DEPARTMENT OF ENGINEERING MATHEMATICS

Public Sentiment Analysis-Based Cryptocurrency Action Recommendation Model Optimization

Yujie Wang

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Saturday 9th September, 2023

Supervisor: Dr. Simon Lock

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Yujie Wang, Saturday 9$^{\text{th}}$ September, 2023

Contents

List of Figures

List of Tables

Ethics Statement

An ethics application for this project was reviewed and approved by the faculty research ethics committee as application 15701.

Abstract

With the surge of interest in cryptocurrencies, the challenge of predicting price fluctuations within the financial markets has come to the forefront. However, the inherent instability of cryptocurrencies makes them particularly difficult to predict with precision. In recent times, a growing awareness has emerged that the pricing patterns of cryptocurrencies can be influenced by the feelings and opinions expressed by the public on various social media platforms. Drawing inspiration from this insight, this study embarks on exploring the fundamental link between public sentiment across a range of social media platforms and the price movements of the leading cryptocurrency, Bitcoin. Moreover, it endeavours to uncover the delayed impact of public sentiment on these price fluctuations. The study further aims to construct proficient models for projecting price returns, predicting price directions, and offering investor action recommendations. This effort utilizes the forecasting capability of public sentiment as a crucial aspect in predicting the price of Bitcoin, all with the aim of maximizing potential profits.

In this research, the study collected hourly Twitter tweets, Google Trends data, and Reddit posts from publicly accessible datasets, focusing on the first half of 2022. At the same time, a comprehensive collection of attributes was assembled using techniques from time-series forecasting and sentiment analysis, with the Long Short-Term Memory (LSTM) model being the primary tool for prediction. The study's findings underscore the tangible impact of public sentiment on cryptocurrency prices, confirming its effectiveness as a reliable predictor. Notably, diverse social media sources yield inconsistent effects on cryptocurrency dynamics. The findings unveil the lagged influence of Google Trends on Bitcoin price returns and price movements, along with the immediate impact of sentiment expressed in tweets on short-term price trends. By utilizing social media data for feature construction, the predictive accuracy of Bitcoin's price movement experiences a noteworthy enhancement. Specifically, the impact of tweets sentiment brings about a substantial 3% to 8% improvement in short-term price movement prediction accuracy. Similarly, it's important to highlight the consistent effectiveness of tweets sentiment scores in increasing the average accuracy by 0.6% over the long term, especially when combined with Google Trends as part of the feature set.

Lastly, an action recommendation model rooted in social media sentiment is constructed. This model empowers investors by furnishing them with actionable recommendations grounded in historical prices, including options like "Buy", "Sell", and "Wait". By fine-tuning the model thresholds, the study highlights the possibility of achieving a notable maximum profits of 21.79% within just one month, and 27.03% within 6 months.

To sum up, the main contributions of the study are as follows:

- Investigation into the connection between Bitcoin price movements and social media data, analysing the delayed impact and causal relationships of diverse public sentiments and interests on Bitcoin price movement direction.
- Emphasis on addressing two distinct prediction challenges regarding Bitcoin prices: regression-based forecasts for price returns and categorical forecasts for price movement direction.
- Concentration on the evaluation and comparison of various models' predictive performance for Bitcoin prices.
- Assessment of the effectiveness of features derived from different sources in predicting Bitcoin prices across both short and long-term trends.
- A trading action recommendation model was devised, utilizing the optimized framework to simulate both short-term and long-term Bitcoin trading. The optimized model furnished actionable advice for future actions, effectively leading to an increase in profits.

Supporting Technologies

- I used Python 3 as the fundamental programming language, along with relevant data science libraries such as Pandas and Numpy.
- I used the Plotly, Matplotlib, and Seaborn libraries, which are public-domain Python libraries, to generate visual charts.
- I used the Scikit-Learn library for machine learning models and metric calculations.
- The early investigative work of the project was guided by the ideas provided by Critien et al. [13], with the code for LSTM model selection: https://github.com/jacquesvcritien/fyp
- I used TensorFlow and Keras libraries for deep learning model building.
- I used NLTK Python library for text cleaning.
- VADER, TextBlob, and Pysentiment2 libraries are used for sentiment analysis.
- I used the open-source visualization software draw.io for creating the diagrams.
- I used Pytrends API to collect the search volume index from Google Trends.
- I used the *Bitcoin_historical_data* API to collect the historical Bitcoin price from Binance.
- I used the public datasets from Kaggle, which include:
  - English Bitcoin Tweets
  - Cryptocurrency-related Reddit Posts
- I used GitHub for version control and code storage.
  - Web address: Project code
- I used OneDrive for storing the video for the oral presentation of my dissertation.
  - Web address: Video
- I used LaTeX to format my thesis via the online service *Overleaf*.

Notation and Acronyms

NLP	:	Natural language processing
CRISP-DM	:	Cross-Industry Standard Process for Data Mining.
SVI	:	Search Volume Index
API	:	Application Programming Interface
VADER	:	Valence Aware Dictionary and sEntiment Reasone
ADF	:	Augmented Dickey-Fuller
ARIAM	:	Autoregressive Integrated Moving Average
AR	:	Autoregressive
MA	:	Moving Average
ARMA	:	Autoregressive Moving Average
ACF	:	Autocorrelation
PACF	:	Partial Autocorrelation
AIC	:	Akaike Information Criterion
CNN	:	Convolutional Neural Network
LSTM	:	Long Short-Term Memory
RNN	:	Recurrent Neural Network
CNN-LSTM	:	Convolutional Neural Network - Long Short-Term Memory
GRU	:	Gated Recurrent Unit
RMSE	:	Root Mean Square Error
MAPE	:	Mean Absolute Percentage Error
ReLU	:	Rectified Linear Unit

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Dr. Simon Lock, who has been the most significant source of support throughout this entire project. His patience, encouragement, and guidance have been invaluable. Whenever the project faced challenges, he offered sincere assistance and concise solutions. Without his help, my journey would have been much more challenging.

I would also like to extend my thanks to the staff at the university. Their patience and assistance for students, along with the services they provide, have been immensely helpful and convenient.

Lastly, I want to express my appreciation to my family for their unwavering support.

Chapter 1

Introduction

This chapter starts by outlining the project's research motivations and primary objectives. It elucidates the reasons behind the widespread interest in investigating the correlation between social media data and Bitcoin's price. Subsequently, the chapter delineates the difficulties associated with examining the forecasting of Bitcoin price fluctuations and employing social media sentiment analysis to investigate its influence on Bitcoin's price. Finally, it provides an overview of the critical matters under consideration in this research, alongside specifying the distinct objectives of the project.

1.1 Motivations

The rise in Internet accessibility has given rise to alternative currencies diverging from traditional monetary systems. Cryptocurrencies, enabled through a distinct method called "mining," have significantly transformed users' economic activities online [32]. Cryptocurrency is a form of electronic cash based on cryptography, allowing users to exchange goods and services similar to traditional currencies [3], which are unique from previous currencies because they operate through a decentralized control system. This system is implemented digitally through a distributed ledger, rather than relying on a centralized banking system [56]. In recent years, cryptocurrencies have gained popularity as an investment option due to security concerns surrounding traditional fiat currencies and strict monetary policies and banking regulations [53]. Since Bitcoin's introduction in 2008, thousands of cryptocurrencies have emerged, sparking significant interest and development in 2017 and early 2018. The increase in the value of Bitcoin and other cryptocurrencies has sparked more interest and involvement.

Given the unique characteristics of cryptocurrency markets, including high volatility, continuous trading, limited capitalization, and abundant market data, predicting prices becomes highly challenging [71]. Cryptocurrency price prediction and its influencing factors have gained substantial interest. Unlike conventional banking systems, Bitcoin's value is driven by public perception, giving it unique attributes. [75]. Behavioural economists clearly showed that decisions in the financial system are influenced by emotional morality and not just by the value of capital [75]. Dolan and Edlin also stated that decisions are influenced by emotions [15]. Thus, sentiment analysis helps to show that commodity prices are effected by values such as sentiment and economic fundamentals [54].

Social media amplifies cryptocurrency market volatility, driven by the challenge of verifying news accuracy. Twitter, in particular, has become a key source for cryptocurrency investors [34]. Regular cryptocurrency enthusiasts share ideas across various platforms, extending beyond Twitter. Kristoufex initially correlated daily Wikipedia traffic with weekly Google search volumes [35]. Later studies highlighted Bitcoin's susceptibility to positive user comments [32].

Building on prior research and considering the current speculative nature of cryptocurrencies, as well as the evident connection between the cryptocurrency market and social media, this study focuses on Bitcoin, the largest and most widely recognized cryptocurrency. The aim is to delve into how public sentiment, gauged through various social media platforms, relates to the fluctuations in Bitcoin's price. This exploration will be undertaken using a variety of modelling approaches. The main goal is to predict cryptocurrency prices based on the analysis of public sentiment. Additionally, the study aspires to devise models that can offer guidance to investors seeking to optimize their gains, suggesting actions like "Sell", "Buy", or "Wait".

1.2 Challenges

Exploring the impact of public sentiment from diverse social media platforms on Bitcoin's price presents several challenges. Sourcing and ensuring data quality is a major obstacle. Gathering substantial data across platforms can be tough, and data integrity might suffer due to misinformation or bot-generated content, adding noise. Text cleaning and sentiment assessment are also complex; words often need context, and language ambiguity can yield inaccurate sentiment analysis.

Understanding lag effects is another challenge. It's crucial to realize that sentiment or price changes might not align instantly; sentiment shifts could follow or precede price shifts. This temporal lag must be considered when building models. Model selection is also a task. Time series problems require data preprocessing, like stabilizing unstable data before feeding it to models. Models need to balance generalizability and the risk of overfitting.

1.3 Objectives

More specifically, the research questions to be explored are as follows:

1. **Is there a relationship between public sentiment on social media and the price movements of Bitcoin that actually exist? If it does, how does public sentiment affect Bitcoin price movements?**

   This study will collect Bitcoin-related posts from various platforms including Twitter, Google Trends, and Reddit. Concurrently, it will gather Bitcoin transaction data for the same time period. Statistical methods will be employed to investigate the potential correlation between public sentiment and cryptocurrency price fluctuations.

2. **Can social media public sentiment analysis be used as a valid indicator for Bitcoin price movement prediction?**

   Assuming that public sentiment on social media does have an impact on Bitcoin price movements, this study will also determine which social media platforms' public sentiment can be used as effective predictive indicators for Bitcoin price changes prediction, and to what extent it can be used to predict cryptocurrency price changes. The selection of appropriate features can maximize the prediction accuracy.

3. **How to use the public sentiment of social media to build a recommendation model of investor behaviour to maximize profits?**

   Finally, this study intends to construct a recommendation model for investor behaviour, use social media sentiment analysis to improve the performance of the action recommendation model, and suggest investors take appropriate actions to maximize profits, such as "Sell", "Buy" and "Wait".

The objectives of this study addressing the proposed questions are as follows:

- Collect Bitcoin-related information from multiple social media platforms, exploring the relationship between public sentiment and Bitcoin price fluctuations and the time lagged impact using statistical techniques and sentiment analysis.
- Select the optimal model to predict price fluctuations and price returns using deep learning models.
- Select related public sentiment as features to optimize model performance and enhance prediction accuracy.
- Construct an action recommendation model using the optimal prediction model, adjusting thresholds to maximize profits within specific time intervals.

1.4 Structure

The rest of this study is organized as follows: Chapter 2 will present the broader context of the cryptocurrency market and offer a comprehensive review of pertinent research. Chapter 3 will focus on the research approaches and explain the workflow of analysis, while Chapter 4 will delve into experiment specifics and the subsequent result discussion. Chapter 5 will consolidate the findings and propose potential future directions.

Chapter 2

Background

This chapter aims to provide essential background information and present a comprehensive review of the current advancements in cryptocurrency price analysis and prediction. Emphasis is placed on the previous studies documented in the literature, particularly focusing on the utilization of social media for predicting cryptocurrency prices.

2.1 Cryptocurrency market

In 2008, the new decentralized cryptocurrency cash system was released. At the same time, it was launched in the form of a cryptocurrency called Bitcoin, the most common application of blockchain technology [46]. In the years following the birth of Bitcoin, many other cryptocurrencies were developed, such as Litecoin and Ethereum. Many alternative coins, known as cottage coins, were created to address Bitcoin's limitations, such as availability and energy consumption [34]. Initially, cryptocurrency literature focused on Bitcoin's security, ethical, and legal aspects. However, in 2017, there was a significant shift in attitude, leading to a surge of interest and a bull market. The quantity of listed cryptocurrencies experienced an exponential growth, surging to 1,865 within one year, marking a more than triple increase. Simultaneously, the total market capitalization underwent a remarkable ascent from $17 billion to a staggering $813 billion [34].

There is ongoing debate about how cryptocurrencies should be classified. Selgin suggested that investors use Bitcoin both as a currency and for investment purposes [66], while Cheah and Fry argued that its volatility contradicts it being a reliable unit or store of value [62]. Yenmack pointed to Bitcoin's scarcity and instability as reasons why it isn't considered a real currency [77]. The SEC has classified Bitcoin and Ethereum as commodities rather than securities [73]. Dwyer's research showed that Bitcoin's volatility exceeds that of gold and major currencies [62], while Dyhrberg found that Bitcoin can be used for risk management similar to gold and the U.S. dollar [16]. Given the unique characteristics of each cryptocurrency, it becomes challenging to generalize findings and assign them to a specific asset class. Some argued that cryptocurrencies may even form an entirely new category [34].

Researchers have raised doubts about the predictability of cryptocurrency prices and whether the efficient market hypothesis (EMH) applies to the cryptocurrency market. According to EMH, the current price should reflect all past information, with future events being the only influencing factor [18]. However, due to the inherent uncertainty of the future, prices tend to follow a random pattern [4]. In terms of weak efficiency, past price changes do not reliably predict future returns [41]. Numerous studies have examined the pricing efficiency of Bitcoin, and early evidence by Urquhart suggested that the Bitcoin market is not weakly efficient and may even become less efficient over time [70].

In the Bitcoin market, prices exhibit complex, chaotic dynamics and, in periods of high prices, significantly increased uncertainty in returns. Markets are complex due to a variety of factors such as economic and political conditions and human behaviour, all of which interact with each other to contribute to uncertainty. Although it can be challenging to forecast market price movements consistently, it is still achievable [71]. Previous studies have also shown that in financial forecasting, earnings do not necessarily have to accurately predict the exact value of future prices. In fact, predicting the market direction can lead to higher profits relative to the exact value of the price [11].

2.2 Google Trends Analytics

This study incorporates data from Google Trends, a platform introduced by Google in May 2006 that showcases the frequency of specific search terms being queried within the Google search engine [1]. Google Trends delivers information in the form of relative search volume scores, representing the popularity of a given search term over a designated time span [75]. Researchers have harnessed this data to prognosticate stock market trends [5].

Melody et al., within a Granger causality framework, reaffirmed the legitimacy of Google Trends as a reliable metric for gauging investor attention, thereby influencing the anticipation of stock market shifts [25]. In 2013, Kristoufek's predictive model underscored Google Trends as a significant factor influencing Bitcoin prices, evidenced by its notably low p-value [36]. The study illuminated that heightened interest corresponded to price surges; conversely, during instances when prices trended lower, intensified interest further contributed to downward pressure.In 2018, Abraham et al. harnessed Google Trends data to engage in linear modelling, thus accurately predicting the direction of cryptocurrency price changes [2]. Subsequently, in 2019, Nico's investigation into Google Trends' search analytics behaviour unearthed nuanced dynamics within cryptocurrency price behaviour. The findings revealed that the connection between cryptocurrency price movements, as identified by Google Trends, and internet search volumes was not universally positively correlated. Specifically, a robust negative correlation emerged between Bitcoin and Ether during June 2018 [68]. This analysis, conducted using an LSTM model, further established Google Trends as a potent predictor of Ether price movements [68]. In 2021, Maximilian et al. meticulously analysed a gamut of metrics to forecast hourly fluctuations in Bitcoin prices, incorporating Google Trends data. Employing models such as Vector Autoregression (VAR), Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM), the study highlighted a moderate correlation between Bitcoin prices and Google Trends data scales [65].

The research found that there are fewer studies on the correlation of cryptocurrency hourly changes with Google Trends data, and most studies focus on the accuracy of predictions with daily prices. However, since cryptocurrencies are highly oscillating markets, further validation and research are necessary to determine the relationship between Google Trends data and Bitcoin.

2.3 Sentiment analysis

2.3.1 Sentiment analysis

Traditional financial markets have a long-established history, providing abundant structured data for research and commercial purposes [72]. For the cryptocurrency market, accessing structured data is not always easy, which means finding alternative sources of data is important for predicting returns [26]. In the context of economics, Kaplanski et al. defined sentiment as any misunderstanding that may lead to mispricing of the underlying value of an asset. In this way, sentiment can make an asset speculative [30].

Natural language processing (NLP) techniques for cryptocurrency return prediction encompass topic modelling [40], semantic analysis [34], and sentiment analysis [47]. This study specifically focuses on sentiment analysis, a subfield of NLP that computationally analyses people's opinions, evaluations, attitudes, and sentiments regarding various entities and topics [39]. The primary objective is to determine the polarity (positive, negative, or neutral) of unstructured text [52]. Sentiment analysis encompasses libraries like "NLTK", "VADER" and "TextBlob" which provide deep learning-based models [58]. The methods for sentiment analysis have progressed from lexicon-based approaches to advanced models like the Transformer model [44].

Sasmaz et al. analysed sentiment in NEO cryptocurrencies and its correlation with daily tweets sentiment and NEO market prices. They employed Random Forest and BERT classification models for sentiment analysis, with Random Forest outperforming BERT. Additionally, they compared daily tweets sentiment with NEO market prices in the study's second phase [64]. Naila et al. performed sentiment analysis and detection, specifically focused on tweets related to cryptocurrencies. They proposed a deep learning integration model called LSTM-GRU, which combines the LSTM and GRU recurrent neural network applications. The GRU was trained using features extracted from the LSTM. The study reported impressive results, with the integrated LSTM-GRU model achieving a high accuracy of 99% for sentiment analysis and 92% for sentiment prediction, especially when utilizing Bow features [9].

Gül et al. recently focused on sentiment classification of cryptocurrency-related tweets, specifically positive and negative sentiments. They introduced a novel hybrid architecture to enhance sentiment analysis accuracy. The architecture incorporated a convolutional layer to extract local features and a grouping enhancement mechanism that assigned weight values to intuitive features. Additionally, in order to extract comprehensive global features from enhanced context vectors, their study employed an attention mechanism and a fully connected layer within a bidirectional layer. Experimental results demonstrated that the proposed architecture achieved an accuracy of 93.77%, outperforming the state-of-the-art architecture [23].

2.3.2 Sentiment analysis-based cryptocurrency prediction

Numerous studies have explored the potential value of incorporating sentiment scores from investment-related social media and online platforms into cryptocurrency return prediction models [26]. For instance, in 2015, Ifigeneia et al. utilized time series analysis to examine the link between cryptocurrency prices and economic variables, technical factors, and sentiment measures derived from Twitter messages [20]. They leveraged the support vector machine (SVM), a machine learning algorithm, for sentiment analysis. The findings indicated that the sentiment ratio on Twitter, Wikipedia search queries, and hash rate positively influenced Bitcoin prices. Conversely, the exchange rates of the US dollar and euro had a negative impact on Bitcoin's value [20]. Furthermore, in 2016, Young et al. investigated user comments in online forums to forecast cryptocurrency prices and transaction volumes. Their findings demonstrated that user comments and replies in these communities affected transaction volumes among users [32]. They further analysed frequently discussed topics by Bitcoin users and connected their significance to Bitcoin transactions [33].

Galen's research revealed a correlation between Twitter sentiment and people's overall emotional state. Interestingly, social media platforms like Twitter were found to have a calming effect on users rather than amplifying their emotions [54]. Building upon this insight, Jethin et al. proposed a method in 2018 to predict Bitcoin and Ethereum price changes with Twitter and Google Trends data. They employed VADER for sentiment analysis and incorporated the results into a linear model for prediction. The study concluded that Google Trends data and tweets volume better reflected the overall interest in owning cryptocurrencies [2]. Krzysztof expanded on this approach by developing a hybrid model based on multiple machine learning models to predict cryptocurrency prices. Their findings highlighted that a combination of Google Trends data and general negative sentiment yielded the strongest predictive power [75]. Similarly, Smuts employed an LSTM model for a binary sentiment-based price prediction method using Google Trends and Telegram sentiment [68]. Their backtesting on the first half of the 2018 test set achieved a 76% accuracy for hourly prices [19].

Subsequently, Tiran et al. employed a big data approach, analysing millions of posts from Twitter, Telegram, and Reddit, to investigate the impact of social media platforms on the price and trading volume of Bitcoin and other top cryptocurrencies. Their study revealed that trading volume, particularly transaction volume, predicted price fluctuations in Bitcoin. Additionally, they found that Reddit and Telegram posts had a more significant influence on Bitcoin trading volume compared to Twitter [62]. In a separate study, Kwansoo et al. utilized a Hidden Markov Model (HMM) and analysed six months of Twitter and Google text data to examine how social sentiment interacts with the cryptocurrency market during bull and bear markets. Their findings suggested that social sentiment exhibited relatively higher correlation during bull markets compared to bear markets. Moreover, they observed that positive social sentiment had a stronger impact on localized downtrends, while negative social sentiment had more influence during localized uptrends [31].

As technology develops, researchers are increasingly enhancing the performance of predictive models by increasing their complexity. Raj et al. proposed a hybrid framework called DL-Gues for cryptocurrency price prediction. This framework integrated LSTM and GRU models, considering interdependencies among cryptocurrencies and market sentiment. By incorporating historical cryptocurrency prices and recent Twitter sentiment, their model achieved a low MSE of 0.0185, indicating accurate price predictions [55]. Similarly, Zi et al. utilized LSTM and GRU models and employed stacking integration techniques to enhance decision accuracy. In addition to social media comments, their model incorporated technical indicators as input. Experimental results demonstrated that near real-time predictions outperformed daily predictions, with a mean error (MAE) improvement of 88.74% [76].

Expanding on earlier studies, Man-Fai et al. developed an advanced system that integrates public sentiment analysis, a price tracking mechanism, and machine learning algorithms to generate trading signals for cryptocurrencies. Their system encompasses a real-time price visualization module for displaying cryptocurrency price data. Additionally, it incorporates a prediction function that offers short-term and long-term trading signals based on sentiment scores derived from tweets about cryptocurrencies from the previous day. To enhance sentiment analysis accuracy, they employed a CNN-LSTM model instead of a TextBlob model, and for cryptocurrency price prediction, they utilized a random forest regression model instead of a decision tree model [37]. Similarly, Duygu et al. discovered that fine-tuning with wake-tags improves the predictive value of text-based features and enhanced the performance of cryptocurrency return predictions [26].

According to the aforementioned studies, it becomes evident that the majority of studies centre around forecasting Bitcoin prices using regression analysis based on Twitter public sentiment. Fewer studies delve into short-term predictions, especially pertaining to the combined impact of various social platforms on Bitcoin. Furthermore, previous research suggests that it's more practical to focus on price direction rather than predicting specific prices. However, many models that use sophisticated deep learning concentrate on predicting specific values rather than price direction. Therefore, this study aims to enhance the existing research by analysing the impact and predictive effects of various social platforms on the short-term direction of Bitcoin prices. The objective is to explore the optimal metrics and models for predicting the direction of cryptocurrency prices.

2.4 Bitcoin Price Prediction

This subsection provides a brief overview of prior research pertaining to Bitcoin price prediction using machine learning and deep learning methodologies. The domain of stock market forecasting using daily and readily accessible high-frequency data has evolved over several decades [12]. Nevertheless, a dearth of comprehensive research focusing on Bitcoin price prediction remains. Previous studies have approached Bitcoin price prediction through empirical analysis and robust machine-learning algorithm evaluation.

Machine learning algorithms have gained prominence for their capacity to yield precise predictions across diverse domains. Yet, the direct replication of these algorithms can introduce challenges, including overfitting, due to the intricate nature of the methodology. Zheshi et al. sought to address this by concentrating on the applicability of distinct modelling techniques to datasets characterized by varying structural and dimensional attributes. Through the classification of Bitcoin data based on daily and high-frequency prices, the study harnessed an array of high-dimensional features encompassing assets and networks, quotations, and the spot price of gold. Employing machine learning methodologies such as Random Forests, XGBoost, and Support Vector Machines, the research achieved a notable accuracy of up to 66% [12].

Cryptocurrency price data, categorized as a type of time-series data, has prompted substantial research employing time-series forecasting models, including ARIMA, to predict Bitcoin's price. In 2019, Amin demonstrated the utility of the conventional Autoregressive Integrated Moving Average (ARIMA) model in projecting Bitcoin's future value. This was achieved through an analysis of price time series spanning a 3-year timeframe [10]. Empirical investigations have indicated the efficacy of this straightforward approach, particularly within sub-periods characterized by relatively stable time series behaviour, especially in the context of short-term forecasting. However, issues arise when employing the ARIMA model for long-term forecasting, following training over a 3-year span. Given Bitcoin's varied price behaviour over that period, the model tends to yield significant forecasting errors, particularly when faced with abrupt price fluctuations. Moreover, the research underscores that the positioning and duration of the time window exert an influence on price prediction outcomes [10]. Triyanna et al. similarly leveraged the ARIMA model for short-term prediction within a day-based context, spanning Bitcoin data from 2013 to 2019. Their study revealed ARIMA's capability to forecast Bitcoin prices 1 to 7 days ahead, yielding favourable results [74]. Sean et al. employed Bayesian-optimized RNN and LSTM networks to predict price movement direction and contrasted their performance with the conventional time series prediction model, ARIMA. The study underscored that deep learning networks exhibit superior performance over the traditional ARIMA model [43]. Similarly, in 2019, Suhwan et al. conducted a comprehensive investigation and comparison of cutting-edge deep learning methodologies. This encompassed DNNs, LSTM models, convolutional neural networks, and deep residual networks, as well as their amalgamations, for predicting Bitcoin price [28]. Empirical findings indicated that while LSTM-based prediction models showcased a slight advantage in Bitcoin price regression forecasting, the DNN-based model excelled in predicting price rises and falls. Notably, the research emphasized that the most effective prediction model for regression might not necessarily be the optimal choice for classification, and vice versa [28].

In 2021, Mohammed and his team suggested using high-latitude features for a machine learning-based time-series prediction of Bitcoin prices [45]. Their study employed various machine learning models like SVM, ANN, and others to forecast the price in the short to medium term. The study included all the price metrics as features, including block size and hash rate. The LSTM model performed the best among all the models and showed up to 65% prediction accuracy for the next day's daily price classification [45].

Furthermore, several studies have explored Bitcoin price prediction using hybrid models and high-dimensional features. Saúl et al. introduced a CNN-LSTM hybrid model that combines CNNs and LSTMs to predict intraday trends in cryptocurrency prices. This hybrid model incorporates technical indicators from the high-frequency market and was compared with standalone CNNs, LSTMs, and multilayer perceptron networks. The experimental results consistently demonstrated the superiority of the proposed hybrid model, particularly in predicting Dash and Ripple prices [6].

Similarly, Yan et al. proposed a CNN-LSTM hybrid model for both regression and classification predictions of Bitcoin prices. This model incorporated external variables like macroeconomic indicators and investor sentiment [38]. Utilizing a two-year dataset, the study found that the CNN-LSTM hybrid model consistently outperformed other approaches in Bitcoin price prediction. This further underscores the potential benefits of hybrid models and the incorporation of diverse data sources for enhancing cryptocurrency price predictions.

Previous research studies have shown that most Bitcoin price prediction studies are biased towards certain days and often span several years. Successful models often incorporate market trading information and require in-depth technical indicator inputs. However, this study aims to shift the focus towards the impact and predictive power of social media information on Bitcoin prices, rather than relying on technical indicators.

2.5 Action Recommendation Model

In the cryptocurrency market, what matters is not only the actual value of the price prediction, but also the maximization of profits in the short term [58]. Nevertheless, existing models often focus on specific value predictions, leaving investors to calculate and determine their actions to maximize profits. To overcome this limitation, recommendation models provide actionable suggestions, such as "Sell", "Buy", or "Wait" to optimize profitability. Nevertheless, research on recommendation models is limited compared to price prediction models, which leads to poorer performance [58].

Nelson et al. conducted a study comparing the performance of pseudorandom-based, MLP, Random Forest, and LSTM action recommendation models for stock prices. They evaluated two action categories ("Sell" and "Buy") and found that the LSTM-based model outperformed the other models, achieving an average accuracy of 56% [48]. Similarly, Sanboon et al. performed a comparable study to evaluate different action recommendation models, including LSTM, decision tree, random forest, logistic regression, MLP, SVM, and KNN approaches. The findings from their experiments revealed that the LSTM-based model outperformed the other models, showcasing superior performance in the evaluated metrics [63].

Building off this study, Park et al. aimed to enhance the performance of the behavioural recommendation model by introducing three additional features ("sellProfit", "buyProfit", and "maxProfit") based on the predicted next-day cryptocurrency price. They conducted experiments using LSTM-based classification models with price data from various cryptocurrencies, including BTC, ETH, etc., over the period of almost 3 years from 2018, to 2021. The results of the experiment showed a statistically validated improvement of around 13% to 21% in performance when utilizing the proposed three input features compared to not using them [57].

Recently, a novel approach was introduced by Park et al. to adapt the outcome of the action recommendation model using sentiment analysis on Twitter data, which improved the performance by roughly 3% compared to the aforementioned approaches [58].

As can be seen from the above studies, research in this area is very poor and lacks depth. Moreover, it is not only Twitter tweets that affect the price of cryptocurrencies, but also many other factors. However, the adjustment method proposed in previous studies solely concentrated on Twitter tweets, presenting a limitation. Hence, further research endeavours are warranted to address these aforementioned constraints.

Chapter 3

Methodology

The main objective of this chapter is to outline the research methodology framework and the various methods used in each stage of the process. The framework is based on the CRISP-DM approach, with emphasis on data preparation, model development, and strategies for recommending actions. Details on the evaluation of the models and their subsequent implementation will be presented in Chapter 4.

3.1 Architecture

One of the main aims of this project is to explore the relationship between hourly social media data and Bitcoin data, investigating the delayed impact of public sentiment on Bitcoin price. Based on this exploration, the project selects features to predict both Bitcoin price and its direction. This is done to validate their relationship. The optimal predictive model is then used to provide trading action recommendations for the next timestep, which includes "Buy", "Sell", and "Wait" actions.

Public sentiment is obtained through sentiment analysis, and beyond that, the study also explores the short-term lagging impact of Google Trends data on Bitcoin price. All data in this study is based on hourly intervals. Finally, backtesting is conducted on the trading action recommendations to obtain the overall total return over a period of time. The entire process adheres to the Cross-Industry Standard Process for Data Mining (CRISP-DM) [49] methodology, a process model used for data mining and machine learning projects. It consists of six stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment [49]. This study adapts this methodology to its needs, and the overall architecture is illustrated in the Fig 3.1. The following sections will provide detailed descriptions of the methods employed in each stage based on the adjusted workflow.

3.2 Data Collection

In order to explore the potential correlation between media-related information and Bitcoin trends, as well as to evaluate the capacity of media data to predict hourly Bitcoin price fluctuations, a comprehensive investigation incorporates multiple diverse data sources beyond historical Bitcoin price data. Three key variables are under scrutiny. The first variable encompasses the sentiment scores and tweets volume associated with Bitcoin-related tweets on Twitter. The second variable involves an examination of posts within the Bitcoin-focused subreddit on Reddit. Lastly, the third variable entails the integration of data derived from Google Trends. This section provides an in-depth exposition of the methodologies adopted for the collection of this array of diverse data.

3.2.1 Financial Data

In contrast to the majority of cryptocurrency studies, this project diverges by gathering Bitcoin data not on a daily basis but at hourly intervals. The historical price data for Bitcoin was sourced from Binance's historical data API. Binance serves as a cryptocurrency trading platform where users can engage in the buying, selling, and trading of a diverse range of cryptocurrencies. Specifically, the *binance_historical_data*[1] API was utilized to access historical cryptocurrency data stored on Binance's servers. The "BTCUSDT" currency pair is widely traded with substantial volume, attracting active

[1] https://pypi.org/project/binance-historical-data

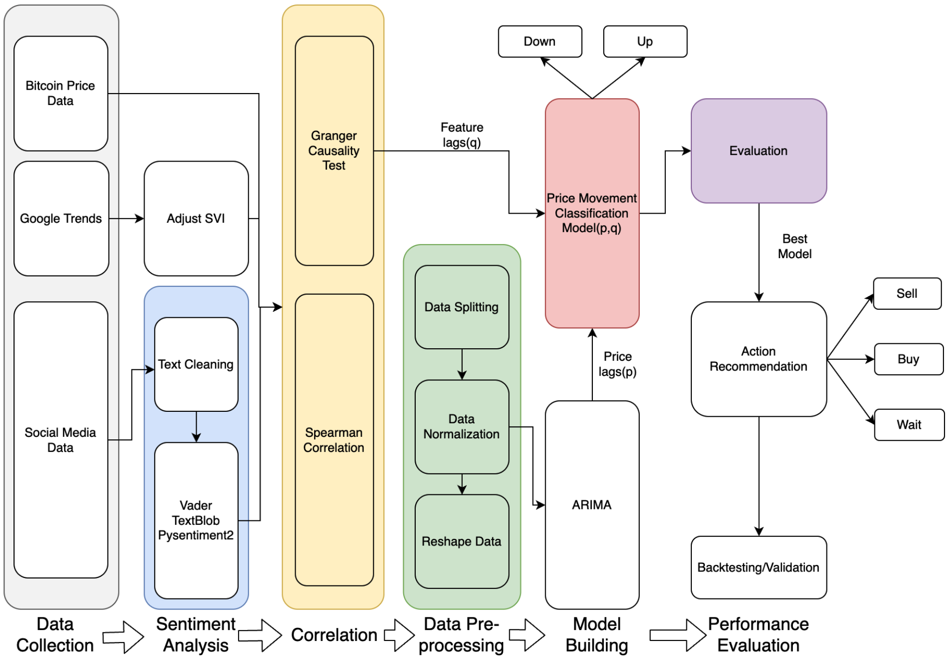


Figure 3.1: An overview of the workflow

traders and investors and generating a wealth of trading data. "USDT" is a stablecoin pegged to the U.S. Dollar. Due to the popularity and stability of the "BTCUSDT" pair, it's commonly used for market and technical analysis. This project aims to gather historical Bitcoin price data in USDT from 1st January to 30th June 2022. The focus is on hourly data for the first half of 2022, capturing price and volume trends. The collected data will be stored as separate monthly ".csv" files locally.

3.2.2 Social Media Data

Social media data will be obtained from the two most popular and high daily traffic social media platforms, Twitter[2] and Reddit[3]. This subsection describes the acquisition of social media data.

Tweets Data

Due to the limitations of Twitter's official API, in this project, tweets data was sourced from the publicly available historical dataset in Kaggle[4], which leveraged the Snscrape Python package to collect more than 22 million English-language tweets mentioning the word "Bitcoin" over the time period from 1st January 2021 to 30th June 2022. The dataset is divided into two files: one containing 15.6 million tweets from 2021 and the other comprising 8.17 million tweets from the first half of 2022. This study is exclusively concerned with the data from January to June 2022. The Table 3.1 illustrates sample data, with the username listed as anonymous.

Reddit Posts Data

The second source of social media data comes from Reddit, with a particular focus on cryptocurrency-related and Bitcoin-related subreddits. Initially, the plan was to directly collect data using the Reddit API and a Python crawler tool. However, the official Reddit API offers only around 1,000 data points per hour for regular users, falling significantly short of meeting the research needs. As a result, the ultimate source of Reddit data came from Kaggle, which provides a publicly accessible historical dataset of Reddit posts related to cryptocurrencies. This dataset spans from 1st January to 30th December 2022, which was compiled by utilizing a combination of a crawler and the pushshift API to gather Reddit submission IDs, followed by the Python Reddit API to retrieve detailed submission information.

[2]https://twitter.com/home
[3]https://www.reddit.com/
[4]https://www.kaggle.com/

datetime	date	username	text
2022-01-01 23:59:56+00:00	2022-01-01	username1	0.4MOT TOKENS IN #LATOKEN airdrop and maybe more! #Bitcoin and #Cryptocurrency
2022-01-01 23:59:53+00:00	2022-01-01	username2	MARA for Bitcoin Exposure: Top Trade Q1 2022 https://t.co/EZ2S4cfFIq https://t.co/srsvNPA2jV

Table 3.1: Sample Tweets Data

Given the diverse landscape of cryptocurrency-related discussions on Reddit, this study specifically concentrates on the subreddits "r/bitcoin", "r/btc", and "r/cryptocurrency", with a combined post count of 243,187 in "r/cryptocurrency", and 66,576 and 21,353 in "r/bitcoin" and "r/btc" respectively. Each dataset consists of 25 columns, covering a comprehensive range of attributes.

3.2.3 Google Trends Data

Google Trends[5] furnishes insights into the relative popularity of specific search terms compared to others, offering a temporal dimension to this comparison. This dynamic has the potential to serve as a representative signal for gauging the current level of widespread interest in bitcoin. Such interest fluctuations might potentially correlate with bitcoin price movements, as interest levels invariably rise and fall [2].

Google Trends, a service by Google, provides trends in search terms without yielding precise search volume. Instead, it offers a Search Volume Index (SVI) [2]. The SVI gauges the frequency of total searches for a given time frame and geographic region, relative to the overall searches for that time period. These values are then normalized to a scale between 0 and 100, factoring in the proportion of the search term to all search terms [2]. For trend data spanning over 90 days, SVIs are presented on a weekly aggregation [2]. In this study, the Pytrends API[6] was harnessed to gather Google Trends data, focusing on the search term "Bitcoin" on a weekly basis. The data encompasses the interval from 24th December 2021 to 1st July 2022, with hourly data increments. Since the data is inherently weekly-scaled, subsequent manipulation is requisite to derive the SVI for each data point, taking into account the broader timeframe.

3.3 Data Preparation

3.3.1 Financial Data

The initial Bitcoin price data uses numeric timestamps. To integrate it with other sources, the "datetime" package in Python converted these timestamps into datetime. Unnecessary columns like "Quote asset volume" were removed. The "Open time" column was chosen as the unique key in time for each dataset. Data from different sources was merged within a loop, and the "yfinance" package was leveraged to retrieve missing Bitcoin price data from Yahoo Finance[7].

3.3.2 Social Media Data

Pre-processing

Prior to delving into social media data analysis, a preliminary step involves pre-processing the raw data. This subsection outlines the initial processing applied to both tweets data and Reddit posts. This preparation paves the way for subsequent sentiment analysis.

For tweets data, it is essential to normalize the date format to "%Y-%m-%d %H:%M". Variables were renamed, and the redundant "date" column was discarded. Each tweets in the raw data corresponds to a distinct time point. To align with Bitcoin's historical price data, tweets were initially sorted and grouped by the hour. The specific time points were rounded off to facilitate hourly grouping and counting of Bitcoin-related tweets.

[5]https://trends.google.com/trends/
[6]https://pypi.org/project/pytrends/
[7]https://uk.finance.yahoo.com/

For each subreddit, data was extracted from the original Reddit posts dataset for the period 1st January to 30th June 2022. Dates were standardized, similar to tweets data. Unlike tweets, each Reddit row represents a submission record, not an actual post. Therefore, empty records were removed, and data from the three subreddits were merged for a preliminary processed Reddit posts dataset. Retained variables and their descriptions are shown in the Table 3.2.

Key	Description
datetime	Time point of post creation (grouped by hour)
upvote_ratio	Percentage of high votes among all votes on the post
selftext	Content of text-based post

Table 3.2: Variables Retained in Preliminary Processed Reddit Posts Data

Text Cleaning

Text cleaning is a crucial preprocessing step for sentiment analysis. Raw text data extracted from social media often contains noise such as special characters, punctuation, and formatting issues. Cleaning helps eliminate these undesired elements, mitigating biased language and offensive content to ensure an impartial analysis.

Tweets data is characterized by its brevity and high level of noise, necessitating extensive text cleaning for subsequent analysis. Firstly, accounts that frequently post tweets are considered noisy, as a substantial portion of their content comprises advertisements. Following Hirad's approach [24], the total tweets count was reduced from 8,173,354 to 5,711,375 by filtering based on usernames and retaining users who posted fewer than 1,000 tweets. Additionally, a list of Bitcoin-related noise words [24] provided by Hirad was utilized to remove noise words. Subsequently, URLs and HTML tags were stripped from tweets, and tagged abbreviations were expanded (e.g., "let's" into "let us"). Tweets often includes slang and acronyms (e.g., "idk" or "iow"), which introduces noise. To address this, a handcrafted list of internet slang [34] was employed to expand the text. Further tokenization and lexical form reduction involved removing punctuation, numbers, and redundant spaces using the WordNet Lemmatizer. English stop words were then eliminated, and other non-letter tokens were removed based on the "stopword" dictionary provided by the NLTK Python library. Case folding was applied to all tokens to achieve uniformity.

Apply the same text preprocessing steps to Reddit posts as for tweets. It's important to note that Reddit posts are generally longer texts. According to histogram statistics, most posts have lengths between 0 and 1,000. To ensure meaningful sentiment analysis, posts with lengths between 7 and 1,000 are retained. This prevents excessively short sentences that can impact sentiment scoring. After processing, the number of posts is reduced from 31,118 to 29,023.

Sentiment Analysis

Sentiment analysis involves extracting opinions or emotions conveyed through messages [69]. The resulting sentiment scores can be studied in relation to historical Bitcoin price data.

Sentiment analysis employs two methods: machine learning and lexicon-based. Both use a sentiment lexicon to assess polarity. This study favours dictionary-based analysis using tools like VADER and TextBlob. These tools were selected for their nuanced understanding of social media sentiment, especially in cryptocurrency discussions. Valence Aware Dictionary and sEntiment Reasoner (VADER) excels in short, informal social media texts, like Twitter. As shown by Kim et al. [32], VADER provides accurate polarity scores for predicting cryptocurrency price fluctuations. VADER is able to generate a compound score normalized between -1 and +1, indicating the level of positivity or negativity in each tweet. On the other hand, TextBlob offers polarity and subjectivity attributes per tweets. Polarity scores range from -1.0 to 1.0, revealing negative or positive sentiment, with 0.0 indicating neutrality. TextBlob's lexicon-based method enables preprocessing-free sentiment analysis.

Furthermore, the Loughran and McDonald Financial Corpus[8] was also explored and tested. Unlike VADER's sentiment score results, this corpus was incorporated into the Pysentiment2 library, generating a dictionary with counts of Positive and Negative tokens. To calculate the sentiment score, normalization is applied using the Formula 3.1 [50], where P and N represent the counts of Positive and Negative tokens per text, respectively:

[8] https://sraf.nd.edu/loughranmcdonald-master-dictionary/

$$\text{sentiment score} = \frac{\text{P - N}}{\text{P + N}} \tag{3.1}$$

Sentiment analysis techniques are applied to tweets, yielding compound and polarity scores. These scores are grouped by hour and averaged for hourly sentiment analysis. For relevant Reddit posts, upvote rates serve as approval metrics. They are incorporated to adjust compound scores and determine final post scores, as shown in the Formula 3.2:

$$\text{weighted sentiment score} = \frac{\sum_{i=1}^{n} u_i c_i}{\sum_{i=1}^{n} u_i} \tag{3.2}$$

where u_i is the upvote ratio and c_i is the compound score of each post. The final adjusted sentiment score is obtained by weighting and averaging the sentiment scores of each post per hour.

3.3.3 Google Trends

To accurately derive Google Trends scaling ratios, Erik Johansson's approach was employed to convert weekly and daily data into daily search volume data [17]. The Google Trends data was organized in weekly lists, ensuring alignment between the final hourly time point of one week and the initial time point of the subsequent week. Ratios for data at overlapping time points were computed by iterating through the list, and these ratios were retained as correction parameters. Two different strategies were employed for processing the data:

1. Within a loop, the correction ratios were employed to adjust the data for each time point in the following week, replacing the original values. When calculating the ratio between the last hour of the second week and the first hour of the third week, the corrected values from the second week were utilized.
2. Ratios for the consecutive two weeks were calculated within the loop and stored in a fresh list. These ratios were then applied to each week in the original list, except the first week, which lacked the last hour's data from the previous week. The corrected data didn't replace the original values.

The corrected data might not fall within the range of 0 to 100, necessitating further scaling. Ultimately, the data obtained through the two processing methods were saved in separate files as features to be selected.

3.4 Correlation Analysis

3.4.1 Spearman correlation

In the initial correlation analysis, the Shapiro-Wilk test [22] is used to check if each feature follows a normal distribution. This test assesses whether a dataset is normally distributed, with a high p-value (>0.05) indicating normal distribution and a low p-value indicating otherwise [22]. None of the individual eigenvalues in the data exhibited a standard normal distribution upon testing. To explore any potential monotonic relationship between adjusted social media features and Bitcoin's closing price, a preliminary correlation is conducted analysis using Spearman's correlation coefficient.

Spearman's correlation coefficient measures the strength and direction of monotonic relationships between variables, suitable for nonlinear or sequential associations. Ranging between -1 and +1, it indicates the degree of monotonic trend. Positive values suggest co-increase, negative values imply co-decrease [29]. Values closer to -1 or +1 indicate stronger correlations. Spearman analysis captures nonlinear and sequential links, improving understanding [8]. However, it can't establish causality or address lagged effects.

3.4.2 Granger causality testing

The Granger Causality test, used in time series analysis, explores potential causal links between two variables by determining if past values of one variable offer statistically significant predictive insights for another [51]. For instance, given time series X_t and Y_t, the test examines whether Y_t predicts X_t. This entails steps like checking data stationarity through the Augmented Dickey-Fuller (ADF) test [67], a prerequisite for reliable results. Once confirmed, the lag order is chosen — past time points (lags) used to gauge potential delays in causality.

Next, the null hypothesis posits that Y_t does not precede X_t. The F-test statistical test is applied to evaluate whether incorporating lagged values of Y_t significantly enhances the predictive capacity for X_t. The null hypothesis is rejected if the p-value for at least one lag is below the significance level (usually 0.05). The autoregressive model for time series X_t and the model equation expanded with increased lagged values of Y_t are illustrated below [59]:

$$\begin{aligned} X_t &= \alpha + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \cdots + \gamma_p X_{t-p} \\ X_t &= \alpha + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \cdots + \gamma_p X_{t-p} + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} \end{aligned} \tag{3.3}$$

where p is the lagged value.

Additionally, note that Granger Causality testing requires both variables to be stationary time sequences. Non-stationarity could lead to misleading results [34], making the test invalid. When analysing the connection between social media messages and Bitcoin price, ensure the stationarity of both variables. Non-stationary series can be transformed into stationary sequences through differencing for accurate testing.

3.5 Features and Targets

After preparing the public sentiment data, which includes social media information and historical Bitcoin prices, the datasets are merged based on time points. This generates sets of features and targets for price prediction. This section explains the initial feature set, along with the three types of predictions for Bitcoin price data.

3.5.1 Initial Feature Set

An initial dataset was created by combining social media sentiment scores, adjusted Google Trends data, and financial information. All features were merged based on timestamps. The social media data occasionally had missing sentiment scores at specific time points. To deal with this, two approaches were explored: either directly removing rows with missing data or assuming contiguous sentiment and filling gaps using scores from the previous time point. Both methods were tested initially, and the final decision was to fill missing values with sentiment scores. This choice was made to maintain continuity in time, as removing rows disrupts the sequence and could impact experiment effectiveness. The initial feature set is displayed in the Table 3.3, encompassing all studied features. A subset of these features will be utilized in subsequent analysis to assess the individual impact of variables.

Feature set (1)	**Feature set (2)**
Close Price	Tweets sentiment compound score
Adjusted google trends	Tweets polarity score
Reddit posts compound score	Tweets subjectivity score
Reddit posts polarity score	Tweets sentiment score from pysentiment2
Reddit posts subjectivity score	Tweets volume

Table 3.3: The initial feature set

3.5.2 Targets

This study examines using social media data for three types of Bitcoin price prediction: Bitcoin closing price, price returns, and price movement direction. In the cryptocurrency field, predicting Bitcoin's price is a common challenge, and initial efforts focus on forecasting the hourly closing price using models. However, due to Bitcoin's volatile nature, directly predicting its price is impractical, especially given the irregularity of financial time series [7]. Non-stationary time series data is hard to predict using linear models, but can be addressed using deep models with consistent price ranges in training and test sets.

In trading, emphasis often shifts to price returns and movement direction rather than exact prices. Price returns capture these dynamics as the difference in closing prices over time, calculated as Formula 3.4:

$$r_{t,t-k} = \frac{p_t}{p_{t-k}} - 1 \tag{3.4}$$

where p_t represents the Bitcoin price at time t, and k indicates the number of cycles for return calculation [27]. In this study, when $k = 1$, it signifies assessing the price change at the current time compared to the previous time point.

Price movement direction prediction involves a binary classification task, assigning a label y to each data point in the dataset. The label is defined as follows:

$$y_t = \begin{cases} 1 & p_{t-1} < p_t \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

where "1" signifies a price increase at a time point t compared to the previous time point [48], in which case the asset is typically held long at the time $t - 1$, implying a "Buy" operation. On the other hand, "0" indicates a stable or decreasing price, suggesting the asset is generally held short, implying a "Sell" operation.

3.6 Data Pre-processing

After feature selection, the time series data requires reconstruction before being fed into the model. For each value to be predicted, the features and corresponding target values at a past time point are utilized as input features. This subsection outlines crucial data preprocessing steps undertaken during the experiment, encompassing the reconstruction of dataset, dataset splitting, and data normalization.

3.6.1 Dataset for time series analysis

Time series forecasting is an analytical technique that leverages historical values and associated patterns to predict future trends. It involves predicting the closing price or other price patterns at the next time point based on historical data from a specific timestep. For instance, to predict the next time point's price using price data from the past 15 time points, the input values and target value would be $[P_1, P_2, \cdots, P_{15}]$ and $[P_{16}]$ respectively, where P represents the price.

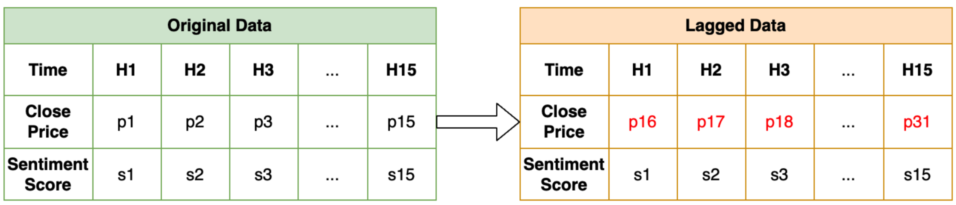


Figure 3.2: The feature of dataset with lag 15

Constructing a time series dataset with a lag or delay value involves organizing the data in a manner that associates each observation not only with its own attributes but also with past observations from previous time steps. These lags are instrumental in capturing temporal dependencies and relationships. A challenge in this study is to determine the optimal lag duration to effectively unveil the influence of Bitcoin-related social media data on Bitcoin's price. The lagged values of public sentiment can be somewhat determined through methods like Granger Causality Testing and Spearman correlation, as discussed earlier. By employing Bitcoin-related public sentiment scores and Bitcoin price data, a time-series dataset was formed with lagged features [13]. This entailed shifting individual features backward by the corresponding lag time being examined, as illustrated in Fig 3.2. Assuming a timestep of 3, the features and target values for each time point in the restructured lagged dataset would resemble those shown in Fig 3.3.

3.6.2 Data Splitting

After preprocessing, 4,344 data points across six months were available. To ensure accurate model training and testing, a common 80:20 training-testing ratio was used. This ratio balances training data volume with a suitable testing subset. A validation set ratio of 0.2 aided in assessing performance and

Original Data				
Time	F1	F2	F3	F4
H1	1	10	11	20
H2	2	9	12	19
H3	3	8	13	18
H4	4	7	14	17
H5	5	6	15	16

Modified Data									
Time	F1(t-3)	F2(t-3)	F3(t-3)	F4(t-3)	...	F1(t-1)	F2(t-1)	F3(t-1)	F4(t-1)
H4	1	10	11	20	...	3	8	13	18
H5	2	9	12	19	...	4	7	14	17

Figure 3.3: The dataset with 3 timesteps

optimization during training. Parameters were tuned using a validation subset extracted from training data, while different shuffled datasets were employed for parameter evaluation to avoid bias. Notably, dividing training and test sets must occur before normalization to prevent information leakage from the test set, preserving model evaluation accuracy.

3.6.3 Data Normalization

Normalization is crucial in preparing the initial dataset for model training, especially when features have varying ranges. This step ensures reliable predictions from machine learning algorithms. By standardizing data into a common range, normalization ensures equal feature influence, while reducing sensitivity to outliers. The use of MinMaxScaler[9] separately on training and test sets establishes consistent scales for all features. This transformation confines data within a uniform 0 to 1 range, effectively reducing errors from price magnitude disparities [21]. The formula of MinMaxScaler can be seen in Appendix A.1.

3.7 Models

This study employs five models, including the ARIMA model as the baseline for regression in price prediction and the Random Forest Classifier as the baseline for classification. The deep learning model is primarily utilized for predicting price trends.

3.7.1 Baseline Models

ARIMA

Autoregressive Integrated Moving Average (ARIMA) is a time series forecasting method [14] designed to predict future trends in time series data. It disregards independent variables in the forecasting process and is suitable for highly correlated data, requiring autocorrelation, trend, and seasonality assumptions to hold. The method is effective in predicting historical data influenced by complex factors, particularly in the short term, with high accuracy. It also handles seasonal fluctuations well. ARIMA comprises four primary model groups: Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) [74]. The AR model focus on the relationship between current and previous period data, determined by the autoregressive order p. The MA model examines the relationship between the error term and the previous error term, indicated by the moving average order q. The ARIMA model combines the effects of both data and the previous error term, represented by the Formula 3.6.

$$\begin{aligned} Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} \\ + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \end{aligned} \tag{3.6}$$

where ε_t is white noise, and c is the constant. However, the ARIMA model requires data must be stationary, meaning that the average change in the data is consistent. Non-stationary data must be transformed into stationary data through differential processing. Therefore, before utilizing the ARIMA model, the data needs to undergo a smoothing process using the Augmented Dickey-Fuller Test (ADF). ARIMA parameters encompass p, d, and q, representing autoregressive, differencing, and moving average

[9] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

orders, collectively shaping model intricacy. Parameter selection often involves the examination of Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots. PACF analysis considers data patterns alongside individual time lag effects, aiding in determination of the autoregressive order (p) [78]. Furthermore, the value of the parameter d is dictated by applied differencing operations, which affect the data's transition into a stationary series.

To choose the optimal ARIMA model, the performance of various models is typically assessed using the Akaike Information Criterion (AIC). The AIC is defined by the Formula 3.7, where a lower AIC indicates a superior model [60]. By comparing AIC values across different parameter combinations, the most suitable ARIMA model can be selected for precise time series forecasting.

$$AIC = 2k - 2l \tag{3.7}$$

where l is a log-likelihood and k is a number of parameters [60].

In summary, the ARIMA model is a reliable approach for regression forecasting of time series data. The process involves autoregression, differencing, and moving averaging to establish the model. By examining ACF and PACF plots and utilizing AIC to determine the optimal model, future trends in time series data can be predicted with precision. For this project, the ARIMA model serves as the baseline model for forecasting actual prices.

Random Forest Classifier

For the binary classification task of predicting price direction, this project employs a random forest classifier as the baseline model, a commonly used approach in the field of classification. A random forest operates by creating multiple decision trees, with the final outcome based on the collective outputs of most trees chosen by the random forest. One key advantage of the Random Forest algorithm is its ability to mitigate overfitting risk and minimize necessary training time. It offers elevated accuracy and efficient handling of sizable datasets, suitable for both classification and regression tasks [4].

Decision trees are the core of random forests and can be applied widely in machine learning. However, decision trees that grow excessively deep to capture intricate patterns can lead to overfitting the training set. Random forests mitigate this by training many decision trees on diverse feature subsets, using metrics like entropy or Gini impurity for splitting criteria. Classifier hyperparameters are tuned via grid search and cross-validation, optimizing "n_estimators" and "max_depth". The final model is trained on the whole dataset. While versatile, random forests may struggle with data having significant sequential dependencies.

3.7.2 Deep learning-based Models

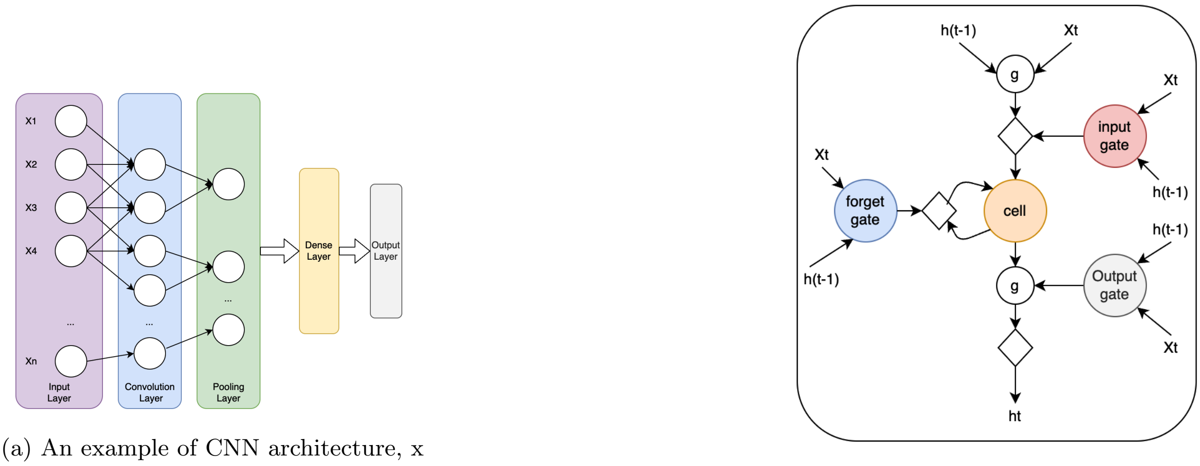
CNN

A neural network called Convolutional Neural Network (CNN) is designed specifically for processing structured lattice data, including images and sequences. It uses convolutional layers to automatically learn spatially hierarchical features from the input data. The typical architecture includes convolutional, pooling, and fully connected layers [6]. CNNs can learn hierarchical features from input data, making them appropriate for image and sequence processing tasks that involve two-dimensional matrices. Convolutional and pooling layers are combined multiple times to extract high-level feature vectors that are processed by the hidden and output layers [42]. Fig 3.4a shows an example of the basic CNN architecture, which includes learnable filters in the convolutional layer that slide over the input data to detect various features.

LSTM

The Long Short-Term Memory (LSTM) neural network is an extension of the Recurrent Neural Network (RNN) architecture that aims to retain temporal information while also enabling long-term memory within its cells. Unlike traditional RNNs, LSTM neural networks consist of memory blocks or cells. As part of the recurrent neural network layers, the LSTM layer enhances information flow control through individual memory units and adaptive gate units such as input, forget and output gates.

Each LSTM memory unit is supported by three key components: input gates, forget gates, and output gates. Forget gates retrieve relevant information from previous states of the memory block (long-term states) using the outputs of previous states or short-term states, controlling what information is to be discarded. The input gate determines the information needed to generate the current state, controlling what information is to be added to the cell or long-term state. Finally, output gates act as filters,



(a) An example of CNN architecture, x denotes the flattened input features

(b) An example of LSTM block

Figure 3.4: The basic architecture of deep learning models

controlling what information in the current state has an effect on the output (prediction) or short-term state. The architecture of the LSTM block is illustrated in Fig 3.4b, where x_t and h_{t-1} refer to inputs and predicted short-term states, and g is the tangent function [6].

One of the significant advantages of the LSTM layer is its ability to identify short-term and long-term correlations in time series data, effectively solving the problem of vanishing gradients. It excels in effectively analysing time-dependent variables and is suitable for classification and regression tasks. In this study, the LSTM model is used for regression and classification tasks related to price forecasting.

CNN-LSTM

This project intends to explore a hybrid CNN-LSTM model, in addition to the CNN and LSTM models. Saúl et al. introduced a CNN-LSTM model that is capable of predicting prices [42]. The model consists of five convolutional layer modules and one LSTM module, followed by eight fully connected layers, and outputs the results. The hybrid CNN-LSTM model has several advantages; it can extract hierarchical features from both spatial and temporal dimensions, enhance the processing of complex patterns, and reduce overfitting. However, this model is more complex and time-consuming, and requires a considerable amount of data. The basic structure of this model is shown in the Fig 3.5. This study tests the dataset with this model and compares its performance with CNN and LSTM models.

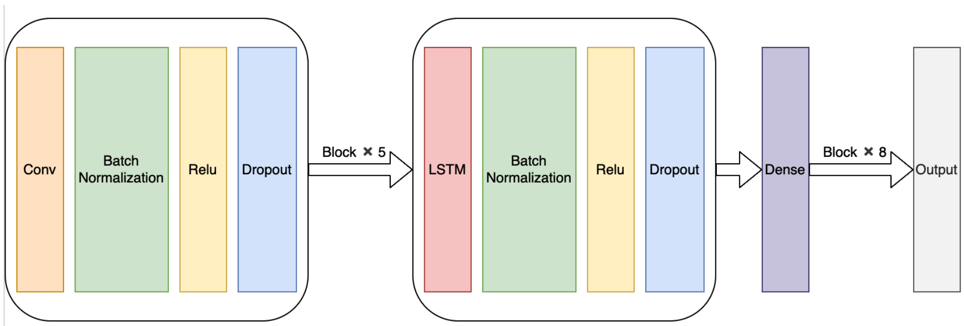


Figure 3.5: CNN-LSTM neural network architecture proposed by Saúl et al.

3.8 Performance Evaluation Methods

To assess predictions, essential regression metrics were employed: RMSE and R^2 Score. RMSE gauges average divergence between predictions and actual values, reflecting fit quality [79]. R^2 evaluates how effectively variables account for variance in models, with a higher score indicating better fit. Mean Absolute

Percentage Error (MAPE) appraises time series model accuracy by calculating average percentage differences between predicted and actual data [74]. The equations for RMSE, R^2, and MAPE are presented in the Appendix A.2, A.3, A.4.

When predicting price movement, a binary classification model is used. Two key metrics – Accuracy and F1-Score – assess the model's performance. Accuracy measures overall correctness, while F1-Score balances precision and recall. Both metrics range from 0 to 1, with higher values indicating better performance. The equations for accuracy and F1-Score are shown in the Appendix A.8 and A.5, respectively.

3.9 Backtesting Strategy For Action Recommendation

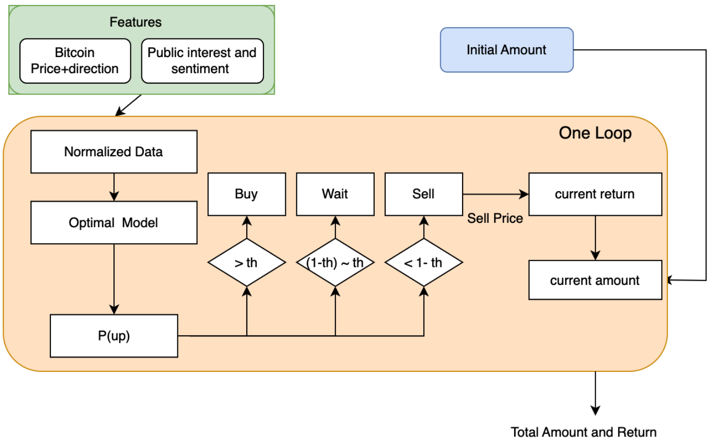


Figure 3.6: The basic framework of action recommendation model: $P(up)$ represents the probability of price going up on the next hour, th represents the threshold.

The action recommendation aims to suggest a buying or selling strategy for the immediate future, using a trained price prediction model. As shown in Fig 3.6, there are three potential actions available: "Buy", "Sell", or "Wait". The backtesting strategy for action recommendation is an iterative process based on historical price data. It's used to make buy or sell decisions based on the predictions from a pre-trained model and a set threshold. This is an iterative process that follows the steps outlined below:

1. To begin with, the flag of trading action is set to False, indicating that no positions are currently held. The historical price data is then iterated through, one timestep at a time, checking the model's predicted price direction for the next timestep.

2. If no positions are currently held and the probability of the model's predicted price direction going up exceeds a predetermined threshold, the strategy triggers a buy operation to purchase the cryptocurrency.

3. Conversely, if a position is currently held and the probability of the predicted price direction going up falls below a certain threshold, which means that the probability of a price decrease has exceeded a certain threshold. The strategy triggers a sell operation to sell the cryptocurrency.

4. If the probability of the model's predicted price direction going up falls between the set thresholds, the strategy waits. For each buy, sell, or wait operation, the strategy records the event and stores it in the "events_list".

5. The strategy will generate a chart that shows changes in price data and trading events. The chart will display buy and sell points, as well as waiting periods. Additionally, the chart will present the return on investment (ROI) and portfolio value over time.

In summary, the strategy utilizes a trained model for cryptocurrency price prediction to execute buy and sell operations based on model predictions and preset thresholds. The threshold values are explored within the range of 0 to 1. Trading events and investment performance are visually represented through charts, which illustrate how the portfolio's performance has changed over time. This approach provides a more formal description of the execution process and presentation of results for the backtesting strategy.

Chapter 4

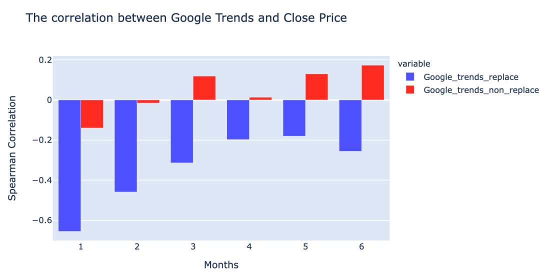
Critical Evaluation

This chapter will implement, discuss, and evaluate the results of the methodology presented in Chapter 3, which involves data selection during the data preparation phase and the analysis of the correlation between public sentiment and Bitcoin price. The chapter will conclude with the evaluation and discussion of the model's performance in relation to the outcomes from the three project phases: the regression prediction of Bitcoin price in the first phase, the classification prediction of Bitcoin price in the second phase, and the backtesting of the action recommendation strategy in the third phase.

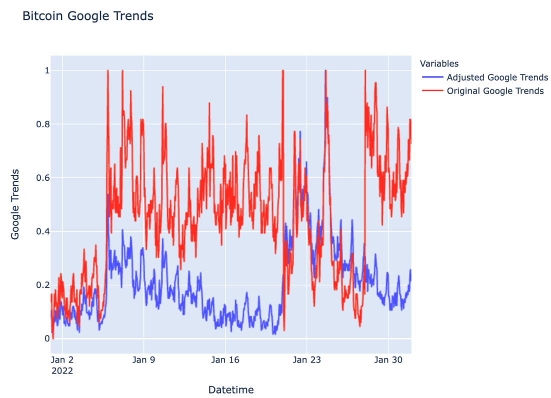
4.1 Results of Data Preparation and Correlation Analysis

4.1.1 Google Trends

The Section 3.3.3 discusses two approaches for correcting Google Trends values. The first method involves substituting corrected values for the original ones and using them in the next week's rate calculation. On the other hand, the second method corrects the values without replacing the original ones. Previous studies have mainly employed the first method, which may result in a smoother time series while maintaining temporal relationships. However, a few studies have also used the second method to preserve the original values' integrity. During the data preparation stage, both methods were separately applied, and the Spearman method was used to calculate the correlation between the adjusted Google Trends data and the closing price. High correlation data with closing prices were prioritized for further research.



(a) Comparison of Replaced and Non-Replaced Values in Correlation with Bitcoin Price



(b) Original Values Vs. Adjusted Google Trends Values

Figure 4.1: Data Analysis and Selection For Google Trends Data

Initially, correlations were examined over a six-month data span. The first method using an alternative to the original values revealed a more robust direct correlation between the closing price and Google Trends at -0.26, signifying a direct negative relationship between the two. In contrast, the second method exhibited an overall weak positive correlation of 0.17 across the entire six-month dataset.

To further validate these findings, the data range was incrementally expanded, commencing with one month of data. The results indicated that within the first five months, a shorter data timeframe led to a stronger direct negative correlation between Bitcoin price and Google Trends data. This trend is depicted in Fig 4.1a. Additionally, instances were observed where the correlation of the Google Trends-adjusted data using the first method consistently surpassed that without the adjustment. Consequently, the decision was made to proceed with the Google Trends data obtained through the first method for subsequent analysis. Fig 4.1b visually contrasts the original trend data with the adjusted Google Trends data, affirming the continuity achieved through the adjusted method.

Granger Causality testing was employed to examine the lagged influence of Bitcoin returns and price fluctuations on changes in Google Trends data over a 48-hour interval. The analysis commenced with a one-month data range, progressively extending to six months with separate tests.

Results revealed a significant causal impact of Bitcoin price returns and changes on Google Trends data alterations over the six-month period. Notably, from the first lag, all p-values of the statistical tests were below 0.05, indicating the capacity of price returns and changes to influence Google Trends shifts.

Consequently, a reverse examination was conducted. Findings demonstrated a causal relationship between Google Trends data and price returns starting from the 10th lag within one to five-month data ranges. However, this predictive ability extended to the second lag when the data range stretched to six months. This implies that the lagged impact of Google Trends on Bitcoin closing prices and returns might differ across various timeframes. Nevertheless, an overarching stable Granger causality emerged between Bitcoin closing price and Google Trends.

4.1.2 Tweets Data

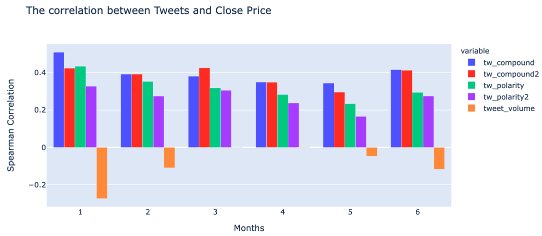
The tweets data encompasses sentiment scores derived from analysed tweets, which include the compound score, polarity score, and subjectivity score obtained through the VADER and TextBlob sentiment analyses. Sentiment score features based on the Loughran and McDonald financial corpus were excluded due to a substantial number of resulting 0-values, rendering meaningful sentiment trends unattainable. Moreover, the prevalence of numerous 0 values (neutral values) within the sentiment score necessitates consideration.

Initially, it was hypothesized that these neutral values might be attributed to robot-generated information. Thus, two approaches were explored to address this issue: the first involved excluding the neutral composite score and averaging hourly, while the second involved a direct hourly average without removing the neutral values. Both methods were separately attempted in this study and validated via correlation testing. At the beginning, a Spearman correlation test was conducted on both datasets to eliminate the subjectivity score, which exhibited an almost negligible correlation. Subsequently, direct correlation analyses were extended to the compound score and polarity score. Analogous to the Google Trends analysis, this study performed correlation analyses on consecutive data spanning 1 to 6 months. The findings highlight that the highest direct correlation is observed within a single month, reaching up to 0.51, as depicted in Fig 4.2a. Fig 4.2b illustrates a distinct similarity between the trends of tweets sentiment values and the changes in Bitcoin price. The correlation coefficient between Bitcoin's closing price and the tweets compound score dwindles during the initial one to five months, but notably increases when six months of data are included. Hence, a hypothesis suggesting a conceivable correlation between Google Trends and tweets sentiment was proposed and validated over a six-month dataset, revealing a significant negative correlation of -0.45.

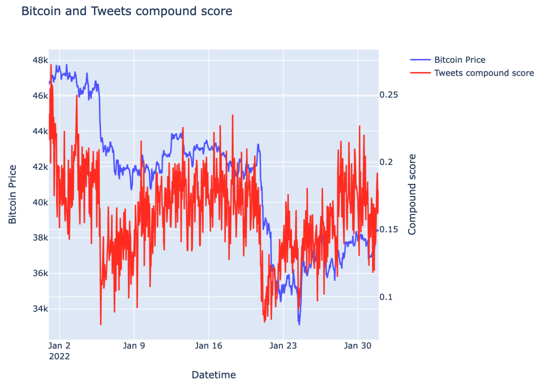
Conversely, when using a three-month data window, outliers emerge, potentially stemming from an excessive number of sentiment-neutral scores for that period. Nonetheless, the overall correlation between the sentiment mean (neutral values excluded) and Bitcoin's price remains lower than the sentiment mean with neutral values retained, implying the necessity of preserving neutral sentiment values. Consequently, this study decided to employ the sentiment mean without eliminating neutral values as a feature.

It is noteworthy that, within a one-month timeframe, the number of tweets exhibits a negative correlation with Bitcoin's price. However, as the data expands to three to four months, the linear correlation loses significance. Nevertheless, this absence of linear correlation does not preclude the existence of a causal relationship between tweets volume and Bitcoin price returns. To validate this, a Granger causality test was applied. Intriguingly, despite the strongest correlation occurring within a single month, there is no bidirectional causality between tweets volume and price returns. Instead, tweets volume exert a lagged effect, whether a 1-lag or spanning 11 to 28 lags, on price returns over consecutive months ranging from 2 to 6 months.

Similarly, in the context of tweets sentiment, despite the robust direct correlation and the potent predictive influence of price returns on tweets sentiment, tweets sentiment itself does not exhibit a causal effect on price returns. Rather, it manifests a lagged effect on the direction of price movement, spanning



(a) Comparison of Sentiment Scores with Neutral Values Retained and Removed in Correlation with Bitcoin Price



(b) Close Price Trends VS. Tweets Compound Score Trends

Figure 4.2: Data Analysis and Selection For Tweets Data

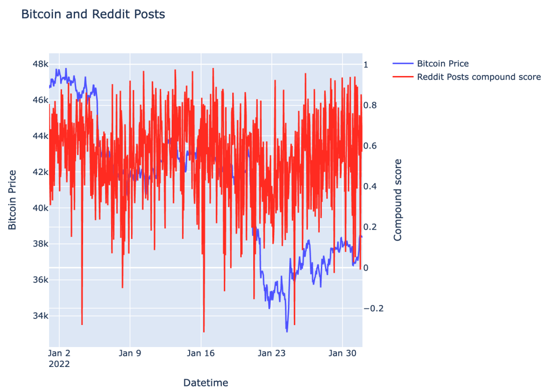
2 to 19 lags within the short term.

In summary, there exists a correlation between tweets sentiment and Bitcoin price, but tweets sentiment lacks the predictive capability for Bitcoin price returns. While a strong linear correlation between tweets counts and Bitcoin price is absent, tweets counts hold predictive power for Bitcoin price returns over a defined period. When assessing correlation and causation for the three distinct features, noteworthy patterns emerged in both short-term and long-term data. Stronger correlation was observed within the one-month short-term data, while it gradually strengthened in the six-month long-term data. Hence, this study will concentrate on the findings within these two time frames.

4.1.3 Reddit Posts

The Section 3.3.2 introduces a methodology inspired by prior approaches for sentiment score adjustment in tweets, utilizing the upvote ratio as a weighting factor to enhance outcomes. Nevertheless, when scrutinizing the Spearman correlation between mean unweighted and weighted sentiment scores and Bitcoin's price (as depicted in the Fig 4.3a), the findings unveil that the correlation scores derived from these two data variations exhibit minimal disparity. Delving deeper, Granger causality tests elucidate the absence of mutual causality between any of the three sentiment scores and Bitcoin price returns across all timeframes.

(a) Comparison of Reddit Posts Sentiment Scores with Weighted and Unweighted Mean Values in Correlation with Bitcoin Price



(b) Close Price Trends VS. Reddit Posts Compound Score Trends

Figure 4.3: Data Analysis and Selection For Reddit Posts Data

However, concerning the prediction of Bitcoin price movement direction, all sentiment scores manifest predictive prowess with a lag of three and beyond, consistently extending to time horizons surpassing three months. Consequently, it can be inferred that the inclusion of the likes ratio as a weight exerts negligible influence on outcomes. Additionally, the linear correlation linking Reddit posts and hourly

Bitcoin price remains modest, emphasizing that the impact of Reddit posts on Bitcoin price alterations primarily stems from long-term influences.

Price movement Prediction		
Months	Social Media Lagged Features	Lags
1	Tweets compound score	2 - 10, 11, 15
6	Google Trends	10 - 19
	Tweets polarity score	4
	Reddit compound score	3
	Reddit polarity score	9, 10
	Reddit subjectivity score	3
Price Returns Prediction		
Months	Social Media Lagged Features	Lags
1	Google Trends	10 - 48
6	Google Trends	2 - 48
	Tweets volume	1, 30, 31, 32

Table 4.1: Lagged Features For the Price Prediction

To sum up, the features and corresponding lag numbers that have predictive power for Bitcoin price returns and price movement direction are shown in Table 4.1. As the correlation judgement and causality tests on the three different social data show prominent features in the short-term data of one month and the long-term data of six months, i.e., the correlation is stronger in the short-term data and picks up in the long-term data. Therefore, this study will focus on the results of these two time horizons. In the feature selection process, for the price prediction problem, Google Trends and Twitter Sentiment can be used as certain predictors due to their strong correlation with the data at hourly intervals respectively. This paper focuses on its impact on Bitcoin price returns and moving direction, and the results show that for the price returns prediction, Google Trends can be selected as one of the features with more than two lags on the six-month-long term data, except for the number of tweets with 1, 30 to 32 lags which may have predictive ability for price returns, and thus selected as one of the features. For price movement direction prediction, on six months long term data, Google Trends influences it with more than 10 lags, tweets polarity score as well as related sentiment score of Reddit posts can have some lag influence on it, so it is also used as one of the features. The Spearman correlation heatmap among features is presented in Fig A.1.

4.2 Phase 1: Hourly Price Prediction

4.2.1 Model Selection

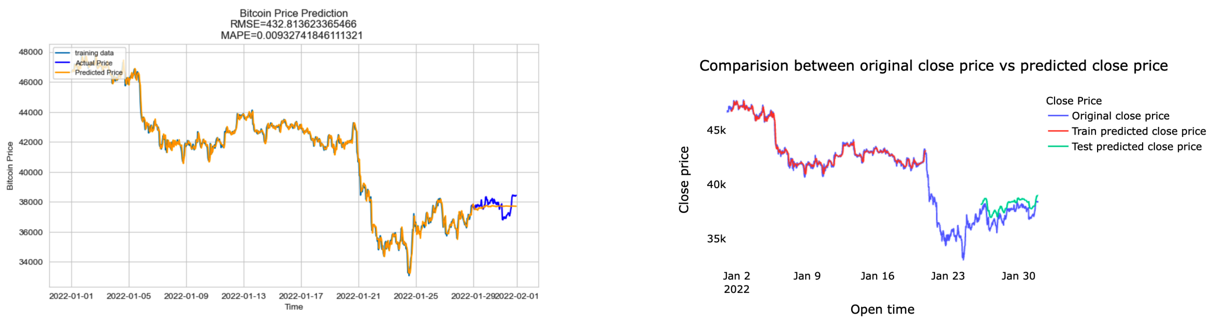
ARIMA

At the beginning of the project, ARIMA served as the baseline model for predicting closing prices. Given the non-smooth nature of closing prices, a first-order difference was applied to achieve smoother data. ACF and PACF plots aided in determining the values of p and q in the ARIMA model, reflecting the correlation between current observations and their lagged past values. For the one-month short-term data, there existed two options for both p and q values: 11 and 24. AIC was utilized to select the optimal model, resulting in ARIMA(11,1,24) due to its lowest AIC value. Employing a similar approach with six-month data, p and q values of 10 and 15 were considered. Through model training across varying lag combinations, the optimal model emerged as ARIMA(10,1,10). Leveraging these two baseline models, solely historical prices were employed as features to predict closing prices for both time horizons.

LSTM

LSTM does not require data smoothing for price regression prediction. In the project's early stage, LSTM was directly employed for closing price prediction. ACF and PACF were used to determine the values of p and q for the timestep selection. For short-term one-month data, timestep was set to 11 or 24, while for long-term six-month data, it was set to 10 or 15. A single-layer LSTM model with 32 neurons and a batch size of 32 was utilized for 5-fold cross-validation on both datasets to identify the optimal timestep.

The results revealed that the short-term data achieved the lowest average RMSE of 376.56 and MAPE of 0.06 with a timestep of 24, while the long-term data achieved the lowest average RMSE of 2724.83 and MAPE of 0.104 at a timestep of 10. Performance comparisons between LSTM and the baseline ARIMA model are presented in Table A.1 and Fig 4.4.



(a) Close Price Prediction Using ARIMA

(b) Close Price Prediction Using LSTM

Figure 4.4: The comparison of ARIMA and LSTM model

LSTM outperforms the baseline ARIMA model in fitting price predictions. ARIMA's prediction outcome presents a linear trend, likely due to its linear nature, which struggles to capture complex features. Conversely, as a nonlinear model, LSTM automatically captures timestep correlations and nonlinear features. For the same timestep, LSTM consistently outperforms the ARIMA model across both time-series datasets. After this preliminary model selection, the LSTM model will be deployed for subsequent price return predictions and assessing the predictive capacity of social media sentiment and Google Trends data.

For price returns prediction, the LSTM model's hyperparameters were chosen via 5-fold cross-validation. These included the number of neurons, batch size, layers, activation function, and optimizer. Neuron counts were tested at 32, 64, and 128, while batch sizes were explored at 30, 64, and 80. The number of layers was assessed at 1, 2, and 3. Activation functions Rectified Linear Unit (ReLU) and Linear were compared, and optimizers Adam and Rmsprop were considered. The optimal LSTM model was selected based on average minimum MAPE and maximum R^2 values. Ultimately, a single-layer LSTM model with 32 neurons, a batch size of 80, ReLU activation, and Adam optimizer was chosen. To prevent overfitting, an Early Stopping callback was incorporated to halt training once validation loss ceased to improve.

4.2.2 Results and Discussion

Price returns predictions were conducted for both 1-month short-term data and 6-month long-term data using the optimal LSTM model and the previously selected features. Five iterations were executed for each feature combination, and the outcomes were evaluated by calculating the average R^2 scores, presented in the provided Table 4.2. Initially, only historical price returns were utilized as input features for the model, and subsequently, features with potential impact were gradually incorporated following the results of the Granger causality test.

Months	Features	Timestep	Mean R2-Score
1	Return	24	-0.04579
	Return	11	-0.02826
	Return, Google Trends (lag 15)	11	-0.0177
6	Return	10	-0.0695
	Return	15	-0.0034
	Return, Google Trends (lag 2)	15	**0.0065**
	Return, Tweet Volume (lag 1)	15	-0.0043
	Return, Google Trends (lag 2) Tweet Volume (lag 30)	15	-0.0015

Table 4.2: Performance Evaluation of Different Feature Combinations

Given the substantial volatility of Bitcoin prices and the generally modest return values, relying solely on LSTM and price return values doesn't yield favourable predictions. Nevertheless, it's worth highlighting that introducing features with lagged values can moderately enhance the R^2 score.

For short-term data, Google Trends features were examined by appending 10, 15, and 25 lags to the original dataset. The findings underscored that the optimal performance was achieved with a 15-lag configuration. Turning to long-term data, the addition of Google Trends features with 2 lags and the inclusion of tweets volume with a 1-lag interval concluded that lagged Google Trends data notably enhances price returns prediction. This observation holds true when considering Google Trends features with other lag values. However, the influence of tweets volume on the predictive ability of price returns remained insubstantial, possibly even yielding counterproductive outcomes. In cases where tweets volume and Google Trends features were combined, a slight improvement in predictive ability was observed only when 30-lag tweets volume were paired with the 2-lag Google Trends feature. However, this combined effect fell short of the predictive performance achieved by exclusively utilizing the value of the Google Trends feature. This phenomenon could be attributed to the introduction of tweets volume, weakening the predictive efficacy.

In conclusion, Google Trends data does exhibit a degree of capability in predicting price returns. In terms of long-term impact, the lagged influence of Google Trends on price returns is concise, typically manifesting after approximately 2 hours. Nevertheless, the influence of tweets volume on price returns remains negligible in both short and long-term contexts.

4.3 Phase 2: Hourly Price Movement Prediction

4.3.1 Model Selection

Random Forest Classifier

A random forest classifier serves as the foundational model for predicting price directions, assessed over a six-month span of extensive historical data. The finest hyperparameters were ascertained via a comprehensive grid search and fine-tuned using a 5-fold cross-validation technique. This process encompassed critical parameters such as the count of decision trees ("n_estimators") and the maximal depth allotted to each individual decision tree ("max_depth"). Ultimately, the optimal setting for the maximum depth of decision trees was identified at 10, accompanied by a decision tree count of 300. These refined parameters were then employed as the bedrock for the baseline model.

Deep Learning-based Model

To select the most effective model for predicting the direction of price movement, three distinct deep learning models were compared using six months of long-term data. Similarly, hyperparameters were fine-tuned for both the CNN and LSTM models, encompassing variations in the number of neurons (16, 32, 64, 128) and batch sizes (30, 64, 80, 200). Ultimately, both the CNN and LSTM architectures were configured as single layers, featuring 32 neurons per layer and a batch size of 200. ReLU served as the activation function for the CNN and LSTM layers, while the fully-connected layer adopted the Linear activation function. The structure of the hybrid CNN-LSTM model was delineated in the Section 3.7.2, revealing optimal performance with a batch size of 80.

Features	Model	F1-score: Down	F1-score: Up	Accuracy
Label, Close Price	Random Forest Classifier	0.5232	0.5243	0.5238
	CNN	0.5327	0.5497	0.5413
	LSTM	0.5665	0.5367	**0.5524**
	CNN-LSTM	0.3814	0.6158	0.5524

Table 4.3: Mean F1-Score and Accuracy of Different Models with Timestep 15

Initial classification forecasts were executed through CNN, LSTM, and CNN-LSTM hybrid models, using historical price trends and closing prices as exclusive features over six months of data. Each of the three models underwent five rounds of assessment at 10 and 15 steps, with the averaged outcomes demonstrating the prevalent superiority of the LSTM's accuracy compared to the other two models. For instance, considering 15 steps, the results, as depicted in the Table 4.3, displayed heightened predictive capability across all three models in contrast to the Random Forest classifier. While CNN-LSTM approached LSTM in terms of accuracy, a conspicuous imbalance was evident in its classification performance over 15 steps. This observation might be attributed to the intricacies of the CNN-LSTM hybrid model's architecture, demanding a more robust data and feature support. This further highlights that complexity in models doesn't necessarily correlate with enhanced classification outcomes; rather, it necessitates substantial underpinning from technical metrics. Consequently, the ultimate decision was to exclusively adopt the LSTM model for subsequent classification predictions.

4.3.2 Results and Discussion

When considering short-term data, only the tweets compound score demonstrates a lagged effect on price direction, as indicated by the Granger causality test. As a result, for predicting price direction over a one-month period of short-term data, only the initial features along with the lagged tweets sentiment score are chosen as predictive elements. The initial features encompass historical price and price movement direction (Label). By incorporating tweets sentiment scores with varying lag counts as features and determining the optimal lag count, as depicted in the table, the utilization of 10 lags of tweets sentiment scores enhances short-term classification prediction accuracy by nearly 5%. Other lagged tweets sentiment scores also enhance model performance to differing degrees. This underscores the substantial predictive capacity of tweets sentiment for price direction in short-term data, with its impact potentially manifesting in the direction of Bitcoin price movement within a span of 2 hours to half a day.

Time range: 1 month				
Timestep	Features	F1-score: Down	Up	Accuracy
24	['Label' 'Close']	0.4568	0.5654	0.5179
11	['Label' 'Close']	0.6202	0.4618	0.5560
11	['Label' 'Close' 'tw_score_lag'](Lag 3)	0.6358	0.5212	0.5886
11	['Label' 'Close' 'tw_score_lag'](Lag 15)	0.6182	0.5792	0.6000
11	['Label' 'Close' 'tw_score_lag'](Lag 10)	0.6387	0.5547	**0.6014**
Time range: 6 months				
Timestep	Features	F1-score: Down	Up	Accuracy
15	['Label' 'Close']	0.5665	0.5367	0.5523
10	['Label' 'Close']	0.5445	0.5723	0.5589
10	['Label' 'Close' 'google_trends_lag']	0.5504	0.5708	0.5610
10	['Label' 'Close' 'tw_polarity_lag']	0.5469	0.5735	0.5607
10	['Label' 'Close' 'google_trends_lag' 'tw_polarity_lag']	0.5599	0.5687	**0.5645**
10	['Label' 'Close' 're_co_lag' 're_po_lag' 're_su_lag']	0.5441	0.5743	0.5603
10	['Label' 'Close' 're_co_lag' 're_po_lag' 're_su_lag' 'tw_polarity_lag']	0.5441	0.5742	0.5598
10	['Label' 'Close' 're_co_lag' 're_po_lag' 're_su_lag' 'google_trends_lag']	0.558	0.5632	0.5612
10	['Label' 'Close' 're_co_lag' 're_po_lag' 're_su_lag' 'google_trends_lag' 'tw_polarity_lag']	0.5519	0.5642	0.5591

Table 4.4: Evaluation for Different Features: 'tw_score_lag' represents lagged tweets compound score. 'google_trends_lag' represents 10 lagged google trends value, 'tw_polarity_lag' represents 4 lagged tweets polarity score, and 're_co_lag', 're_po_lag', 're_su_lag' represent the 3 lagged Reddit posts compound score, 10 lagged polarity score, 3 lagged subjectivity score, respectively

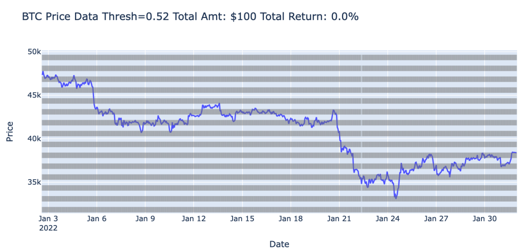
For long-term data, features that might influence price direction prediction encompass Google Trends data, Twitter sentiment scores, and sentiment scores from Reddit posts. Following validation of the lagged effects of these features, the most suitable lag count is selected for Google Trends data. The results reveal that a Google Trends model with 10 lags attains optimal performance, thereby elevating model accuracy to a certain degree. Concurrently, sentiment scores from Reddit posts are juxtaposed with Google Trends data and Twitter sentiment scores. The findings presented in the table illustrate that the inclusion of either public sentiment or Google Trends data enhances predictive capability to a certain extent compared to the original features. Concerning the impact of individual features, the lagged influence of Google Trends data holds more significance, contributing to an approximate 0.3%

increase in average accuracy. However, coupling Google Trends data with either Twitter sentiment scores or sentiment scores from Reddit posts yields even greater accuracy improvements. Among these, the combination with Twitter sentiment scores yields the highest accuracy, boosting overall accuracy by around 0.6%. Nevertheless, combining features from three distinct sources leads to a decline in accuracy. This phenomenon might be attributed to intricate interactions among the features.

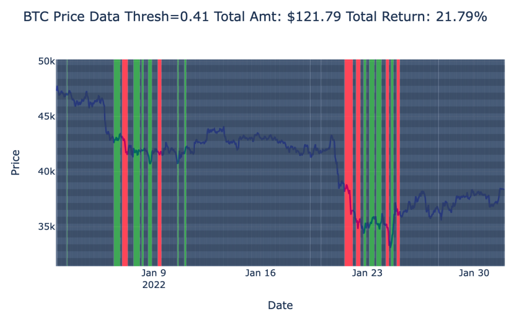
In summary, it can be deduced that in the context of long-term data, Google Trends data also influences price direction after approximately 10 hours, which is similar to findings from earlier price return forecasts. The combination of Google Trends data with public sentiment fosters heightened accuracy. Features stemming from these three distinct sources exert varying degrees of influence on price direction, with Reddit posts yielding the weakest influence. Analysis suggests this could be attributed to the relatively scant volume of Reddit posts, contrasting with platforms like Twitter where posts are often extensive, and users lack the same immediacy to express sentiment. Additionally, the limited volume of data might introduce noise when addressing data gaps. Correlation analyses also underscore a relatively weak relationship between sentiment scores from Reddit and Bitcoin price. The structure of the best models obtained through feature selection can be seen in Fig A.4.

4.4 Phase 3: Backtesting for Action Recommendation Strategy

In Phase 2, the optimal LSTM model trained was utilized to develop a Bitcoin trading behaviour recommendation model. Notably, this phase revealed significant insights into the predictive capabilities concerning short-term (one month) and long-term data (six months). Specifically, when employing short-term data for model construction, the compound score from tweets demonstrated a notable enhancement in predicting price movement directions, achieving a remarkable accuracy of up to 63%. Similarly, with the utilization of long-term data for model development, a combination of Google Trends and tweets polarity score effectively contributed to heightened prediction accuracy. Consequently, two separate models were established for short-term and long-term action recommendations, presenting a comprehensive strategy for maximizing profits during specific timeframes. The proposed strategies have been outlined in Section 4.4, and a simulator was employed for backtesting the action recommendations.



(a) The baseline model: Only include basic features - Bitcoin historical close price and price movement direction

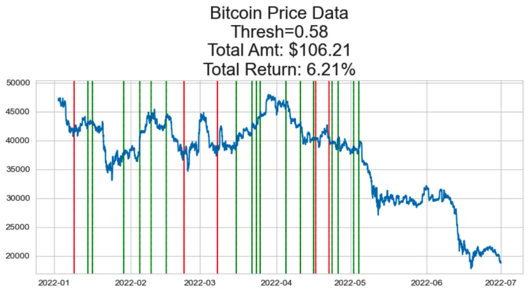


(b) The optimized model: Basic features+ Tweets compound score

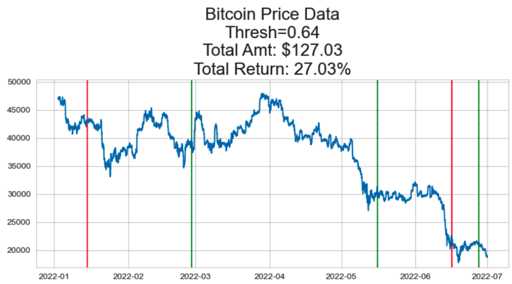
Figure 4.5: The trading simulation of action recommendation model for short-term data (1 month)

It's worth mentioning that due to the probabilistic nature of the price direction predictions generated by the model, a threshold of 0.5 is typically employed. When the probability of price increase surpasses this threshold, a "Buy" recommendation is made. By adjusting the threshold, the percentage of price returns can be obtained. Backtesting was conducted on short-term data spanning a month, and by iteratively adjusting the threshold, the optimal threshold for achieving maximum price returns was identified. As demonstrated in the Fig 4.5b, the red window denotes the instance of purchasing Bitcoin, followed by a subsequent decline in its price. On the other hand, the green window signifies the correct decision of purchasing Bitcoin when its price subsequently increased. The range of threshold testing was between 0.35 and 0.65. Starting with an initial threshold of 0.5, the thresholds of 0.35 and 0.65 were tested for their corresponding price return rates. The threshold that yielded an increase in return rates compared to the initial threshold was selected. Further testing was conducted within the range between this selected threshold and the initial threshold, narrowing down the threshold incrementally to identify the optimal threshold. Ultimately, a threshold of 0.41 was determined to yield the maximum return. As depicted in the Fig 4.5b, assuming an initial investment of $100 in Bitcoin at the first hour of the month, a total return rate of 21.79% was achieved after a month of simulated trading, resulting in $121.79. The prevalence of green windows in the graph demonstrates the accuracy of the decision-making process.

To establish a comparison with the LSTM model trained on the original features, excluding social media sentiment, an additional trading behaviour recommendation model was devised. This model was optimized by fine-tuning the threshold to attain the maximum return rate. As depicted in the Fig 4.5a, the optimal threshold emerged at 0.52 or higher, leading to a return rate of zero. This signifies that following the initial Bitcoin purchase in the first hour of the month, subsequent actions remained in a "Wait" state throughout the month. At a threshold of 0.51, a loss of 0.67% was observed (See Fig A.2), indicating that the output probabilities of the classification model primarily clustered within the 0.48 to 0.52 range. This range highlighted challenges in distinguishing between upward and downward price trends. However, the model enriched with supplementary features showcased an augmentation in the probabilities of both categories, yielding subsequently amplified returns.



(a) The baseline model: Only include basic features - Bitcoin historical close price and price movement direction



(b) The optimized model: Basic features+ Google Trends and Tweets polarity score

Figure 4.6: The trading simulation of action recommendation for long-term data (6 months)

A parallel strategy was employed in constructing a long-term action recommendation model and executing testing. In the end, the optimized model, integrating social media data, achieved a remarkable maximum price return rate of 27.03% after six months of simulated trading, utilizing a threshold of 0.64. By comparison, the baseline model, unoptimized with social media data, exhibited an optimal threshold of 0.58, yielding a maximum return rate of merely 6.21%. As illustrated in the Fig 4.6, the optimized model demonstrated superior returns over the extended term. Notably, it maintained accurate decision-making even during pronounced price declines, while the baseline model struggled to discern price trends in the later stages. This validation further accentuated the effectiveness of optimizing the Google Trends and tweets sentiment score recommendation model. Other profits obtained at various threshold values are shown in the Appendix A.3.

4.5 Summary

This chapter provides a comprehensive implementation and evaluation of the study, focusing on social media sentiment, Google Trends data, and historical Bitcoin hourly price data from public sources. The data was split into short-term (one month) and long-term (six months) segments for detailed examination. Google Trends data showed lagged impact on price returns in both periods, with the effect emerging approximately two time units after the initial data point in long-term data. Short-term tweets sentiment scores significantly influenced predictions, raising accuracy by 3% on average, sometimes up to 8%. However, this impact was short-lived (2 to 15 hours) due to social media's real-time nature. In contrast, Google Trends data displayed longer-lasting influence. Despite training on long-term data, accuracy gains remained limited due to Bitcoin's extreme volatility, notably between January and June 2022.

Utilizing the optimal LSTM model trained on short-term and long-term data, the action recommendation model was created and backtested via simulation. Optimal thresholds, determined at 0.41 and 0.64, maximized short-term and long-term trade returns, respectively. Furthermore, these results further validate the predictive capabilities of Google Trends and Tweets sentiment scores in determining the direction of Bitcoin price fluctuations.

Chapter 5

Conclusion

5.1 Contributions and Achievements

This study gathers public sentiment and interest data from various social media platforms, encompassing metrics like tweets frequency from Twitter, the Google Trends index, and Reddit posts. It delves into the influence of public sentiment on hourly Bitcoin price movements across these platforms, examining both short-term (one month) and long-term (six months) impacts. Correlations between each media type and hourly Bitcoin price movements were investigated, and relevant features were selected through statistical methods.

These correlation-derived features were then utilized to construct models predicting Bitcoin price returns and movement direction, considering both short and long-term data. This step further validated the lagged impact and predictive capacity of diverse sources of public sentiment and interest on Bitcoin's price behaviour. The findings reveal that Google data exhibits a certain lagged impact on Bitcoin's price returns in both short and long-term scenarios, while displaying predictive power primarily for price movement direction in the long term.

On the contrary, tweets sentiment demonstrates significant predictive power for short-term price movement direction, enhancing accuracy by 3% to 8%. However, this sentiment's impact wanes in the long term, though in conjunction with Google Trends as a feature set, it effectively boosts average accuracy by 0.6%. The tweets count doesn't exhibit substantial connections with price returns or movements in either the short or long term. Conversely, Reddit posts indicate minimal influence on price movement direction in long-term trends, providing marginal improvement in price return and movement prediction.

The study identifies the optimal historical timestep for predicting the next hour's Bitcoin price to be between 10 and 24 hours. Google Trends' lagged impact is observed after 2 or 10 hours, while the short-term lagged impact of tweets sentiment surfaces after 10 hours. Ultimately, the study constructs a trading action recommendation model utilizing the optimized price movement prediction model, enhanced by tweets sentiment scores and Google Trends. This adjustment significantly improves maximum price returns compared to using the original feature set, post-threshold optimization.

In summary, the key contributions of this study encompass:

- Exploring correlations between social media data and Bitcoin price movement, elucidating lagged impact and causality between public sentiments, interests, and price trends.
- Focused on two types of Bitcoin price prediction problems: regression prediction of price returns and classification prediction of price movement direction.
- Comparing the predictive capabilities of diverse models for Bitcoin prices, highlighting the efficacy of single-layer LSTM models over complex hybrid models for short-term forecasting.
- Evaluating the potential of features from different sources for predicting short and long-term Bitcoin prices.
- Crafting a trade action recommendation model based on the optimized framework, simulating a month's worth of Bitcoin trading to propose actionable recommendations like "Buy", "Sell", or "Wait".

5.2 Project Status

The initial objectives and completion of this study are outlined below:

- **Collect Bitcoin-related information from multiple social media platforms, exploring the relationship between public sentiment and Bitcoin price fluctuations and the time lagged impact using statistical techniques and sentiment analysis.**
  - Because of limitations in accessing social media, this study ultimately used historical datasets that are available to the public instead of collecting data directly from online sources. The data collection methodology and platform have been detailed in 3.2, the statistical techniques employed have been explained in 3.4, and the process of conducting correlation and causation tests has been executed and described in 4.1.
- **Select the optimal model to predict price fluctuations and price returns using deep learning models.**
  - The LSTM model is finally chosen for this study, a description of the model has been shown in 3.7 and its prediction and evaluation are elaborated in 4.2 and 4.3.
- **Select related public sentiment as features to optimize model performance and enhance prediction accuracy.**
  - Across various time periods and distinct price prediction tasks, the study has identified diverse optimal feature combinations. The outcomes and discussions pertaining to these findings are expounded upon in 4.2 and 4.3.
- **Construct an action recommendation model using the optimal prediction model, adjusting thresholds to maximize profits within specific time intervals.**
  - The application plan of the model is detailed in 3.9, and its execution, evaluation, and comparison are covered in 4.4.

5.3 Further Work

As mentioned in 5.2, due to limitations in accessing data, this study focused solely on media data from Google Trends, Twitter, and Reddit concerning Bitcoin's price. However, it's important to note that cryptocurrencies encompass more than just Bitcoin, including well-known options like Ether and Litecoin. As a result, future research could consider expanding the dataset to cover a broader array of sources and a longer timeframe to gain insights into more extended trends.

Furthermore, various models were employed for analysis. Although the performance evaluation favoured LSTM-based models, there's room for exploration of alternative models. For instance, possibilities include experimenting with augmented learning techniques or investigating the potential of GRU models. Additionally, hybrid models could be considered to leverage their combined strengths.

In terms of generating trading action recommendations, further testing with diverse datasets remains necessary. In addition, the current approach involves manually determining the threshold, continuously validating different thresholds within a certain range to find the optimal one. Typically, this range falls between 0.35 and 0.65. However, this method is time-consuming. Therefore, further research is needed to explore alternative methods for threshold determination. Developing software that facilitates user interaction and manipulation through simulation tools could enhance practical applications.

Bibliography

[1] Bipin Aasi, Syeda Aniqa Imtiaz, Hamzah Arif Qadeer, Magdalean Singarajah, and Rasha Kashef. Stock price prediction using a multivariate multistep lstm: A sentiment and public engagement analysis model. pages 1–8, 2021.

[2] Jethin Abraham, Daniel Higdon, John Nelson, and Juan Lbarra. Cryptocurrency price prediction using tweet volumes and sentiment analysis. 1, 2018.

[3] Gourang Aggarwal, Vimal Patel, Gaurav Varshney, and Kimberly Oostman. Understanding social factors affecting the cryptocurrency market. Jan 2019.

[4] Erdinc Akyildirim, Ahmet Goncu, and Ahmet Sensoy. Prediction of cryptocurrency returns using machine learning. *Ann Oper Res*, 297:3–36, 2021.

[5] L. Alessandretti, A. ElBahrawy, L. M. Aiello, and A. Baronchelli. Machine learning the cryptocurrency market. 2018.

[6] Saúl Alonso-Monsalve, Andrés L. Suárez-Cetrulo, Alejandro Cervantes, and David Quintana. Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. *Expert Systems with Applications*, 149, 2020.

[7] Argimiro Arratia and Albert X. López-Barrantes. Do google trends forecast bitcoins? stylized facts and statistical evidence. *J BANK FINANC TECHNOL*, 2021.

[8] Gupta Aryan. Spearman's rank correlation: The definitive guide to understand, Feb 2023.

[9] Naila Aslam, Furqan Rustam, Ernesto Lee, Patrick Bernard Washington, and Imran Ashraf. Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model. 10:39313–39324, 2022.

[10] Amin Azari. Bitcoin price prediction: An arima approach. 2019.

[11] An-Sing Chen, Mark T Leung, and Hazem Daouk. Application of neural networks to an emerging financial market: forecasting and trading the taiwan stock index. 30:901–923, 2003.

[12] Zheshi Chen, Chunhong Li, and Wenjun Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, page 112395, 2020.

[13] Jacques Vella Critien, Albert Gatt, and Joshua Ellul. Bitcoin price change and trend prediction through twitter sentiment and data volume. *Financ Innov*, 8:45, 2022.

[14] Estela Bee Dagum. The x-ii-arima seasonal adjustment method. 1980.

[15] Paul Dolan and Richard Edlin. Is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis? 21:827–843, 2002.

[16] Anne Haubo Dyhrberg. Hedging capabilities of bitcoin. is it the virtual gold? 16:139–144, 2016.

[17] Johansson ErikJJ. Creating daily search volume data from weekly and daily data, Dec 2014.

[18] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[19] Fan Fang, Carmine Ventre, Michail Basios, Leslie Kanthan, David Martinez-Rego, Fan Wu, and Lingbo Li. Cryptocurrency trading: a comprehensive survey. *Financial Innovation*, 8:13, 2022.

[20] Ifigeneia Georgoula, Demitrios Pournarakis, Christos Bilanakos, Dionisios N. Sotiropoulos, and George M. Giaglis. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. *SSRN Journal*, 2015.

[21] Bremmer Gerlof. Predicting tomorrow's cryptocurrency price using a lstm model, historical prices and reddit comments. Master's thesis, Tilburg University—, 2018.

[22] Malato Gianluca. An introduction to the shapiro-wilk test for normality, May 2023.

[23] Gül Cihan Habek, Mansur Alp Toçoğlu, and Aytuğ Onan. Bi-directional cnn-rnn architecture with group-wise enhancement and attention mechanisms for cryptocurrency sentiment analysis. *Applied Artificial Intelligence*, 36:2145641, 2022.

[24] Dolatzadeh Hirad. 01-tweet cleaning, 2023.

[25] Melody Y. Huang, Randall R. Rojas, and Patrick D. Convery. Forecasting stock market movements using google trend searches. *Empir Econ*, 59:2821–2839, 2020.

[26] Duygu Ider and Stefan Lessmann. Forecasting cryptocurrency returns from sentiment signals: An analysis of bert classifiers and weak supervision, 2023.

[27] Patrick Jaquart, David Dann, and Christof Weinhardt. Short-term bitcoin market prediction via machine learning. *The Journal of Finance and Data Science*, 7:45–66, 2021.

[28] Suhwan Ji, Jongmin Kim, and Hyeonseung Im. A comparative study of bitcoin price prediction using deep learning. 7:898, 2019.

[29] Frost Jim. Spearman's correlation explained, 2021.

[30] Guy Kaplanski and Haim Levy. Sentiment and stock prices: the case of aviation disasters. 95:174–201, 2010.

[31] Kwansoo Kim and Sang-Yong Tom Lee. The dynamics of cryptocurrency market behavior: sentiment analysis using markov chains. pages 365–395, 2021.

[32] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, TTae Hyeong Kim, Shin Jin Kang, and Change Hun Kim. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLOS ONE*, 11:e0161197, Aug 2016.

[33] Young Bin Kim, Jurim Lee, Nuri Park, Jaegul Choo, Jong-Hyun Kim, and Change Hun Kim. When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PLOS ONE*, 12:e0177630, 2017.

[34] Olivier Kraaijeveld and Johannes De Smedt. The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, 65:101188, Mar 2020.

[35] Ladislav Kristoufek. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific Reports*, 3:3415, Dec 2013.

[36] Ladislav Kristoufek. Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. 3, 2013.

[37] Man-Fai Leung, Lewis Chan, Wai-Chak Hung, Siu-Fung Tsoi, Chun-Hin Lam, and Yiu-Hang Cheng. An intelligent system for trading signal of cryptocurrency based on market tweets sentiments. 2:153–169, 2023.

[38] Yan Li and Wei Dai. Bitcoin price forecasting method based on cnn-lstm hybrid neural network model. pages 344–347, 2020.

[39] Bing Liu and Lei Zhang. *A Survey of Opinion Mining and Sentiment Analysis*. Springer US, 2012.

[40] Ekaterina Loginova, Wai Kit Tsang, Guus van Heijningen, Louis-Philippe Kerkhove, and Dries F. Benoit. Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data. *Machine Learning*, 2021.

[41] Benoit B. Mandelbrot. When can price be arbitraged efficiently? a limit to the validity of the random walk and martingale models. *The Review of Economics and Statistics*, 53(3):225–236, 1971.

[42] Mishra Mayank. Convolutional neural networks, explained, Aug 2020.

[43] Sean McNally, Jason Roche, and Simon Caton. Predicting the price of bitcoin using machine learning. pages 339–343, 2023.

[44] Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8:131662–131682, 2020.

[45] Mohammed Mudassir, Shada Bennbaia, Devrim Unal, and Mohammad Hammoudeh. Time-series forecasting of bitcoin prices using high-dimensional features: a machine learning approach. *Neural Comput Applic*, 2020.

[46] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Cryptography Mailing list at https://metzdowd.com*, 03 2009.

[47] Sergey Nasekin and Cathy Yi-Hsuan Chen. Deep learning-based cryptocurrency sentiment construction. *Digital Finance*, 2:39–67, 2020.

[48] David M. Q. Nelson, Adriano C. M. Pereira, and Renato A. de Oliveira. Stock market's price movement prediction with lstm neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1419–1426, 2017.

[49] Hotz Nick. What is crisp dm?, Jan 2023.

[50] Nilsson and Wikingsson Andersson. Using sentiment analysis on reddit to predict stock returns. *Attitydanalys på Reddit för att förutspå aktie avkastning*, 2022.

[51] Padav Pallavi. Granger causality in time series – explained using chicken and egg problem, Aug 2021.

[52] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. 2(1–2), 2008.

[53] Yan Pang, Ganeshkumar Sundararaj, and Jiewen Ren. Cryptocurrency price prediction using time series and social sentiment data. pages 35–41, Dec 2019.

[54] Galen Panger. Emotion in social media. 2017.

[55] Raj Parekh, Nisarg P. Ratel, Nihar Thakkar, Rajesh Gupta, and Sudeep Tanwar. Dl-guess: Deep learning and sentiment analysis-based cryptocurrency price prediction. 10:35398–35409, 2022.

[56] Han Woo Park and Youngjoo Lee. How are twitter activities related to top cryptocurrencies' performance? evidence from social media network and sentiment analysis. 28:435–460, 2019.

[57] Jaehyun Park and Yeong-Seok Seo. A deep learning-based action recommendation model for cryptocurrency profit maximization. *Electronics*, 11, 2022.

[58] Jaehyun Park and Yeong-Seok Seo. Twitter sentiment analysis-based adjustment of cryptocurrency action recommendation model for profit maximization. 11:44828–44841, 2023.

[59] V.A. Profillidis and G.N. Botzoris. *Chapter 7 - econometric, gravity, and the 4-Step methods*. Elsevier, 2019.

[60] Bevans Rebecca. Akaike information criterion — when how to use it (example), Mar 2020.

[61] Kundu Rohit. F1 score in machine learning: Intro calculation, Dec 2022.

[62] Tiran Rothman and Chen Yakar. Empirical analysis towards the effect of social media on cryptocurrency price and volume. 15:52–52, 2019.

[63] Thaloengpattarakoon Sanboon, Kamol Keatruangkamala, and Saichon Jaiyen. A deep learning model for predicting buy and sell recommendations in stock exchange of thailand using long short-term memory. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pages 757–760, 2019.

[64] Emre Sasmaz and F. Tek. Tweet sentiment analysis for cryptocurrencies. pages 613–618, 09 2021.

[65] Maximilian Schulte, Mathias Eggert, Frederik Ahlemann, Reinhard Schütte, and Stefan Stieglitz. Predicting hourly bitcoin prices based on long short-term memory neural networks. pages 754–769, 2021.

[66] George Selgin. Synthetic commodity money. 17:92–99, 2015.

[67] Prabhakaran Selva. Augmented dickey fuller test (adf test) – must read guide, 2022.

[68] Nico Smuts. What drives cryptocurrency prices?: An investigation of google trends and telegram sentiment. *SIGMETRICS Perform. Eval. Rev.*, 46:131–134, 2019.

[69] Naman Srivastava, Omkar Gowda, Shreyas Bulbule, Siddharth Bhandari, and Animesh Chaturvedi. Bitcoin evolution analytics: Twitter sentiments to predict price change as bearish or bullish. pages 3547–3556, 2022.

[70] Andrew Urquhart. The inefficiency of bitcoin. 148:80–82, 2016.

[71] Franco Valencia, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. 21:589, June 2019.

[72] Huiwen Wang, Shan Lu, and Jichang Zhao. Aggregating multiple types of complex data in stock market prediction: A model-independent framework. 164:193–204, 2019.

[73] Hinman William. Digital asset transactions: When howey met gary (plastic), June 2018.

[74] I Made Wirawan, Triyanna Widiyaningtyas, and Muchammad Maulana Hasan. Short term prediction on bitcoin price using arima method. pages 260–265, 2019.

[75] Krzysztof Wołk. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. 37:e12493, 2020.

[76] Zi Ye, Yinxu Wu, Hui Chen, Yi Pan, and Qingshan Jiang. A stacking ensemble deep learning model for bitcoin price prediction using twitter comments on bitcoin. 10:1307, 2022.

[77] David Yermack. *Chapter 2 - Is Bitcoin a Real Currency? An Economic Appraisal.* Academic Press, 2015.

[78] Verma Yugesh. Quick way to find p, d and q values for arima, May 2022.

[79] Zach. Rmse vs. r-squared: Which metric should you use?, June 2021.

Appendix A

Appendix

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{A.1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f(\vec{x_i}))^2} \tag{A.2}$$

where N is the total number of data points. y_i is the i-th actual value and $f(\vec{x_i})$ is its corresponding predicted value.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(\vec{x_i}))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{A.3}$$

where $\bar{y}$ is the mean of the target values y_i.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - f(\vec{x_i})|}{|y_i|} \tag{A.4}$$

The F1 score is calculated by combining precision and recall using the Equation A.5 [61].

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{A.5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{A.6}$$

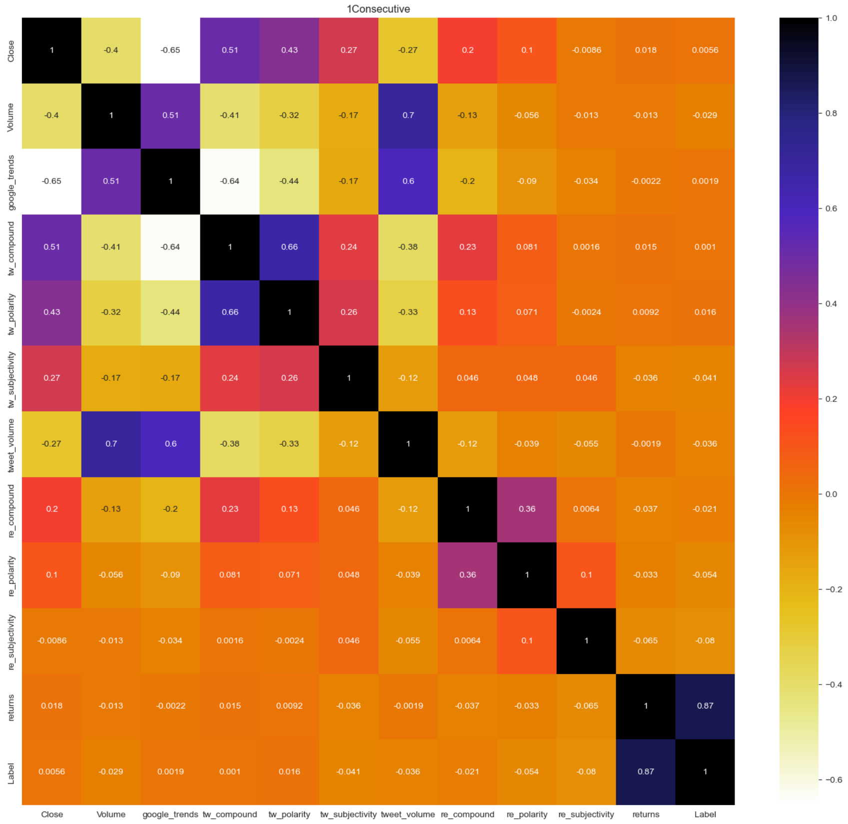
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{A.7}$$

Another frequently utilized evaluation metric is accuracy [45]. It is computed using the Equation A.8:

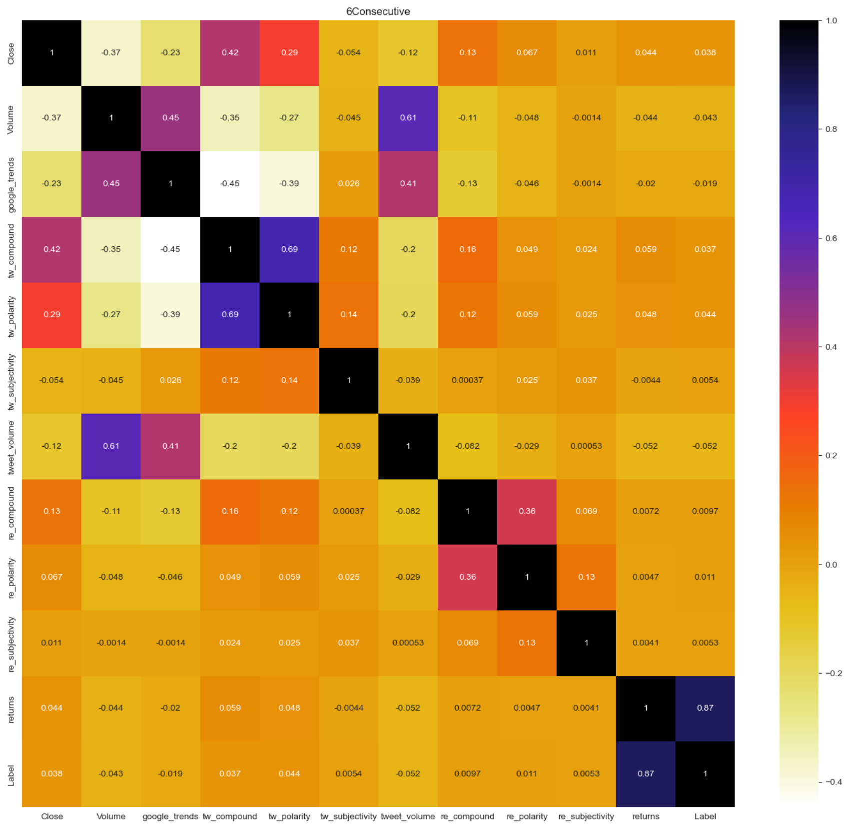
$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{A.8}$$

Model	Month	RMSE	MAPE
ARIMA			
ARIMA(11,1,24)	1	432.81	0.009
ARIMA(10,1,10)	6	6324.78	0.226
LSTM			
Timestep: 24	1	376.56	0.006
Timestep: 10	6	2724.83	0.104

Table A.1: Performance Comparison of ARIMA and LSTM Models



(a) 1 consecutive month



(b) 6 consective months

Figure A.1: The Spearman Correlation Heatmap for the Features

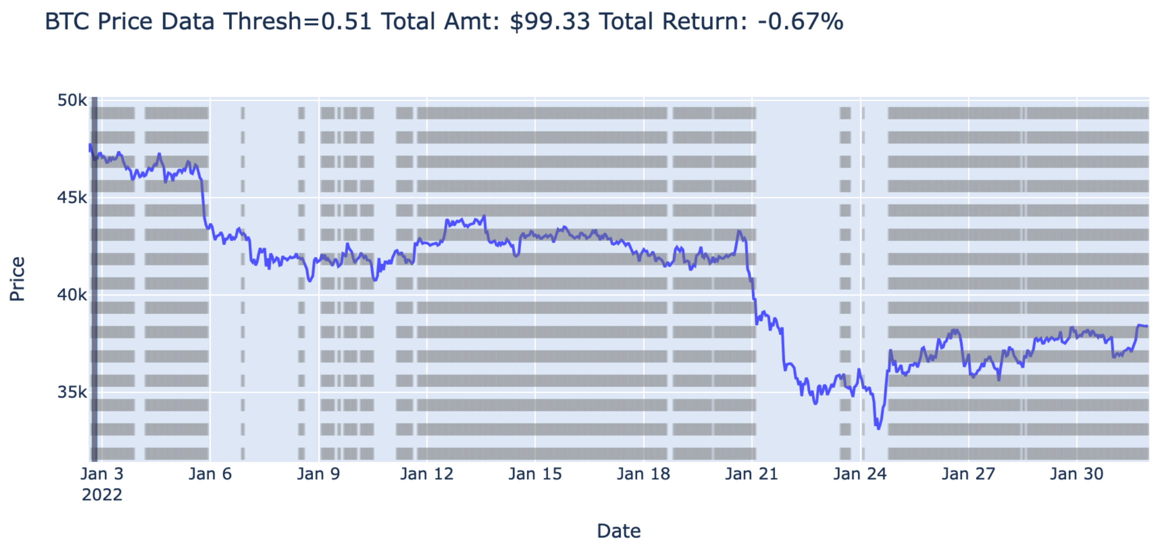
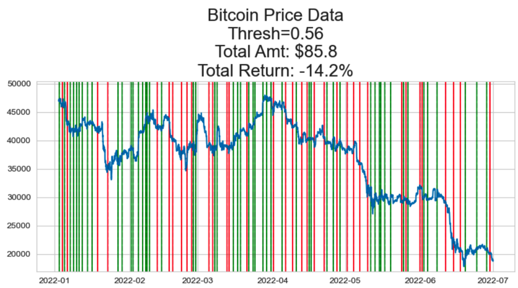
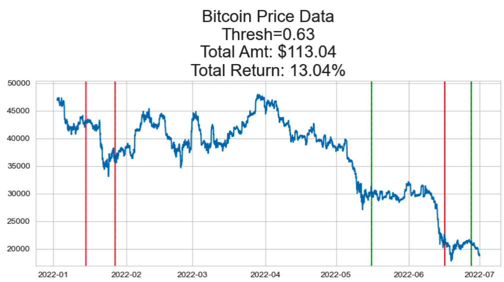


Figure A.2: The trading simulation when threshold is 0.51 for short-term data

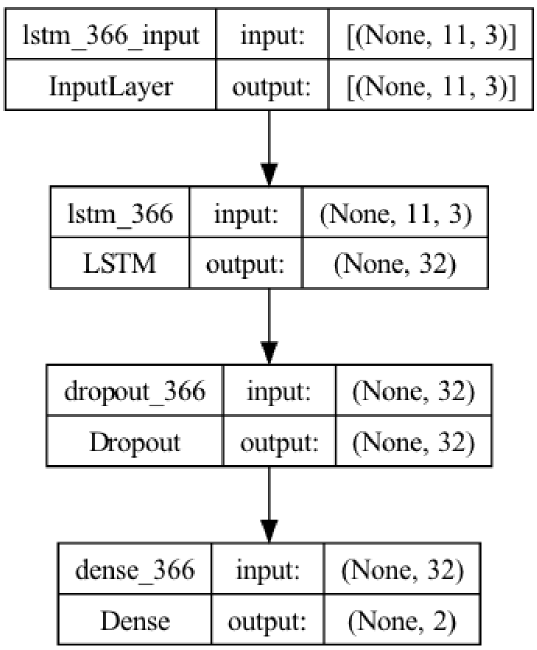


(a) Other thresholds for baseline model

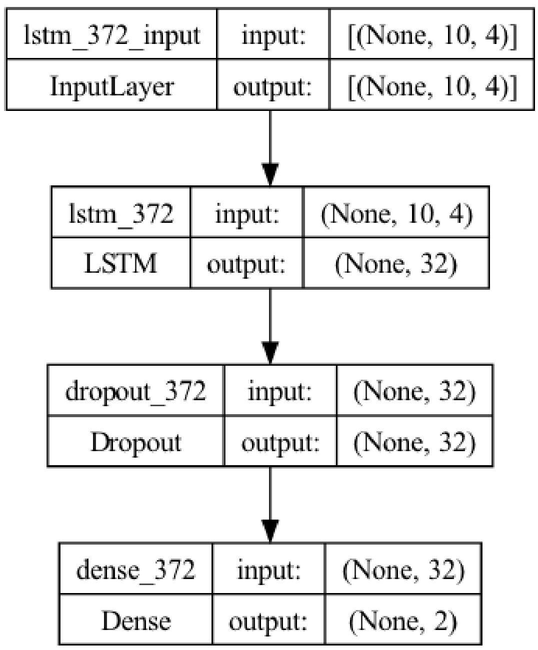


(b) Other thresholds for optimized model

Figure A.3: Other results for long-term simulation



(a) For short-term data (only 1 month)



(b) For long-term data (6 consecutive months)

Figure A.4: The structure of the best models