures. Through the synthesis of current knowledge and the gaps that exist when combining sentiments to mere time series data, we aim to set a foundation of subsequent empirical investigation. This project's aim remains to be that it wants to contribute to the goal of predictive modelling, which in our context remains to be highly volatile and speculative cryptocurrency markets and stock markets data with its impact evaluated with and without market sentiments embedded.

2.1 Machine Learning Approach to Value Predictions

The first study that we see provides an explanation of how cryptocurrency trends are not seasonal and hence are hard to predict statistically. In order to model this problem statistically there need to be made several noisy and unrealistic assumptions. Hence, in their research, Khedr et al. (2021), discuss using Bayesian Regression, Linear Regression or Support Vector Machines. Derbentsev et al. (2021), also verify the applicability of Machine Learning techniques in predicting cryptocurrency values of Bitcoin, Ripple and Ethereum using Random Forest and Stochastic Gradient Boosting Machine. They touch a Mean Absolute Percentage Error (MAPE) value ranging between 0.92% and 2.61%. Another study done by Akyildirim et al. (2021), claims to have produced better results through Random Forests, as compared to ANNs on the weekly analysis of Bitcoin based on an hourly frequency dataset. By collecting the data that included seven-day a week daily information by obtaining from https://coinmarketcap.com/, Chowdhury et al. (2020) work on predicting the closing price of the cryptocurrency Index 30, and nine constituents of the cryptocurrency. Through a strong comparative analysis of conventional Machine Learning approaches they found that the best approach is to ensemble all methods that outperform the best of works from the literature. Gupta & Nain, (2021) predict the value of Bitcoin through the use of several key features like the amount of cryptocurrency in circulation, the volume of the cryptocurrency exchanged each day and the overall demand. They do this forecasting through a series of comparative models starting from the Time Series techniques like Moving Average, ARIMA to Linear Regression, Support Vector Machines, and ending at Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). In a more simplistic yet effective approach, the authors in Alessandretti et al. (2018) discuss the use of nontrivial but simple algorithmic techniques with state-of-the-art Machine Learning algorithms that can outperform all standard benchmarks. Most of the methods applied use the adoption of a Simple Moving Average strategy with feature-target pairs trained over a Recurrent Neural Network, and over the XGBoost Algorithm. A research on the stock market dataset compares five algorithms, the K-Nearest Neighbors, Support Vector Regression, Linear Regression, Decision Tree Regression and Long Short-Term Memory that predict stock prices of 12 leading Indian stock market companies The authors find that while Machine Learning approaches may be convenient, the Deep Learning based approach always outperforms them as discussed by the authors in Bansal et al. (2022). Another

research focused on bridging the gap of using Gaussian Naive Bayes in Time Series models. The authors combine feature extraction and feature scaling with stock price movements to show how the GNB algorithm with LDS, PCA and Min-Max scaling produces better performance than previous standard techniques Ampomah et al. (2021).

The review of these above-mentioned previous research will help a lot in addressing the first research question that talks about the best ML algorithms that can possibly be applied. As we have seen, conventional Machine Learning algorithms, while acting as a strong starting point are actually limited in performance due to their statistical nature; that defines everything linearly. This limits the dynamic, multidimensional nature of cryptocurrency trends and hence calls for Deep Learning implementations that will be reviewed in the next subsection.

2.2 Sequence-Driven Deep Learning Approach to Value Predictions

Li & Dai (2020) propose the usage of a hybrid CNN-LSTM architecture that leverage CNN for feature extraction and putting in those feature vectors into LSTM for training and forecasting the short-term price values of Bitcoin. Apart from the transactional data, the researchers also used external information like macroeconomic variables and investor attention. They reported an MAE value of 209.89 which is much lesser than standalone models, and also achieved a F1-score of 0.69. Another study by Cavalli and Amoretti (2021) focuses just on Bitcoin where they illustrate a novel approach based on a One-Dimensional CNN. They claimed that a scalar, one dimensional window of convolutions can heavily outperform LSTM's memory and forgetting mechanism. Similar to the work by Li and Dai (2020), Mounika et al. (2021) also proposed using a hybrid CNN-LSTM approach in practice. However, Hamayel & Owda (2021), conducted a more comprehensive study of what models to apply to this problem, they use models like GRU, LSTM, Bi-LSTM and also propose using ARIMA which is a common stock-market prediction algorithm. Their contribution is not limited to only Bitcoin, they show implementations on Litecoin and Ethereum as well. These particular researchers reviewed above discuss and illustrate the evolving focus towards Deep Learning approaches that align with the non-linearity, and sequentiality (that captures trends) in the data. Jaquart et al. (2022) present a trained series of Sequential models for predicting the binary relative daily market movements of the 100 largest cryptocurrencies. However they discuss and show how a long-short portfolio strategy outperforms the buy-and-hold benchmark strategy by using a Long-Short Term Memory (LSTM) and a Gated Recurrent Unit (GRU). The average accuracy values near 60% and the out-of-sample Sharpe ratio yields an annual cost of 3.23 for the LSTM, and 3.12 for the GRU. To overcome the traditional methods of using K-line diagrams, the authors in a study discuss a new technique for stock forecasting, this includes adding technical indicators, investor sentiment indicators, and financial data with dimension reduction with LASSO and PCA approaches. The experiments are done using GRU and LSTM where the LASSO approach yields better results than the PCA as discussed in Gao et al. (2021). In terms of stock market forecasting, the a study evaluates different Deep Recurrent Neural Network based architectures like the LSTM and GRU, and test both, bbi-directional and unidirectional, with multivariate inputs. When compared with shallow neural networks over the S&P 500 index data, the stacked LSTM architecture gives the highest forecasting results as we see in Althelaya et al. (2018).

We can see that previous literature confirms the strengths of Deep Learning, and we will use these findings to further lay foundations of our own study and showcase the performance efficacy of different, simple and hybrid, Sequential Neural Networks to analyse and predict future trends of highly volatile entities like crypto currencies, and stock markets.

2.3 Research Niche and Twitter Sentiments

While existing literature plays an important role in exploring, building and reporting various useful approaches that predict cryptocurrency prices and values. There is a series of studies that also suggest that Twitter posts and sentiments can impact cryptocurrency prices as well. In their research, Kraaijeveld and De Smedt (2020), conducted causality tests to indicate that the sentiments on twitter have predictive powers that affect prices, but it is majorly focused on just the return values of Bitcoin and Litecoin. To make a sturdy forecasting model, Parekh et al. (2022) say that all cryptocurrencies belong to the same class, and thus we should be able to infer that change in one cryptocurrency's price may lead to changes of the other too. The authors propose a hybrid framework called DL-GuessS, which looks for interdependence on other cryptocurrencies and the sentiments of the market. They considered the price prediction of Dash, using its price history, and tweets of not only Dash, but also Litecoin and BitCoin for loss function validation. They also check the performance of DL-GuessS on other cryptocurrencies by inferring results for price prediction of Bitcoin-Cash using histories of Bitcoin-Cash, Litecoin and Bitcoin. In a study the authors apply sentiment analysis and Machine Learning principles for finding correlation between public sentiments and market sentiments. The authors use the twitter data for predicting public mood and apply to the previous days' DJIA values for prediction of movements in the stock market. The cross validation performed over the financial data gets an accuracy of 75.56% and uses Self Organizing Fuzzy Neural Networks as in Mittal & Goel, (2012) .

Reviewing all of this previous work in this complex domain we have realised that there still remains a significant gap in comparing and contrasting different Sequential Neural Networks and building Hybrid approaches that build the best of different models together. In our research, we will use a feature engineering approach to create more comprehensive and the most relevant features only to add indicative factors that affect the performance of our solution; and we will be exploring the development of various Sequential Neural Networks like RNN, LSTM and GRU; including their hybrids and including a solution that promises robustness across time series data like the cryptocurrencies and stock market data. A critical part of our study would be to incorporate social media trends, investors sentiments and time-series trends together to forecast values based on several statistically unrelated but truly related factors for predictions.

3. Research Methodology

In our work where we explore cryptocurrency price prediction and how social media sentiments are integrated for analysis, we use two distinct but comprehensive datasets. The first dataset, *Bitcoin +233 Crypto Coins Prices*, holds the historical price and volume data for many cryptocurrencies. The second dataset, Stock Tweets for Sentiment Analysis and Prediction, captures the sentiments that are expressed in tweets regarding different stocks. It provides an interesting and invaluable relativity of how sentiment analysis can impact asset valuation. In this chapter, we will discuss the intricacies of these datasets, the data preparation and preprocessing and the relevance they have to the study.

3.1 Bitcoin +233 Crypto Coins Prices Dataset

The *Bitcoin +233 Crypto Coins Prices* dataset offers an extensive collection of the trading data from history of 234 most traded cryptocurrencies. The dataset is pivotal in nature as it analysed price movements and helps understand market dynamics to develop predictive models. The dataset is sourced from the Binance Exchange. The dataset is structured with key financial metrics like the Open, High, Low, Close and Volume (OHLCV) prices. These metrics are also made available across various timeframes like the weekly (W1), daily (d1), hourly (H1) and in

even more granular intervals like 30-minute (M30), 15-minute (M15) and 5-minute (M5) incremental timeframes.

3.1.1 Data Structures and Data Attributes

The dataset comprises several very essential columns that each represent a critical aspect of the trading data. The data structures:

- **Datetime:** This column records the exact date and the time when that data point was captured which is crucial for time series analysis.
- **Open:** This is the price of the cryptocurrency at the beginning of that specified time interval.
- **High:** This is the highest price that was recorded during that specified time interval
- **Low:** This is the lowest price that was recorded during that specified time interval
- **Close:** The price of the cryptocurrency at the closing of that specified time interval.
- **Volume:** The total amount of the cryptocurrency traded during the specified time interval which indicates the size of the market activity.

These attributes are captured for each of the cryptocurrency in the dataset. They enable comprehensive analysis taken across multiple dimensions. For instance, the dataset has the trading information for top cryptocurrencies such as Bitcoin (BTC USDT), Ethereum (ETHUSDT) and Binance Coin (BNB USDT) and many more. The breadth of data spans from the initial records of 2017 to the early 2024. This offers a substantial temporal window that allows historical and trend analysis.

3.1.2 Data Acquisition and Sources

The dataset was obtained directly from the Binance Exchange through the use of a set of Python scripts that ensured accurate and up-to-date data collection. The data is granular. Its granularity ranges from hourly to weekly intervals and allows for us to conduct some flexible analysis that will accommodate both short-term and long-term predictive models.

3.1.3 Applications in Predictive Modelling

The OHLCV data is instrumental for constructing the predictive models. It is particular for time series and sequential data forecasting. Through the use of historical price data and trading volumes these models can be trained and taught to anticipate future price movements based on past trends. This aids us in investment decisions and risk management. The data is diverse and covers a big range of cryptocurrencies that also facilitates cross-sectional studies and comparisons between different digital assets that are drawn for comprehending the broader market behaviours.

3.2 Stock Tweets for Sentiment Analysis and Prediction Dataset

The dataset chosen for consideration called Stock Tweets for Sentiment Analysis and Prediction discusses the relation between social media sentiment and financial market fluctuations. It has more than 80,000 tweets in relation to stock market including focus on the top 25 most active stock tickers according to yahoo finance. This dataset is complemented by price and volume data for the stock market making this data set particularly valuable for analysis of the impact of public opinion on the market.

3.2.1 Data Structure and Attributes

The dataset is divided into two primary components:

1. **stock_tweets.csv:** This file includes the following columns:

- **Date:** The time and date when the specific tweet was sent to the official Twitter account of the company or organisation.
- **Tweet:** The actual text of the tweet providing the original material for the Sentiment Analysis.
- **Stock Name:** This is the name of the stock which was mentioned on the tweet.
- **Company Name:** The actual name of the company which corresponds to the mentioned stock market symbol.

2. **stock_yfinance_data.csv:** This file contains the stock market data corresponding to the stocks mentioned in the tweets, with the following columns:
   - **Date:** The specific date on which stock data entry will be recorded.
   - **Open:** The initial price that is offered for the stock in the market on the specified date.
   - **High:** The maximum price observed during the given trading day.
   - **Low:** The lowest price of the given stocks during the trading day.
   - **Close:** It represents the final price of the stock in the trading session.
   - **Adj Close:** The adjusted closing price, accounting for dividends and stock splits.
   - **Volume**: The number of shares of stock that is bought and sold within the market.
   - **Stock Name:** The unique identifier as it appears on the stock market's ticker.

3.2.2 Data Acquisition and Sources

The tweets were gathered from Twitter, concentrating on the 25 top stock symbols according to the Yahoo Finance watchlist, from September 30, 2021, to September 30, 2022. This period includes different market conditions, thus offering a range of characteristics that may be used for analysis. Yahoo Finance was used for the market data corresponding to the respective stocks to obtain accurate and consistent data.

3.2.3 Application in Sentiment Analysis and Prediction

The main use of this set is in the field of opinion mining. Through the positive or negative tweets, sentiment determines the situation in relation to specific stocks of stock markets. This can than be linked to real market swings and provide insight into how fundamental changes influence the stock.

For instance, if there is an inflation in the number of posts that have an optimistic view of a certain stock, it is advisable to buy it as the market seems to be bullish. Alternatively, if there are many negative posts then it may be interpreted that the expectations of the market are bearish. It increases the value of the predictive models when used in coordination with the standard financial ratios giving better understanding about the factors that are influencing the markets. The Table 1 demonstrates samples from our dataset that help us incorporate market sentiments with the market trends.

Date	**Tweet**	**Stock Name**	**Company Name**
2022-09-29 23:41:16+00:00	Mainstream media has done an amazing job at brainwashing people. Today at work, we were asked what companies we believe in & I said @Tesla because they make the safest cars &	TSLA	Tesla, Inc.

	EVERYONE disagreed with me because they heard“they catch on fire & the batteries cost 20k to replace”		
2022-08-09 13:45:01+00:00	4/ When it comes to growth stocks, as we have seen in the past 7 months, you have to be able to stomach the BIG moves both up and down Here are 10 popular Growth stocks: $GOOGL $TSLA $AMZN $META $NVDA $AMD $CRM $DIS $ABNB $PYPL	AMZN	Amazon.com , Inc.

Table 1: *Example tweets and how they can significantly impact the values or inclinations of people buying or selling a stock based on public sentiments on X.com (formerly Twitter).*

3.3 Integration and Challenges

The integration of these datasets offers certain advantages and at the same time it includes certain difficulties. It has been seen that the cryptocurrency price datasets offer strong fundamentals to analyse the market trends and construct the price forecasting models. At the same time, the sentiment of data obtained as a result of using the Twitter API also helps to expand the understanding of market activity by covering the psychological aspect. Thus, it can be seen that threats such as low quality of the data under analysis, presence of noise, and ambiguity as to what constitutes sentiment could hamper the reliability of the prediction.

Given the fact that the Bitcoin +233 Crypto Coins Prices and Stock Tweets for Sentiment Analysis and Prediction contain arrays of variables, the foundation for higher-level modelling is laid. As a result, they have used historical price and volume data along with real-time sentiments derived from social media to identify and model a complex features of financial markets. The specifics of data pre-processing, sentiment analysis and predictive model implementation will be described in the following chapters, referencing the practical innovative solutions and challenges that have been met along the way.

3.4 Data Engineering and Pre-processing

We take into account the prices dataset first. The first step is to parse the date and time to extract various temporal components that we use for feature engineering. Year, Month, Day, Hour and Minute were extracted as well as the Day of the Week (0-6).

3.4.1 Feature Engineering of prices dataset

For the enrichment of the dataset, we added the technical indicators and derived new features, Mann, (2022). They are:

- **Relative Strength Index (RSI):** A momentum oscillator that is able to measure the speed and change in the price movements. The calculation is done using a window of

14 periods that have values ranging between 0 and 100. The RSI value helps to identify overbought or oversold conditions of the market.

- **Price Differences:**
  - **Open-Close:** The difference of the opening and the closing prices. This shows the net price change within the trading period.
  - **High-Low:** The difference between the highest and lowest prices that provide a measure of volatility in this period.
- **Previous Price Differences:**
  - **O - PO:** Difference between the current opening price and opening price of the previous period.
  - **H - PH:** Difference between the current high price and current high price of the previous period.
  - **L - PL:** Difference between the current low price and current low price of the previous period.
  - **C - PC:** Difference between the current closing price and current closing price of the previous period.

To finally ensure that the features make equal contribution, we apply normalisation. It also expedited the training process. The MinMaxScaler is used to scale the features to a range between 0 and 1. All columns of the dataset go through this transformation and ensure that variance in the scales of data does not add or influence any bias to the model's learning.

3.4.2 Stock Tweets Dataset Preparation

The dataset contains the tweets related to several stocks. For the sake of this experiment, we primarily focus on the "AMZN" (amazon) stock for evaluation and analysis. We filter the dataset to only hold the records regarding "AMZN".

We quantify the sentiments that are expressed in each tweet using the VADER (Valence Aware Dictionary for Sentiment Reasoning) model. VADER is a lexicon and rule-driven sentiment analysis tool that is specially attuned for sentiments expressed in social media.

For each tweet, VADER gives us for key metrics, Elbagir & Yang, (2019):

- **Compound Sentiment Score:** A single value that represents the overall sentiment ranging between -1 (very negative) to 1 (very positive).
- **Negative:** The proportion of text that is seen as negative.
- **Neutral:** The proportion of text that is seen as neutral.
- **Positive:** The proportion of text that is seen as positive.

In this part, a separate stock prices dataset is also used. It comprises the columns Date, Open, High, Low, Close, Adj Close and Volume.

For each of the tweets in the sample, we integrate those sentiment scores with other financial details of the stock that day. To enhance this dataset's robustness further, we calculate some key technical indicators.

- **Moving Average (MA):**
  - **MA(7):** The average closing price over the past 7 days
  - **MA(20):** The average closing price over the past 20 days
- **Exponential Moving Average (EMA):** A weighted moving average that gives more significance to recent prices.

- **Bollinger Bands:**
    - **Middle Line:** standard deviation value of the 20-day moving average
    - **Upper Band:** MA(20) plus two standard deviations
    - **Lower Band:** MA(20) minus two standard deviations
- **Moving Average Convergence Divergence (MACD):** the difference between the 26-period and the 12-period EMAs.
- **LogMomentum:** The logarithm of the difference in the closing price. This helps to capture the momentum of the stock price changes.

See Fig 1. for the comparison of MA(7) with MA(20) for mapping the trend. See how MA(7) performs over the MA(20) for modelling the trend well.

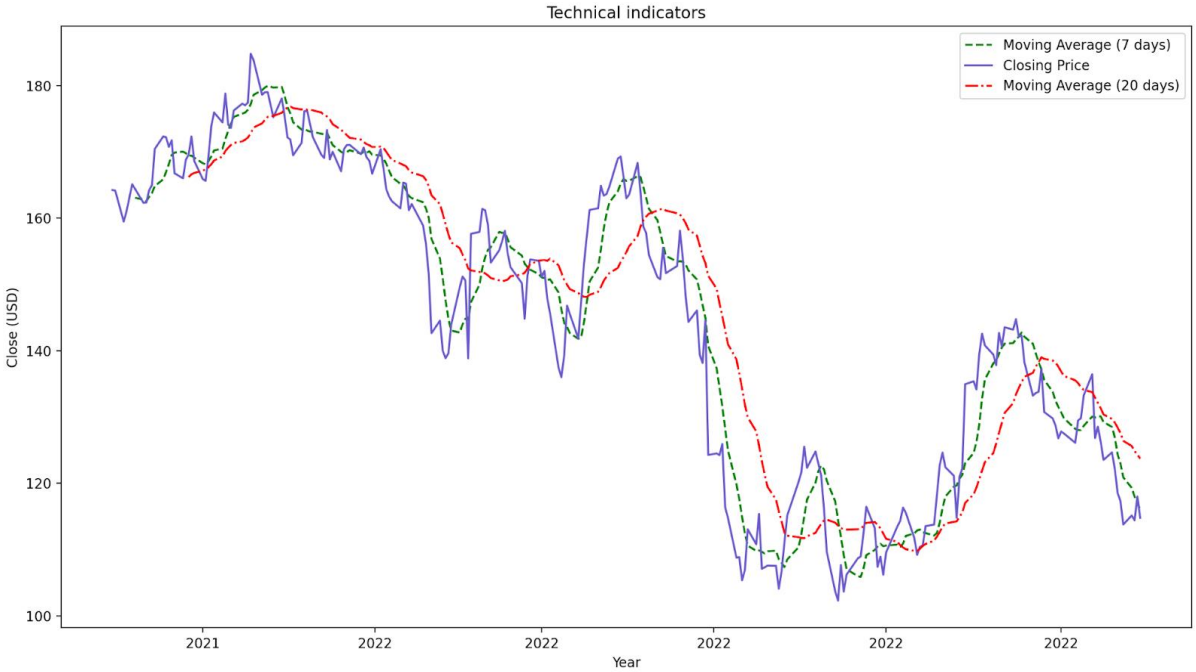

Fig 1: *A comparison of how the Moving Averages (a) MA(7) and (b) MA(20) map against the actual moving trends of the closing prices of the Amazon stock.*

4. Design Specification

The advanced Machine Learning and Deep Learning strategies that we employed here in this research are listed, expanded and explained in this chapter. This chapter will cover the complete discussion of the theory, applications, advantages and limitations of each of the models in the context of how we use them for cryptocurrency and stock market prediction.

4.1 Basic Models

4.1.1 Recurrent Neural Network (RNN)

4.1.1.1 Theory and Architecture

The Recurrent Neural Networks are a class of the Artificial Neural Networks and they are designed to process the sequential data. Unlike the feedforward artificial neural networks, the RNNs have connections that are able to form directed cycles. These directed cycles allow them to maintain an internal state or “memory” of the previous inputs Zhang & Man, (1998). This

makes the RNNs more suitable, in particular, for solving time series problems such as the financial market prediction, which applies to our case.

The basic architecture of an RNN model can be described as follows:

Given an input sequence $\boldsymbol{x = (x_1, x_2, \ldots, x_t)}$,
An RNN model computes a sequence of the hidden states
$\boldsymbol{h = (h_1, h_2, .., h_t)}$ and outputs
$\boldsymbol{y = (y_1, y_2, \ldots, y_t)}$ as follows:

$\mathbf{h_t} = \boldsymbol{\sigma(W_{hh}h_{t-1} + W_{xh}x_t + b_h)}$
$\mathbf{y_t} = \boldsymbol{\sigma(W_{h\gamma}h_t + b_\gamma)}$

Where:

- $\boldsymbol{\sigma}$ is an activation function, which is often tanh or ReLU.
- $\mathbf{W_{hh}}$, $\mathbf{W_{xh}}$, and $\mathbf{W_{h\gamma}}$ are all weight matrices.
- $\mathbf{b_h}$ and $\mathbf{b_\gamma}$ are both bias vectors.

The main feature of the RNN is the recurrent connection that is represented by $W_{hh}h_{t-1}$, it allows the information to persist across time steps Sherstinsky, (2020).

4.1.1.2 Applications to Financial Time Series

In the context of cryptocurrency, and making stock market price predictions, we use the RNNs to model temporal dependencies to map price movements. As an example:

Input(x_t): A vector that represents various features at time t, such that:

- Closing price of the previous day,
- Trading volume at that time step,
- Other technical indicators, like Moving Average.
- Sentiment scores from the social media platforms.

Output(y_t): The predicted price of the next time step, or the price movement.

The RNN model processes all of these inputs sequentially and updates the hidden state at each time step. This allows it to capture the evolving market dynamics and integrate into the mathematical-only analysis of trends.

4.1.1.3 Advantages and Limitations

Advantages:

- RNNs have the ability to handle variable-length sequence,
- RNNs can capture temporal dependencies as discussed in Fang at al. (2021),
- RNNs have a relatively simple architecture than other models.

Limitations:

- RNNs suffer from the vanishing gradients or the exploding gradients problem. This makes it difficult to capture long-term dependencies,
- RNNs may struggle with sequences that are too long,
- RNNs are sometimes computationally expensive, especially when we have a very large dataset.

4.1.1.4 Example: Bitcoin Price Prediction

To better understand the RNNs, let's consider a simple model and predict Bitcoin prices with it, Awoke et al. (2020).

Input Features (daily):

- x_1: Normalised closing price
- x_2: Normalised trading volume
- x_3: 7-day moving average
- x_4: 30-day moving average
- x_5: Sentiment scores from social media

The RNN model would process the sequence of these features, let's say for the past 30 days, to make a prediction on the next day's closing price. The model would learn to attribute weightage to new information more heavily when it is still considering the longer-term trends that were captured in the moving averages.

4.1.2 Gated Recurrent Unit (GRU)

4.1.2.1 Theory and Architecture

The Gated Recurrent Unit model is an advanced version of the RNN architecture. It is designed to address the vanishing or exploding gradient problem. GRU model uses an update gate and a reset gate. These gates control the flow of information that allow the network to capture longer dependencies more effectively, Elsayed et al. (2019).

The architecture of the GRU model can be mathematically described as follows:
$z_t = \sigma(W^z x_1 + U^z h_{t-1} + b^z)$ (Represents the Update gate)
$r_t = \sigma(W^r x_t + U^r h_{t-1} + b^r)$ (Represents the Reset gate)
$\tilde{h}_t = tanh(W x_t + U(r_t \odot h_{t-1}) + b)$ (Represents the Candidate activation)
$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$ (Represents the New hidden state)

In these notations:

- σ is the sigmoid function
- $\odot$ denotes the element-wise multiplication
- W, U are the weight matrices and
- b are the bias vectors

The update gate (z_t) determines what proportion of the previous hidden state must be retained. The reset gate (r_t) controls the proportion of the previous state that must be forgotten when computing and calculating activation.

4.1.2.2 Applications to Financial Time Series

The GRU models are particularly useful for making financial time series predictions. GRUs have the ability to capture both short-term and also long-term dependencies in given data. The GRU model can effectively model complex market dynamics that are needed and are commonly seen in the finance world. These dynamics include:

- Trend persistence,
- Cyclic patterns,
- Sudden regime changes,
- Geopolitics,
- Cultural and local factors.

In cryptocurrency, and stock market price predictions, the GRU models can process sequences of technical indicators. They can also process fundamental data with market sentiments to make more accurate forecasts.

4.1.2.3 Advantages and Limitations

Advantages:

- GRUs are better at capturing long-term dependencies compared to the standard RNN model.
- GRUs are more efficient in computation than LSTMs, they have fewer parameters.
- GRUs tackle the vanishing gradient problem better.

Limitations:

- GRUs may struggle with really long sequences.
- GRUs are more challenging to interpret.
- GRUs require careful hyperparameter tuning to get optimal performance.

4.1.2.4 Example: Stock Market Sector Rotation Prediction

Considering that a GRU model is designed to make predictions on the sector rotation in the stock market, as presented in Karatas & Hirsa, (2021). Let's see how we address this.

Input features (weekly):

- x_1**:** Sector performance that is relative to the broader market
- x_2**:** Changes in the interest rate
- x_3**:** Economic indicators (e.g. growth of GDP, rate of unemployment)
- x_4**:** Sector-specific sentiment scores

The GRU model would process the sequences of these features for making a prediction on which sectors are likely to outperform in the coming weeks. The model's gating techniques allow the process to capture not just recent market trends, but also longer-term economic cycles. Thai will provide a more nuanced prediction of sector rotations.

4.1.3 Long Short-Term Memory (LSTM)

4.1.3.1 Theory and Architecture

The Long Short-Term Memory network models are another advanced form of the RNN architecture. They are designed to learn long term dependencies of the sequential data. LSTMs introduce the memory cell along with three new gates: the input gate, the forget gate and the output gate. These gates are there to control the flow of information in an LSTM framework, this is discussed in Siami-Namini et al. (2019).

In mathematical terms, the LSTM architecture is described as follows:

$f_t = \sigma(Wf \cdot [h_{t-1}, x_t] + bf)$ (Represents the forget gate)

$i_t = \sigma(Wi \cdot [h_{t-1}, x_t] + bi)$ (Represents the input gate)

$\tilde{o}_t = tanh(Wo \cdot [h_{t-1}, x_t] + bo)$ (Represents the output gate)

$\tilde{C}_t = tanh(Wc \cdot [h_{t-1}, x_t] + bc)$ (Represents the candidate memory cell)

$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$ (Represents the new memory cell)

$h_t = o_t \odot tanh(C_t)$ (Represents the new hidden state)

In these notations:

- σ is the sigmoid function
- $\odot$ denotes the element-wise multiplication

- $\boldsymbol{W}$ are the weight matrices
- $\boldsymbol{b}$ are bias vectors

In the LSTM, the forget gate ($\boldsymbol{f_t}$) determines what information should be discarded in the cell state and what information must flow through. The information gate ($\boldsymbol{i_t}$) decides the new information that needs to be stored. The output gate ($\boldsymbol{o_t}$) will control the information that goes from the cell state to the output. This is discussed in Yadav et al. (2020).

4.1.3.2 Applications to Financial Time Series

LSTMs in particular are well-suited for financial time series modelling and predictions. Their ability to capture long-term and complex dependencies make them well suited for the task. They can effectively model:

- Long-term market trends
- Seasonal patterns
- The impact of infrequent yet significant events (economic crises, or policy changes)

In the cryptocurrency space, the LSTMs can process extensive historical data that includes price movements, volumes of trade and more external factors too to build a robust forecast.

4.1.3.3 Advantages and Limitations

Advantages:

- LSTMs are excellent when capturing long-term dependencies.
- LSTMs are robust against the vanishing gradient problem.
- LSTMs are flexible to adapt to various tasks in the Time Series.

Limitations:

- LSTMs are a lot more complex and expensive for computations, when compared to the GRUs.
- LSTMs are prone to overfitting, especially when data is limited.
- LSTMs require careful initialization and a proper procedure for training.

4.1.3.4 Example: Cryptocurrency Portfolio Optimization

Let's take an LSTM model that is designed to optimise cryptocurrency portfolio.

Input features (daily, for multiple cryptocurrencies):

- $\boldsymbol{x_1}$**:** Price returns
- $\boldsymbol{x_2}$**:** Trading volume
- $\boldsymbol{x_3}$**:** Market capitalization
- $\boldsymbol{x_4}$**:** Measures of volatility
- $\boldsymbol{x_5}$**:** Cross-correlation with other cryptocurrencies

In this, the LSTM model would process sequences of the features simultaneously, taking into account multiple cryptocurrencies. The LSTM's ability for long-term dependencies allow the model to map complex interactions between different cryptocurrencies and the overall trends of the market. The model can potentially output an optimal portfolio weights set that is set for the next investment period and is able to balance risk with expected returns as seen in Sen et al. (2021).

4.2 Hybrid Models

4.2.1 LSTM-RNN

4.2.1.1 Theory and Architecture

The hybrid LSTM-RNN model combines the long-term memory capabilities of the LSTM model and the simplicity provided by the RNNs. The architecture of this hybrid approach aims to capture the dependencies of a long sequential data and also maps the short-term local patterns. The model can be described mathematically as follows:

Firstly, the input is processed through the LSTM layers and then the output from the LSTM layers is fed into a standard RNN layer:

$$h_t = \sigma(W_{hh}h_{t-1} + W_{xh}x_t' + b_h)$$

$$y_t = \sigma(W_{h\gamma}h_t + b_\gamma)$$

In this, the x_t' term is taken as the output from the LSTM layer. The architecture used in our experimentation is demonstrated in Fig 2.

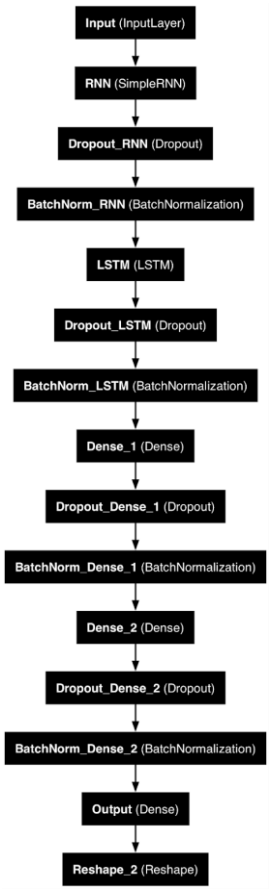

Fig 2: *The LSTM-RNN hybrid architecture for the experiment*

4.2.1.2 Applications to Financial Time Series

The LSTM-RNN hybrid model is a particular model that is useful for financial time series that are exhibiting both long-term relational trends and some short-term fluctuations. In our context of the project, this model can:

- Capture long term cycles of the market and the trends through the LSTM layers,

- Model short term movements in price and their volatility with the RNN layer,
- Combines the fundamental analysis (long-term) along with the technical analysis (short-term).

4.2.1.3 Advantages and Limitations

Advantages:

- LSTM-RNN balances the strengths of both LSTM and RNN architectures,
- LSTM-RNN captures long-term and short-term patterns effectively,
- LSTM-RNN are more interpretable than pure LSTM models.

Limitations:

- LSTM-RNN has increased complexity in the model as compared to the individual ones (LSTM or the RNN),
- LSTM-RNN needs a careful training pipeline with balanced tuning to strike the right balance of performance,
- LSTM-RNN is computationally expensive for really large datasets.

4.2.1.4 Example: Multi-asset Class Market Prediction

Let's take an LSTM-RNN model that is designed to make movement predictions across multiple asset classes:

Input features (daily, for stocks, bonds, commodities and cryptocurrencies):

- x_1**:** Class returns of Asset,
- x_2**:** Macroeconomic indicators (interest rates, inflation),
- x_3**:** Market sentiment scores,
- x_4**:** Cross-asset correlations.

In a hybrid approach,the LSTM layers would process these inputs for modelling long-term trends and relationships between asset classes. The LSTM's output would then be fed into the RNN layer that could focus on short-term market dynamics and even regime shifts. The hybrid approach will provide a full view of market movements across different asset classes. This will account for both long-term cycles of economics and short-term market sentiment.

4.2.2 LSTM-GRU

4.2.2.1 Theory and Architecture

The hybrid model LSTM-GRU combines the strengths of the LSTM and the GRU model architectures. This hybrid model aims to leverage the dependable long-term memory of the LSTMs and the computation efficiency of the GRUs. The model can be denoted mathematically as follows:

The input is first processed through the LSTM layers, and then the output of the LSTM layer is fed into the GRU layers. The architecture used in our experimentation is demonstrated in Fig 3.

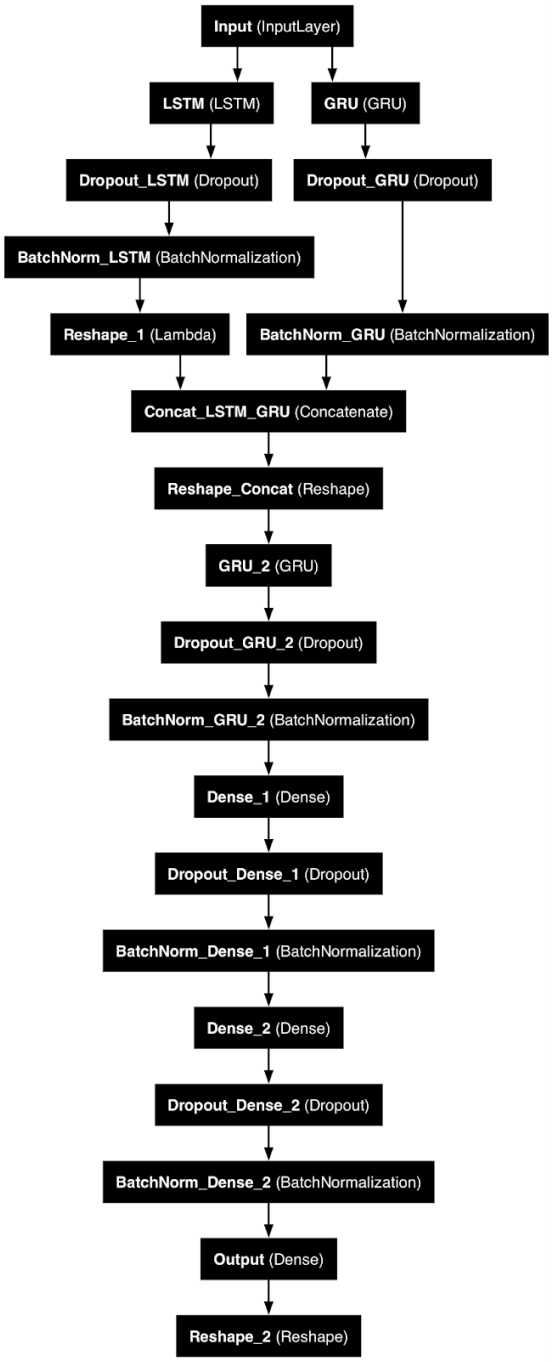

Fig 3: *The LSTM-GRU hybrid architecture for the experiment*

4.2.2.2 Applications to Financial Time Series

The LSTM-GRU hybrid approach is well suited for making complex financial time series analysis which require long-term memory and adaptive short-term processing. In cryptocurrency and stock market prediction use cases, the model can:

- Use LSTM layers for capturing long-term market trends, economic cycles and persistent patterns.
- Employ GRU layers to adapt quickly into the recent markets' information with sudden changes.

4.2.2.3 Advantages and Limitations

Advantages:

- LSTM-GRU combines the strengths of both individual architectures,
- LSTM-GRU is potentially more flexible and adaptive in nature than a single architecture model,
- LSTM-GRU can handle a wide range of dependencies in the temporal space.

Limitations:

- LSTM-GRU has increase model complexity and can lead to potential overfitting,
- LSTM-GRU requires a careful design that determines the optimal arrangement of the LSTM and the GRU layers,
- LSTM-GRU may be computationally expensive, majorly in terms of large-scale applications.

4.2.2.4 Example: High-Frequency Trading Strategy

Let's take an LSTM-GRU model that is designed to conduct high-frequency trading in the markets of cryptocurrency:

Input features (per minute):

- x_1**:** Price tick data,
- x_2**:** Order book depth,
- x_3**:** Volume of trading,
- x_4**:** Short-term technical indicators (1-minute, 5-minute moving averages),
- x_5**:** Market microstructure metrics (bid-ask spread, imbalance of order-flow).

In a hybrid approach, the LSTM layers would process these inputs for modelling long sequences and long-term market dynamics. The LSTM's output would then be fed into the GRU layer that could focus on short-term market dynamics, recent market conditions and the microstructure changes. The hybrid approach will be able to generate a rapid trading signals series that can account for persistent markets and immediate market conditions.

4.2.3 Transformer-LSTM

4.2.3.1 Theory and Architecture

The Transformer-LSTM hybrid approach combines the self attention mechanism of Transformers with the sequential processing capability of the LSTM model. The architecture aims to capture complex and nonlinear relationships in the data to maintain the ability for processing sequential information effectively, also demonstrated in Chen et al. (2022).

The Transformer's Self-Attention mechanism:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k})V$$

Here, the Q, K, and V are the query, key and value matrices that are derived from the input, Vaswani, et al,. (2017).

The Transformer's Multi-head attention mechanism:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$
$$where\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

The Transformer's output is fed into the LSTM layers. The architecture used in our experimentation is demonstrated in Fig 4.

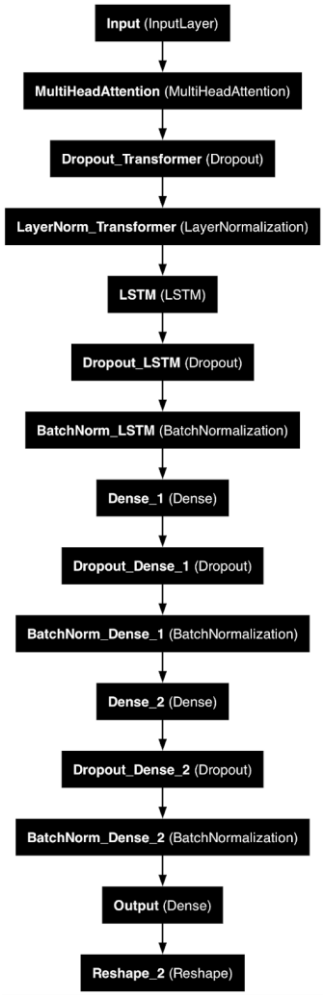

Fig 4: *The Transformer-LSTM hybrid architecture for the experiment*

4.2.3.2 Applications to Financial Time Series

The Transformer-LSTM hybrid model is a powerful approach for exhibiting complex and non-linear relationships with longer dependencies. In our project's use case, this model can:

- Use the Transformer's self-attention technique to capture complex relationships between different market factors.
- Employ LSTM layers for processing the sequential aspects of the data points for maintaining temporal coherence.

4.2.3.3 Advantages and Limitations

Advantages:

- Transformer-LSTM combines pattern recognition and data inter-relatability of the Transformers with sequential processing of the LSTMs,
- Transformer-LSTM can capture long-range dependencies well,
- Transformer-LSTM can be highly flexible and adaptable to apply to many financial prediction tasks.

Limitations:

- Transformer-LSTM will increase model complexity and computational requirements,
- Transformer-LSTM requires large amounts of data for effective training,
- Transformer-LSTM is challenging to interpret due to high complexity,

4.2.3.4 Example: Global Macro Trading Strategy

Let's take a Transformer-LSTM model that is designed to get a global macro trading strategy:

Input features (daily, for multiple countries and asset classes):

- x_1**:** Asset returns (stocks, bonds, currencies, commodities),
- x_2**:** Economic indicators (GDP growth, inflation, unemployment),
- x_3**:** Central bank policy rates,
- x_4**:** Geopolitical risk indices,
- x_5**:** Global trade flow.

In a hybrid approach, the Transformer layers would process the diverse inputs for capturing complex relationships of countries, asset classes and economic factors. The self-attention approach allows the model to focus on relevant factors. The output of the Transformers is fed into the LSTM layers for sequential processing. This approach generates signals accounting for complex global macro relationships while also maintaining a coherent view of the evolving market with time.

4.3 Adversarial Networks

4.3.1 Generative Adversarial Networks (GANs)

4.3.1.1 Theory and Architecture

Generative Adversarial Networks are a class of deep learning models that consist of two neural networks that are competing against each other in a game-theoretic conceptual scenario. While the GANs were originally designed to generate realistic images, they have been used and adapted for several applications that include time series generation and predictions, as presented in Goodfellow et al. (2020). The architectures of the two neural networks, used in our experimentation, are demonstrated in Fig 5.

The two important components of the GAN are:

1. **Generator (G):** This is the network that generates synthetic data samples. In the financial time series context, it would be able to generate synthetic price sequences or some other market data.
2. **Discriminator (D):** This network tries to distinguish between the real samples and the synthetic samples that are produced by the Generator.

The Generator and the Discriminator are expressed as a two-player minmax game.
min_G max_D V(D, G) = E_xp_data(x)[log D(x)] + E_zp_z(z)[log(1 - D(G(z)))]
In this notation:

- **G** is the generator network,
- **D** is the discriminator network,
- **X** represents all real data samples,
- **z** represents random noise input that goes into the generator,
- **p_data(x)** is the real distribution of data,
- **p_z(z)** is the Gaussian noise distribution.

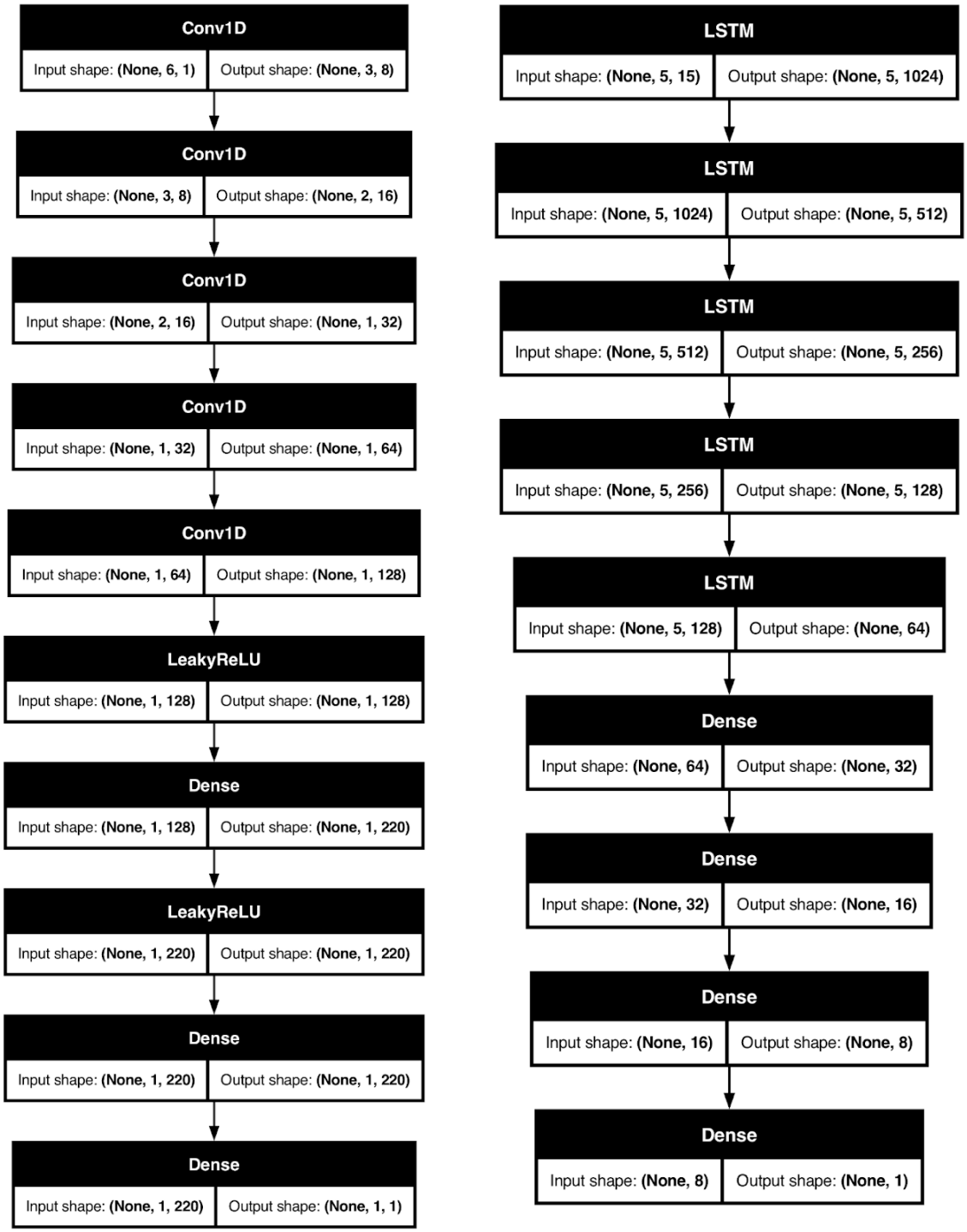

Fig 5. *(a) The Discriminator architecture (b) The Generator architecture used in our experiments.*

4.3.1.2 Applications to Financial Time Series

GANs are a significant approach in our project as they present an innovative approach.

1. GANs can create scenarios and multiple possible trajectories of price for risk assessment and stress testing.

2. GANs can also generate synthetic examples, allowing to augment data for the training set.
3. GANs can detect anomalies by learning the distribution of a normal market behaviour and using that to identify unusual market conditions.
4. GANs can be adapted for achieving tasks of direct prediction through conditioning the generator on past market data.

GANs are powerful tools for generating and training over realistic financial scenarios to capture complex market dynamics. In our project, the GANs approach will play an important role in the financial modelling and projections part to allow for better risk management of such systems.

5. Implementation & Evaluation

5.1 Experimentation Setup and Resources

We developed this entire system and ran these experiments on a standalone MacBook M1 Pro. Thai showcases the capability of how low-resourced and easy to replicate and execute our research is for anyone looking to develop and deploy such systems. Even for the most complex training like those of the GANs and the Transformers, we were able to execute this on a single M1 chip's neural engine.

With the completion of this project, we demonstrate that such similar setups can be replicated on other personal computers. This makes Machine Learning, especially for financial advisory, accessible to small-scale companies, applications or individuals around the world. This research shows that advanced predictive systems can be deployed without accessing extensive computational resources and make this usage practical for a wide range of users.

All sequential Deep Learning approaches, like the RNN, LSTM, GRU and Transformers (along with their hybrid approaches) were trained on 50 epochs. However, the GAN model has been trained over 500 epochs before final testing.

5.2 Evaluation Metrics

5.2.1 Mean Squared Error (MSE)

Mean Squared Error (MSE) is the metric that is able to quantify the average squared difference between the model's predicted values and the actual values. The MSE is calculated by the formula shown:

$$\text{MSE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{Mean}} \underbrace{(Y_i - \hat{Y}_i)}_{\text{Error}}^{\underbrace{2}_{\text{Squared}}}$$

In this equation, $\boldsymbol{n}$ is the number of the observations, while $\boldsymbol{Y_i}$ is the actual value and $\boldsymbol{Y\hat{}_i}$ is the predicted value by the model.

MSE gives us a measure of the variance between actual and predicted values. The errors are squared. This indicates that MSE gives a higher weight to larger errors, hence making it sensitive to outliers. A lower MSE value indicates that the model has a great predictive accuracy due to the fact that the average squared difference between predictions and actual outcomes is very small.

In our project, the MSE was used to evaluate how well the model is able to predict stock prices and cryptocurrency prices accurately. It is a crucial metric that assesses the models' ability to capture patterns in the data with minimised prediction errors.

5.2.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error is very similar to the MSE calculator, it is just the square root of the MSE. It is defined as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

The RMSE calculation indicates the average magnitude of errors that are in the same units as the original data. This makes it more interpretable than the MSE. By just taking the square root of the MSE, the RMSE modulates the influence of large errorsts and provides a balanced view of the model's predictive performance. A lower RMSE means that the model has better accuracy.

We also use the RMSE in some experiments to give a more intuitive understanding of the prediction errors. Given that the RMSE expresses errors in the same units as the predicted variable (stock prices or cryptocurrency prices), it offers a direct interpretation of how the model performs in financial terms. This is beneficial for stakeholders who require clear insights and are demanding precision of predictions.

5.3 Results & Analysis

We will discuss the results of our experiments in this section. Our results were divided into two approaches of experimentation:

1. **Multiple Cryptocurrencies Price Prediction:** Compared the performance metrics across standalone Deep Learning models and Hybrid approaches.
2. **Stock Price Prediction with Sentiment Analysis from Twitter:** Compared the performance of a GAN model approach with the other best performing Hybrid models.

5.3.1 Cryptocurrency Price Prediction

5.3.1.1 Basic Models

The most basic models for cryptocurrency price predictions were trained on a normalised dataset that used the MinMaxScaler for consistency across currency values. Each model was trained on one currency at a time, and evaluated to test over multiple currencies. The models applied were the Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU) and the Long-Short Term Memory (LSTM). This evaluation is conducted and quoted for the WAXPUSDT cryptocurrency.

Model Name	Mean Squared Error (MSE)
Recurrent Neural Network (RNN)	2.4196e-04
Gated Recurrent Unit (GRU)	8.5267e-05
Long-Short Term Memory (LSTM)	1.2132e-04

Table 2: *A summary of results for all three basic models evaluated.*

The GRU model demonstrates its supremacy over the other models as it has the lowest MSE with one one exponent lower in the MSE score than the rest.

We also evaluated the models on all columns, by setting each of the columns as the target columns from the dataset. This evaluation is based on the Mean Absolute Error. These results are summarised in Table 3.

Model Name	MAE_open	MAE_high	MAE_low	MAE_close
LSTM	0.010813042740905	0.007039306449982 1	0.006022837825977 7	0.011650317737664
GRU	0.005193252240762 7	0.005595012202545	0.008756472975519 2	0.010204375062517 9
RNN	0.015153893353321 3	0.008018436158277	0.014642899817310 4	0.010899295022935 8

Table 3: *A summary of results for all three basic models evaluated column-wise.*

The LSTM and GRU's performances are also compared visually through a bar chart as demonstrated in Fig 6. It shows how the GRU model outperforms the LSTM model on most target columns.

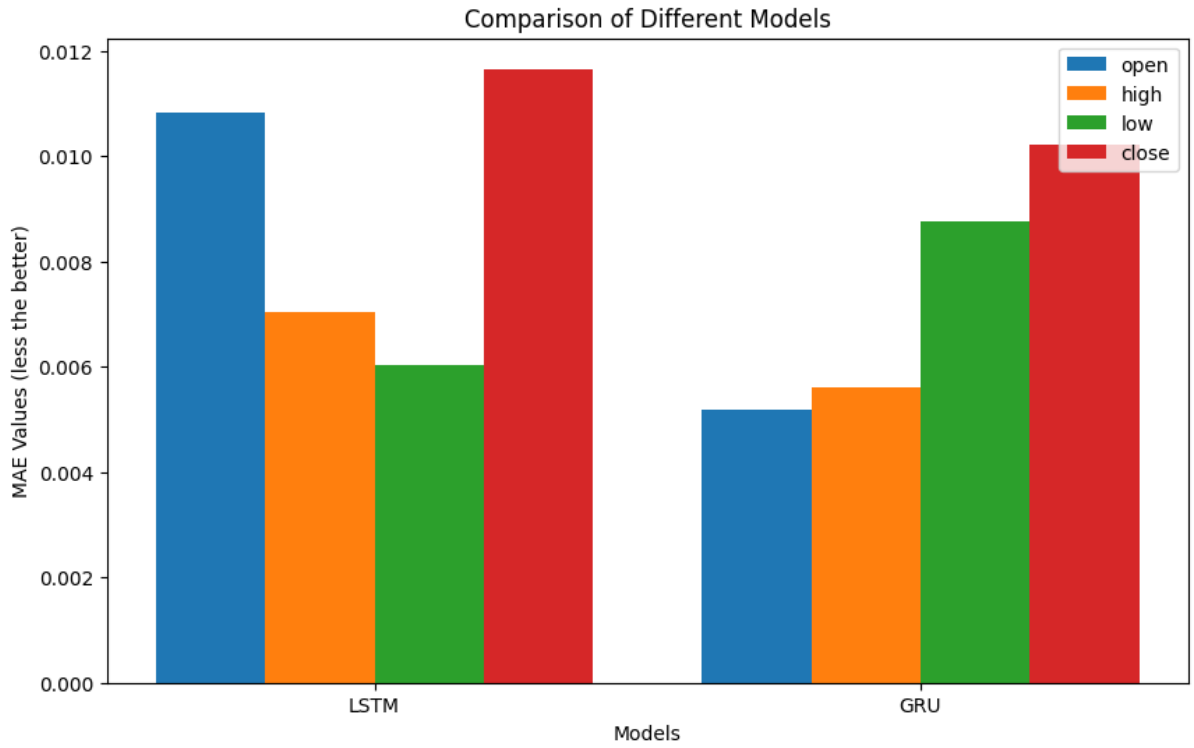

Fig 6: *Bar chart for comparing LSTM with GRU over all columns.*

5.3.1.2 Hybrid Models

The hybrid models for cryptocurrency price predictions were trained under different scenarios where we trained on a dataset combining multiple cryptocurrencies simultaneously, trained on some cryptocurrencies and evaluated some unseen ones and trained on individual trends. Combining multiple currencies in training could not yield a foreseeable pattern, hence our approach of individual cryptocurrency training proceeded further through our experimental phase. With advanced Exploratory Data Analysis and data integration, the approach aimed to leverage different cryptocurrencies' diverse patterns. The tests included model combinations like the LSTM with RNN, GRU and Transformers.

Model Name	Mean Squared Error (MSE)
LSTM-RNN	5.7440e-04
LSTM-GRU	0.0014
Transformer-LSTM	0.0015

Table 4: *A summary of results for all three hybrid models evaluated.*

In the hybrid models, the LSTM-RNN has outperformed all other hybrid models, however, as far as their comparison with the basic model is concerned, they were unable to outperform them. This suggests that a single GRU model can be deemed effective over all of these approaches. Only the LSTM-RNN approach has yielded results that stand comparable to the results to the basic models.

The hybrid models' performances are also compared visually through a bar chart as demonstrated in Fig 7. It shows how each model performs on each target column.

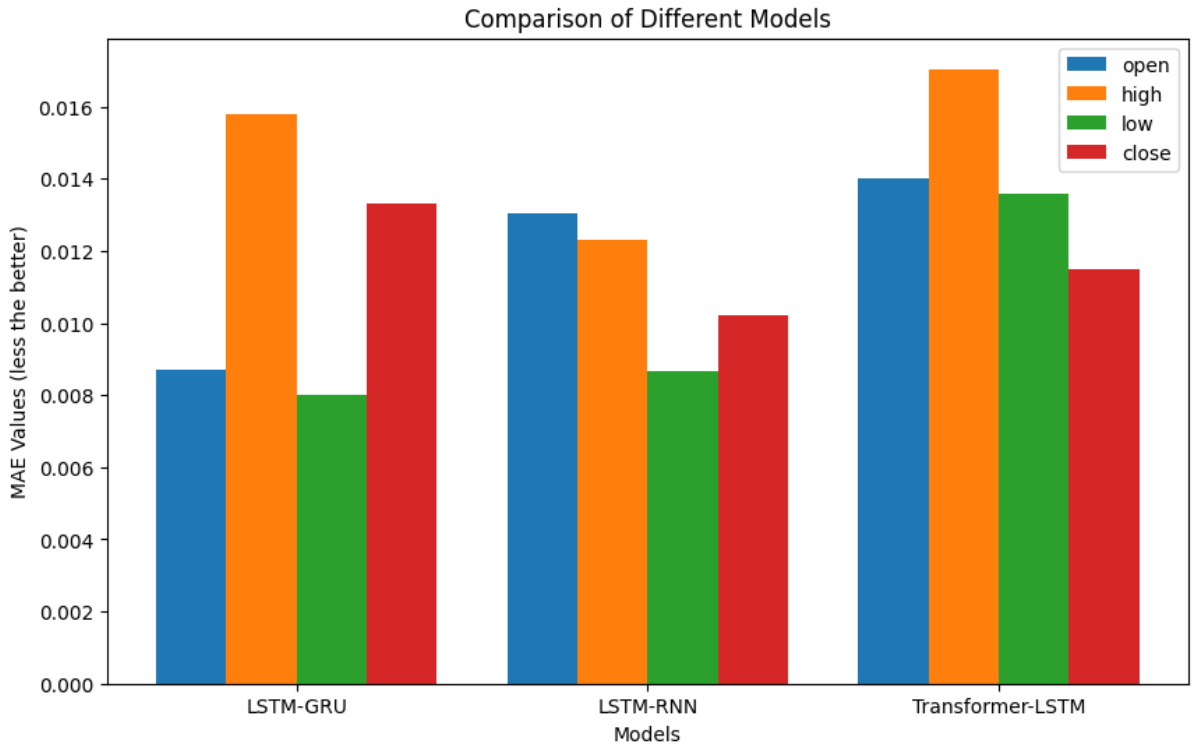

Fig 7: *Bar chart for comparing hybrid models over all columns.*

5.3.2 Stock Price Prediction using Sentiment Analysis

In this experiment, we predict stock prices' (closing values) and incorporate them with sentiment analysis from the tweets on Twitter (now X.com). We assigned each tweet with a sentiment score, and added that score as a feature alongside market trend values. The integration of the market data (OHLCV) with the sentiments creates a market sentiment aware forecasting plan.

Unlike our previous tests, this test is conducted over a Stock Market dataset instead of a Cryptocurrency dataset, as we lacked the availability of Twitter sentiments revolving around Cryptocurrencies. Even if the sentiments existed, their legitimacy was to be questioned, and could be unethical, as most cryptocurrency Twitter sentiments are affected through tweets made on hacked accounts, and do not account for the actual public sentiment.

Our comparative evaluation uses the same models as tested above with Cryptocurrency datasets to establish a relatability factor that cryptocurrency data and stock market data addresses the same nature, and builds over the same concept. The comparison is created on the basis of first evaluating scores over only the financial data, of time series, where we do not add sentiments. The next experiment in this evaluation is on the basis of adding sentiment scores to the stock data.

For this task we evaluated the results on a Generative Adversarial Network (GAN), LSTM-RNN and a LSTM-GRU hybrid technique. We also evaluated our best performing model from the cryptocurrencies experiments, the GRU model. Table 5 shows the results without sentiments, and Table 6 shows results with sentiments

Names	Root Mean Squared Error (RMSE)	Mean Squared Error (MSE)
Generative Adversarial Network (GAN)	3.4587	11.96
LSTM-RNN	0.6351	0.4033
LSTM-GRU	0.6003	0.3604
GRU	0.0509	0.0026

Table 5: *A summary of results for all models over the stock dataset evaluated before incorporating sentiments.*

Names	Root Mean Squared Error (RMSE)	Mean Squared Error (MSE)
Generative Adversarial Network (GAN)	2.6279	5.194
LSTM-RNN	0.1374	0.0189
LSTM-GRU	0.1328	0.0176
GRU	0.0328	0.0011

Table 6: *A summary of results for all models over the stock dataset evaluated before incorporating sentiments.*

The inclusion of sentiment scores from Twitter shows some promising improvements in the model's prediction of stock prices. The LSTM-GRU model demonstrates a lower RMSE and MSE value as compared to other Hybrid and GAN approaches, but it is the GRU model that has a significantly lower score than the rest of the models in the experiment.

The GAN model took the longest training, however originally used for generating synthetic but realistic images, it fails to outperform the specialised sequential models for this task. Its training process has been demonstrated for both: Generator and Discriminator, over the 500 epochs in Fig 8.

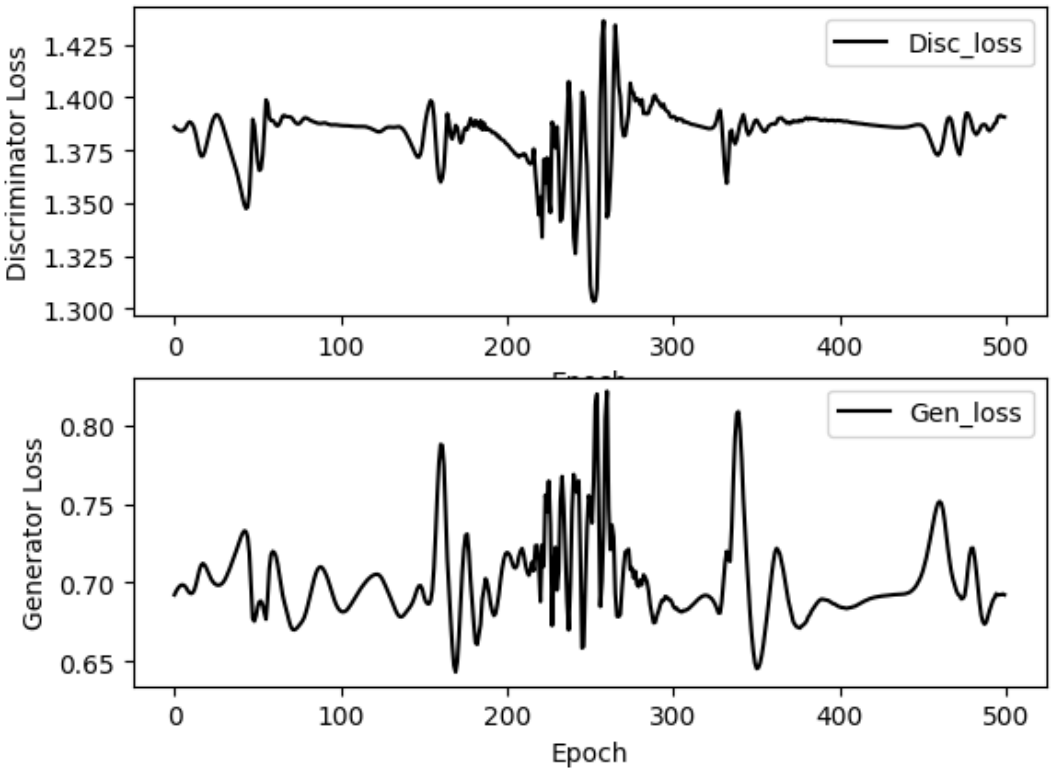

Fig 8: *The training loss, epoch-wise for the (a) Discriminator and the (b) Generator.*

The GAN model can be seen to perform well with Time Series and Stock Data, if we see Fig 9. However, it gives very poor predictions when using raw historical data. Through the use of technical indicators and Twitter sentiment analysis, it improves in performance but still does not reach the capacity of being as accurate as the sequential Deep Learning approaches. GANs do not work well for the "less popular" stock tickers as their number of tweets would be lower than some of the most famous stocks like Amazon or Tesla. In cases like these, the sentiment scores may fail to capture the bigger picture and lead to reduced or worse results of the model as we saw in the reported results.

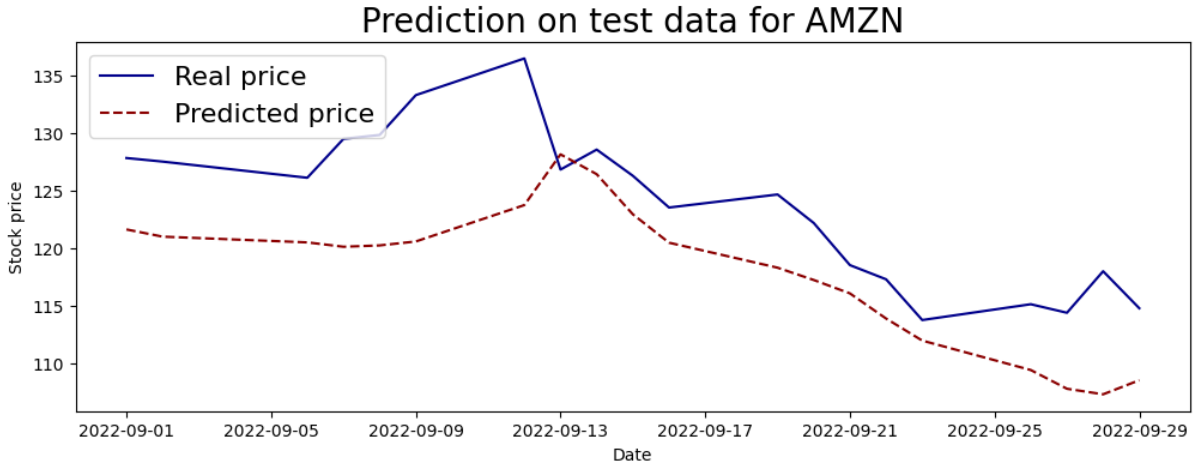

Fig 9: *The real price of the stock plotted against the predicted price by the GAN model*

We also demonstrate the results of the hybrid models, LSTM-RNN and LSTM-GRU, that are evaluated on each of the four target columns that we can have in stocks, in order to determine

their price and analyse their trends. We can see that the LSTM-GRU is a much closer approximator of the actual values and has a lot of overlapping points in the plot, indicating its efficiency as an approach for making Stock Market predictions using sentiments of Twitter users from their tweets.

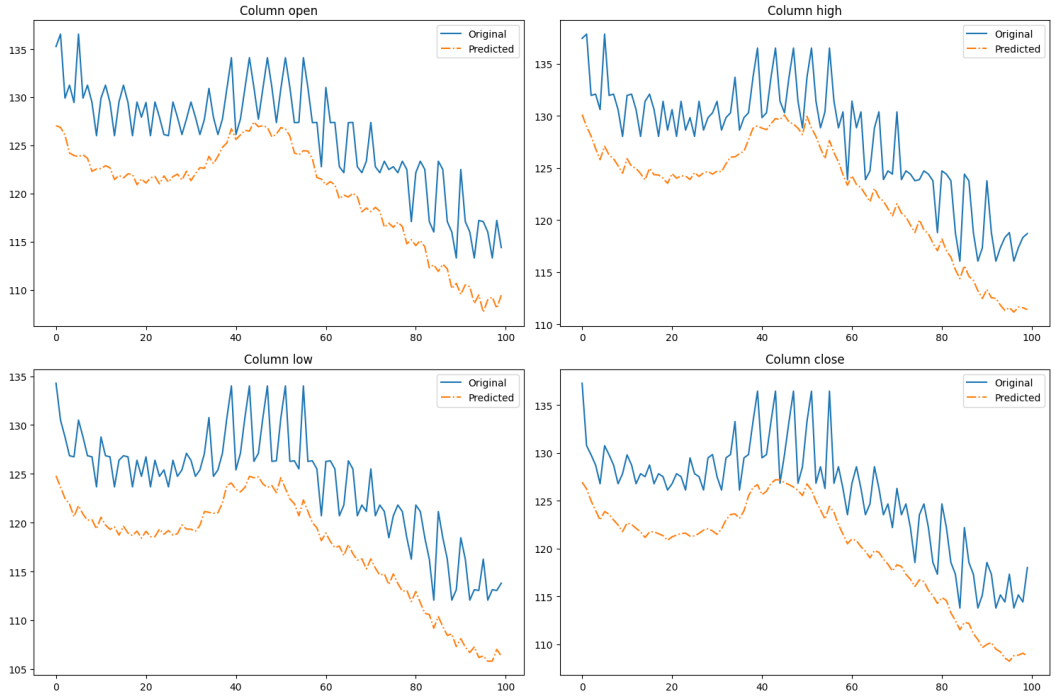

Fig 10: *The LSTM-RNN evaluated over the four main target columns after incorporating sentiment scores*

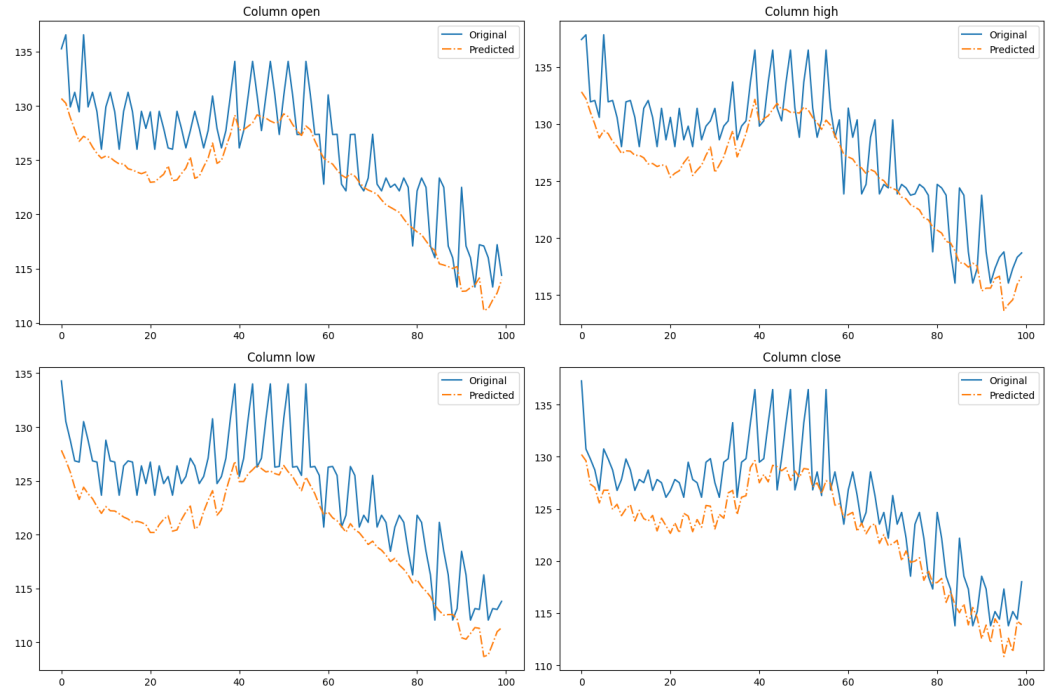

Fig 11: *The LSTM-GRU evaluated over the four main target columns after incorporating sentiment scores*

5.3.3 Ablation Studies

To understand how each of the components of our experiment impact performance, the ablation studies were conducted. We initially identified that using just the base features and no data normalisation lead to terrible results. Thus, we indulged in the process of Feature Engineering as discussed in Chapter 3.

For cryptocurrencies, when all cryptocurrencies were trained simultaneously, the models yielded much lower performance as the model was unable to learn and focus on one primary cryptocurrency. In the finance world, each currency (or entity) has their own trends, patterns and market sentiments that affect their value. There is very little commonality between trends across various cryptocurrencies.

If we exclude sentiment analysis from the stock price prediction models, the models give us higher RMSE and MSE values. This indicates that sentiment scores from Twitter and the trends that we follow on social media create a considerable and effective impact and contribute towards the accuracy understanding of stock price predictions.

5.4 Discussion

Our experiments demonstrate the effectiveness of different neural network architectures for predicting cryptocurrency values and stock prices (closing) with varying trends, patterns. For predicting cryptocurrency, the basic models like the GRU performed the best, and even outperformed all hybrid models. While hybrid models aim to be more nuanced in nature, their over complexity, that is introduced through the combination of multiple architectures, is not always needed. The model may get too complex that it cannot capture underlying patterns in an effective manner.

As an alternative, or contrasting use case, the stock price prediction that incorporates the sentiment analysis from Twitter tweets has shown that it can be of benefit. The LSTM-GRU model has the lowest MSE and RMSE among all experiments of advanced models, except the GRU standalone model that has a significant performance boost in the cryptocurrencies data, as well as the stocks' data (with sentiment, and without sentiment).

This highlights the potential of leveraging and integrating real-time trends from social media to enhance market predictions. Although, in our project's scope, the concept and implementation focused on the stocks being predicted with sentiments, the unavailability of cryptocurrency sentiment data was not a problem. The approach remains to be the same, and we have successfully built a proof of concept that is applicable to all time series datasets that have a similar nature to the problem, as stocks and cryptocurrency price movements do.

6. Conclusion and Future Work

This research presented in our dissertation provides an exploratory approach towards testing multiple basic, and advanced Deep Learning approaches to predict prices of cryptocurrencies and stocks. Our experiments demonstrated how the models like the GRU, RNN and LSTM are effective in their capacities and how hybrid models are also a good approach for this use case.

Out of our many key findings, we were clear to identify that not all problems require a complex, hybrid and superior solution. Among the basic models, the GRU model had the lowest MSE and it indicates a superior performance over all. None of the hybrid models outperform the single models in the specific context. Apart from this, by incorporating sentiment analysis from

Twitter, we have seen a significant performance increase in the prediction loss. The LSTM-GRU model has shown that it is a valuable approach to combining market data with social media sentiment, and it yields the lowest loss value of the entire experiment showing how adding sentiments truly impacts how we evaluate the stock prices and their changing trends.

Addressing our initial research questions, we established the success of multiple sequential neural network architectures, with the GRU model being on top for Cryptocurrencies, and Stock Market data. Although not at the top of the performance metrics, the hybrid approaches LSTM-GRU and LSTM-RNN also provide an interesting tradeoff in performance and selection. While LSTM-RNN outperforms in the cryptocurrency space, the LSTM-GRU is a better estimator for stock prices data, showcasing their interchangeable usage and calling for a careful curation of even more sophisticated hybrid approaches made to perfection, with GRU components embedded as needed. We also see how incorporating the sentiments from social media has significantly improved the model performances. Simply looking at the GRU, by embedding sentiments we noticed a drop from 0.0026 MSE to 0.0011 MSE score. This is a 57.6% increase in the model's performance.

Our results, overall, confirmed the effective use of the neural networks for financial market predictions, and how market sentiments can enhance the performance of such tools.

6.1 Limitations

Despite showing a promising result and experiment, our study had several challenges like:

- **Data Availability:** The unavailability of a comprehensive cryptocurrency sentiment data had limited our ability to embed sentiments with cryptocurrency details for value predictions. Having access to public sentiments about cryptocurrencies with timesteps would have helped us get a more holistic view.
- **Model Complexity:** The hybrid models were known to be theoretically advantageous but they introduced more complexity to the experimentation and training without always translating to a better performance. This complexity had an uneven tradeoff of lower performance and higher compute times taken.
- **Sentiment Analysis Accuracy:** The accuracy of the sentiment analysis is also contingent to the quality of the NLP techniques that we used. The VADER sentiment analysis classifier and scoring mechanism has a syntactic approach that, if misinterpreted, can potentially lead to skewing of results.

6.2 Future Work

There are several new directions that one can potentially work towards, this can be added to future work and build over our findings.

- **Expanding Sentiment Analysis:** The potential future researchers may implement sentiment analysis specifically for cryptocurrency sentiments and source more comprehensive social media data that can provide a more detailed and deeper insights and improve the prediction accuracy.
- **Enhanced Hybrid Models:** The researchers may next build over these findings to create more sophisticated hybrid models that are optimised in their configurations to yield better performance. Techniques like model ensembling and transfer learning can be of benefit for building highly accurate systems.
- **Real-Time Predictions:** Development of real-time systems that continuously train and update the model at set time stamps, can be done by adding daily's market and sentiment information making the model's learnings most up-to-date and the most responsive to market changes in real-time.

- **Broader Data Sources:** Not just the tweets' sentiments, the future studies or production systems may be able to incorporate additional data sources like the news articles, economic indicators, opinions of economic institutions and experts, and can enrich the overall feature set that improves model's accuracy.

References

Akyildirim, E., Goncu, A. and Sensoy, A., 2021. Prediction of cryptocurrency returns using machine learning. Annals of Operations Research, 297, pp.3-36.

Alessandretti, L., ElBahrawy, A., Aiello, L.M. and Baronchelli, A., 2018. Anticipating cryptocurrency prices using machine learning. Complexity, 2018(1), p.8983590.

Althelaya, K.A., El-Alfy, E.S.M. and Mohammed, S., 2018, April. Stock market forecast using multivariate analysis with bidirectional and stacked (LSTM, GRU). In 2018 21st Saudi Computer Society National Computer Conference (NCC) (pp. 1-7). IEEE.

Awoke, T., Rout, M., Mohanty, L. and Satapathy, S.C., 2020. Bitcoin price prediction and analysis using deep learning models. In Communication Software and Networks: Proceedings of INDIA 2019 (pp. 631-640). Singapore: Springer Singapore.

Bansal, M., Goyal, A. and Choudhary, A., 2022. Stock market prediction with high accuracy using machine learning techniques. Procedia Computer Science, 215, pp.247-265.
Cavalli, S. and Amoretti, M., 2021. CNN-based multivariate data analysis for bitcoin trend prediction. Applied Soft Computing, 101, p.107065.

Chen, Y., Wang, Y., Liu, X. and Huang, J., 2022, May. Short-term load forecasting for industrial users based on Transformer-LSTM hybrid model. In 2022 IEEE 5th International Electrical and Energy Conference (CIEEC) (pp. 2470-2475). IEEE.

Chowdhury, R., Rahman, M.A., Rahman, M.S. and Mahdy, M.R.C., 2020. An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. Physica A: Statistical Mechanics and its Applications, 551, p.124569.

Derbentsev, V., Babenko, V., Khrustalev, K.I.R.I.L.L., Obruch, H. and Khrustalova, S.O.F.I.I.A., 2021. Comparative performance of machine learning ensemble algorithms for forecasting cryptocurrency prices. International Journal of Engineering, 34(1), pp.140-148.

Elbagir, S. and Yang, J., 2019, March. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, No. 16). Sn.

Elsayed, N., Maida, A.S. and Bayoumi, M., 2019, July. Gated recurrent neural networks empirical utilization for time series classification. In 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 1207-1210). IEEE.

Fang, W., Chen, Y. and Xue, Q., 2021. Survey on research of RNN-based spatio-temporal sequence prediction algorithms. Journal on Big Data, 3(3), p.97.

Garg, A., Shah, T., Jain, V.K. and Sharma, R., 2021, December. Cryptop12: A dataset for cryptocurrency price movement prediction from tweets and historical prices. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 379-384). IEEE.

Gao, Y., Wang, R. and Zhou, E., 2021. Stock prediction based on optimized LSTM and GRU models. Scientific Programming, 2021(1), p.4055281.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM, 63(11), pp.139-144.

Gupta, A. and Nain, H., 2021. Bitcoin price prediction using time series analysis and machine learning techniques. In Machine Learning for Predictive Analysis: Proceedings of ICTIS 2020 (pp. 551-560). Springer Singapore.

Hamayel, M.J. and Owda, A.Y., 2021. A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms. Ai, 2(4), pp.477-496.

Jaquart, P., Köpke, S. and Weinhardt, C., 2022. Machine learning for cryptocurrency market prediction and trading. The Journal of Finance and Data Science, 8, pp.331-352.

Karatas, T. and Hirsa, A., 2021. Two-stage sector rotation methodology using machine learning and deep learning techniques. arXiv preprint arXiv:2108.02838.

Khedr, A.M., Arif, I., El-Bannany, M., Alhashmi, S.M. and Sreedharan, M., 2021. Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. Intelligent Systems in Accounting, Finance and Management, 28(1), pp.3-34.

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. Journal of International Financial Markets, Institutions and Money, 65, 101188.

Li, Y. and Dai, W., 2020. Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model. The journal of engineering, 2020(13), pp.344-347.

Mann, A.D., 2022. Machine Learning Methods to Exploit the Predictive Power of Open, High, Low, Close (OHLC) Data (Doctoral dissertation, UCL (University College London)).

Mittal, A. and Goel, A., 2012. Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15, p.2352.

Mounika, S., Yadav, P.A., Yashaswi, T., Krishna, C.Y. and Reddy, V.K., 2021. Crypto-currency Price prediction using CNN and LSTM models. International Journal for Research in Applied Science and Engineering Technology, 9(3), pp.107-114.

Parekh, R., Patel, N.P., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., Davidson, I.E. and Sharma, R., 2022. DL-GuesS: Deep learning and sentiment analysis-based cryptocurrency price prediction. IEEE Access, 10, pp.35398-35409.

Sen, J., Dutta, A. and Mehtab, S., 2021, October. Stock portfolio optimization using a deep learning LSTM model. In 2021 IEEE Mysore sub section international conference (MysuruCon) (pp. 263-271). IEEE.

Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, p.132306.

Siami-Namini, S., Tavakoli, N. and Namin, A.S., 2019, December. The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International conference on big data (Big Data) (pp. 3285-3292). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Yadav, A., Jha, C.K. and Sharan, A., 2020. Optimizing LSTM for time series prediction in Indian stock market. Procedia Computer Science, 167, pp.2091-2100.

Zhang, J. and Man, K.F., 1998, October. Time series prediction using RNN in multi-dimension embedding phase space. In SMC'98 conference proceedings. 1998 IEEE international conference on systems, man, and cybernetics (cat. no. 98CH36218) (Vol. 2, pp. 1868-1873). IEEE.