Unveiling Underlying Patterns, Drivers and Anomalies in Cryptocurrency Price Dynamics through Feature Fusion of Financial Indicators and Sentiment Fluctuations

by

Md. Nafis Rabbi
22366046

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2024

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Nafis Rabbi
22366046

Approval

The thesis titled "Unveiling Underlying Patterns, Drivers, and Anomalies in Cryptocurrency Price Dynamics through Feature Fusion of Financial Indicators and Sentiment Fluctuations"

Submitted by:

Md. Nafis Rabbi (22366046)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on June 11, 2024.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Examiner:
(External)

Dr. Shamim H Ripon
Professor
Department of Computer Science and Engineering
East West University

Examiner:
(Internal)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

____

Dr. Md Sadek Ferdous
Associate Professor
Department of Computer Science and Engineering
BRAC University

Chairperson:
(Chair)

____

Sadia Hamid Kazi, Ph.D.
Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

The research provides a deep exploration of cryptocurrency price dynamics by blending technical analysis, sentiment analysis, and backtesting, aiming to reveal the hidden patterns, drivers, and irregularities in their price behaviors. As the field of cryptocurrencies gains importance, characterized by extreme price volatility and sensitivity to sentiment shifts, understanding these dynamics is vital for developing effective financial models and investment strategies. Cryptocurrencies are infamous for their unpredictable nature, often influenced by market sentiment as much, if not more, than fundamental or technical indications. This study aims to bridge the gap by evaluating the effectiveness of combining sentiment analysis with traditional technical analysis to enhance predictive accuracy and investment returns. We use various predictive models, including Support Vector Machine (SVM) and Random Forest, to evaluate their performance in different scenarios. Our findings reveal that the SVM model significantly outperforms other methods when sentiment analysis is merged. Specifically, sans sentiment analysis, the Random Forest model achieves an annual return of 3.59. Nevertheless, with sentiment analysis, the SVM model generates a distinctly higher annual return of 10.112. These results underscore the crucial role of sentiment analysis in boosting the predictive power of financial models concerning cryptocurrencies. Backtesting these models offers pragmatic insights into their effectiveness. The backtesting results show that including sentiment analysis in financial models not only enhances return metrics but also improves risk management. The superior performance of the SVM model with sentiment analysis underscores the impact of market sentiment on cryptocurrency prices, indicating that investor sentiment is a potent force that should not be ignored. The implications of these findings are substantial for both academia and practice. For researchers, this study adds to the growing body of literature on financial modeling in unstable and emerging markets, like cryptocurrencies. It presents empirical evidence supporting the merging of sentiment analysis into predictive models, thereby advancing theoretical understanding and methodological approaches in the field. For practitioners, particularly investors and financial analysts, the results provide actionable insights into optimizing investment strategies. By utilizing sentiment analysis, they can develop sturdier models that better capture market movements and investor behavior, leading to improved investment outcomes. The ability to predict price movements with increased accuracy permits more effective portfolio management and risk mitigation, which are crucial in the highly volatile cryptocurrency market. Additionally, the research accentuates the importance of continuous innovation in financial modeling techniques. As the cryptocurrency market evolves, so must the methods employed to analyze and predict its behavior. The integration of sentiment analysis represents a significant leap forward in this aspect, offering a robust tool to navigate the complexities of this emerging asset class. This research highlights the value of integrating sentiment analysis into financial models for cryptocurrencies. The findings indicate that such integration not only boosts predictive accuracy but also enhances investment returns and risk management. By advancing financial modeling techniques and providing practical insights for investment strategies, this study presents a significant contribution to both academic research and practical applications in the swiftly evolving world of cryptocurrencies.

Keywords: Bitcoin; Sentiment Analysis; Support Vector Machine; Random Forest; Backtesting; Naive Bayes Gaussian; Logistic Regression

Acknowledgement

All praise to the Great Allah for whom my thesis have been completed without any major interruption.

I want to express my sincere gratitude for the unwavering support during this research work from my supervisor, **Dr. Md. Golam Rabiul Alam** Associate Professor, Department of Computer Science and Engineering, BRAC University.

I want to express my gratitude to my parents and my entire family for their assistance. Without them, I would not be in this situation right now.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$\%R$	Williams Percent Range
A/D	Accumulation/Distribution
ANN	Artificial Neural Network
ATR	Average True Range
CCI	Commodity Channel Index
EMA	Exponential Moving Average
FSO	Fast Stochastic Oscillator
HMM	Hidden Markov Model
$LSTM$	Long Short Term Memmory
$MACD$	Moving Average Convergence Divergence
MLP	Multilayer Perceptron
$MRMR$	Maximum Relevance Minimum Redundancy
$OHLCV$	Open, High, Low, Close, Volume
RF	Random Forest
RSI	Relative Strength Index
SMA	Simple Moving Average
SVM	Support Vector Machine
TE	Technical Indicators
$TRIX$	Triple Exponential Moving Average
$VADER$	Valence Aware Dictionary and Sentiment Reasoner

Chapter 1

Introduction

The world of cryptocurrency has captured the imagination of millions, offering a glimpse into a decentralized financial future. However, for many, the volatile price movements of these digital assets remain an enigma. It is important for financiers, entrepreneurs, and everyone interested in developing this dynamic terrain to comprehend the factors behind these oscillations. It is important for financiers, entrepreneurs, and everyone interested in developing this dynamic terrain to comprehend the factors behind these oscillations. The economic ecosystem has been significantly disrupted by cryptocurrencies, which are rewriting conventional ideas about money, investing, and market behavior. The decentralized nature of cryptocurrencies, underpinned by blockchain technology, has ignited a paradigm shift in how financial transactions are conducted. As the market for cryptocurrencies continues to evolve, so does the complexity of understanding and predicting price movements within this dynamic ecosystem. While the field of cryptocurrency research is relatively young, a burgeoning body of literature has emerged, reflecting the growing interest and complexity of the cryptocurrency landscape. Previous studies have delved into various aspects, contributing valuable insights that form the foundation for this research. Early research in cryptocurrency markets often focused on assessing market efficiency and identifying anomalies. Studies explored whether traditional financial theories, such as the Efficient Market Hypothesis, hold in the context of cryptocurrencies. Researchers sought to uncover patterns and behaviors that deviate from the expectations of traditional financial models. Understanding the drivers of cryptocurrency prices and the factors contributing to their inherent volatility has been a central theme. Past studies investigated the impact of macroeconomic indicators, regulatory developments, technological advancements, and market sentiment on price movements. These inquiries laid the groundwork for comprehending the intricate dynamics shaping cryptocurrency valuations. As cryptocurrencies gained traction, researchers explored the factors influencing their adoption and the behavior of users within the ecosystem. Studies examined user motivations, the role of social networks, and the impact of educational initiatives on fostering widespread acceptance and use of cryptocurrencies.

1.1 Aims and Objectives

The research contributes to the field of cryptocurrency analysis by integrating a comprehensive methodology that encompasses technical indicators, sentiment analysis, and machine learning models to analyze cryptocurrency price movements. This study goes beyond traditional approaches by incorporating sentiment data alongside historical price data and technical indicators, offering a more nuanced understanding of the factors influencing cryptocurrency markets. Specifically, the contribution lies in the following key aspects:

1. Integrated Approach: The study adopts an integrated approach by combining historical price data, sentiment data, and technical indicators. By integrating these diverse sources of information, the research provides a more holistic view of cryptocurrency price movements, capturing both market dynamics and investor sentiment.

2. Sentiment Analysis: Leveraging Vader's algorithm for sentiment analysis, the study quantifies the positive and negative sentiment surrounding cryptocurrencies. This allows for a deeper analysis of how sentiment fluctuations impact price movements, providing valuable insights for traders and investors.

3. Data Fusion: The research merges sentiment data with historical price data, enabling the exploration of correlations between sentiment trends and cryptocurrency price movements. By integrating sentiment analysis into the analysis framework, the study enhances the predictive capabilities of traditional technical indicators.

4. Machine Learning Models: The study applies multiple machine learning models to forecast cryptocurrency price movements. The goal of the research is to increase forecast accuracy and detect trends that might not be apparent using only conventional analytic techniques by utilizing advanced algorithms.

5. Backtesting: Finally, the research conducts thorough backtesting to evaluate the performance of the proposed methodology. By backtesting the integrated approach against historical data, the study provides empirical evidence of its effectiveness in analyzing cryptocurrency price movements and informing trading strategies.

1.2 Structure of the Study

Our study paper is structured into several sections and subsections to enhance clarity and facilitate a thorough explanation of the various components included in our research. We provide a concise overview of each section within this framework. The Literature review section encompasses a review of existing academic research published in papers. In the methodology section, we detail our proposed model and research approach, covering key aspects such as data collection, preparation, and feature extraction, as well as feature selection, backtesting, implementation, and subsequent discussion of results. In the Result and discussions section, we analyze the outcomes of our ensemble model architecture, dividing our analysis into two subsections: performance metrics and discussion. Finally, in the conclusion section, we summarize our research findings and outline our strategies for future endeavors.

Chapter 2

Related Work

Cryptocurrency markets have been a focal point of academic inquiry as the decentralized digital assets continue to redefine the financial landscape. This literature review synthesizes existing research on various aspects of cryptocurrency price movements, emphasizing the integration of technical indicators, sentiment analysis, and backtesting.

The analysis of trends and factors influencing cryptocurrency prices, such as Bitcoin, Ethereum, and Ripple, has garnered considerable attention in recent research. Various methodologies have been proposed to enhance the predictability of cryptocurrency price movements. In [10], The authors has developed a statistical approach based on the Random Walk theory, specifically focusing on forecasting the real-time price of Bitcoin. Additionally, for currencies like Bitcoin, Ethereum, and Litecoin, the methodology incorporates Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks. Expanding on this research, another study [11] employs noise-correlated stochastic differential equations to establish a framework for understanding cryptocurrency price fluctuations, particularly in correlation with social media activities. The authors claim to forecast data over three months (April, June, and August) by drawing parallels between cryptocurrency price dynamics and those of traditional stock markets. Furthermore, [13] proposes a model for predicting Bitcoin prices using various neural network approaches. However, a notable drawback highlighted in this study is the independent prediction technique, which necessitates the model to establish correlations with known data for quantifying predictions accurately.A multi-input architecture-based deep neural model for bitcoin price prediction is presented in [23]. As said in [11], the study recognizes the extraordinary fluctuation of the cryptocurrency market and stresses the need to consider hidden variables like the dissemination of false information and the effect of social media on price changes. In [21], the authors utilize three Recurrent-Neural-Network models to forecast cryptocurrency prices: Bidirectional LSTM, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). Among these models, GRU demonstrates superior performance. However, the study overlooks factors such as trading volume and social media activity, which can significantly impact cryptocurrency prices. In [45], Hidden Markov Models (HMM) are proposed to describe previous cryptocurrency market trends, with a focus on incorporating new interaction features to enhance price prediction accuracy. Additionally, the study suggests that the Gradient Boosting Algorithm outperforms traditional machine learning models [27] due to its function-based optimization approach. Turning our attention to stock market forecasting, In **b7**, the author uses Random Forest, Bi-LSTM, and LSTM models to link sentiment scores with stock prices. Various features including stock prices, Gold prices, Oil prices, USD exchange rates, and Indian Government Securities yields are considered for model training. In [7], a related study predicts cryptocurrency prices through sentiment analysis using multiclassification using the Keras library and a Random Forest regressor for prediction. Notably, Random Forest exhibits superior accuracy compared to other models tested in the study, such as Decision Tree, Support Vector Machine, and Naive Bayes. The authors suggest potential improvements through techniques like Part-of-Speech (POS) tagging and word weighing. In the domain of stock market analysis, research **b7** utilizes historical stock prices from Yahoo Finance and incorporates sentiment scores to predict stock prices. Various features including Open, High, Low, and Close prices, as well as external factors like Gold prices, Oil prices, USD exchange rates, and Indian Government Securities yields are considered. In [15], the author proposed an innovative big data platform for price prediction that combines sentiment and pricing with standard machine learning models. Tweets from Twitter were captured in real time. The disadvantage was that deep learning models, such as RNN, were not evaluated for prediction. In [16], the LSTM-GRU hybrid model was used to forecast the price of Litecoin and Monero using different window widths. The hybrid model helps to mitigate the loss. The disadvantages were dependency between bitcoin and sentiment as a feature not considered. In [17], ARIMAX and LSTM-based RNNs were used to forecast bitcoin prices. The disadvantages were that hybrid models were not investigated, and feature fusion and sentiment were not examined. In [25], a hybrid LSTM and GRU-based deep learning model beat cutting-edge approaches in predicting the price of Litecoin and Zcash using the effect of big currencies like as Bitcoin. The disadvantages were perspectives of major crypto currencies is not regarded to predict the price of an impacted cryptocurrency. In [24], an ensemble model comprising LSTM, GRU, and TSN (Temporal Convolutional Networks) was employed to forecast the price of Ether using historical price data. Disadvantages included cryptocurrency interdependence and features not being included as a factor for predicting ether price. In [22], The author presented an LSTM and GRU-based method for predicting the prices of Bitcoin, Ethereum, and Litecoin. The price forecast was based on two types of data samples: Bitcoin, Ethereum, and Litecoin. One of the most essential components of market analysis is sentiment analysis, which is not included, as well as interdependence across currencies. In a recent study by Gülmez (2023) [32], a novel approach to stock price prediction is presented, leveraging a deep LSTM network optimized with the artificial rabbits optimization algorithm, highlighting its potential for improving forecasting precision. In [2], the authors delve into the realm of cryptocurrency trading by employing artificial neural network (ANN) methods to forecast the exchange rate between Bitcoin and the American dollar. The significance of this work lies in its exploration of predictive modeling techniques applied to a volatile and rapidly evolving asset class like Bitcoin. By leveraging ANN methods, the study aims to provide insights into the potential for forecasting cryptocurrency exchange rates, a task fraught with challenges due to the unique characteristics of digital currencies. The utilization of ANN methods in predicting Bitcoin exchange rates underscores the interdisciplinary nature of the research, bridging computer science and finance domains. This interdisciplinary approach is increasingly relevant as digital currencies gain prominence in global financial markets, necessitating innovative methodologies for analysis and prediction. The paper contributes to the growing body of literature on cryptocurrency trading and forecasting techniques, shedding light on the applicability of ANN methods in this context. As such, it serves as a valuable reference for researchers and practitioners seeking to understand the dynamics of cryptocurrency markets and develop effective trading strategies. In the broader context of algorithmic trading and financial market analysis, studies like this one highlight the importance of leveraging advanced computational methods to navigate the complexities of modern financial systems. By incorporating insights from the paper into the literature review, researchers can gain a comprehensive understanding of the evolving landscape of cryptocurrency trading and predictive modeling techniques. In [4], the authors' exploration of various machine learning models for Bitcoin price prediction underscores the importance of employing advanced computational techniques in understanding and forecasting the behavior of digital assets. By comparing the performance of different models, the paper provides insights into the efficacy of machine learning algorithms in capturing the complex dynamics of cryptocurrency markets. The significance of this work lies in its practical implications for traders, investors, and researchers seeking to navigate and capitalize on the fluctuations of Bitcoin prices. By identifying the strengths and weaknesses of various machine learning approaches, the study offers valuable guidance for selecting suitable predictive models in the context of cryptocurrency trading. In [5], the authors adopt a holistic perspective in developing their predictive model, recognizing the interconnectedness of various factors influencing cryptocurrency prices. By considering a wide range of variables, including market sentiment, technical indicators, and macroeconomic factors, the study aims to capture the complex dynamics driving cryptocurrency price movements. This work is significant because it focuses on creating a strong prediction model to help shed light on how the global Bitcoin market behaves. Integrating multiple data sources and employing advanced analytical techniques, the paper offers a novel approach to forecasting cryptocurrency prices, addressing the inherent challenges posed by the volatile and decentralized nature of digital assets. In [6], The authors' hybrid model combines the strengths of both HMM and LSTM networks, leveraging the temporal dependencies captured by LSTM networks and the probabilistic framework offered by HMM to improve prediction performance. By integrating these two techniques, the study addresses the challenges associated with modeling the complex and non-linear nature of cryptocurrency price movements. The significance of this work lies in its adoption of a hybrid approach, which reflects the growing trend towards combining multiple methodologies to achieve more accurate predictions in financial markets. The study provides a fresh approach to the cryptocurrency price prediction problem by taking advantage of the complimentary properties of HMM and LSTM networks, furthering the development of predictive modeling methods in this field. The creators of [9] employ sentiment analysis, a method for assessing the sentiment or emotion conveyed in textual data, to extract valuable insights from news articles, social media, and other data sources about cryptocurrency markets. The study aims to include market sentiment as a prediction component in addition to conventional financial indicators by assessing the mood surrounding cryptocurrencies. This paper is vital because it acknowledges the influence of market condition on the dynamics of bitcoin prices. By integrating sentiment analysis into machine learning models, the authors provide a holistic framework for forecasting price movements, capturing the influence of investor sentiment on market behavior. In [3], the authors apply the Adaptive Market Hypothesis, which posits that financial markets are not inherently efficient but rather adapt and evolve over time in response to new information and participant behavior. Through an analysis of Bitcoin price data, Khuntia and Pattanayak examine the degree of predictability in Bitcoin prices and how it evolves over different market conditions. To reflect cryptocurrency markets' intrinsic volatility and unpredictability, the authors [12] suggest using stochastic neural networks, which combine neural network topologies with stochastic processes. By incorporating stochasticity into the modeling process, the study aims to develop more robust and accurate predictive models for cryptocurrency price movements. In [8], Sharma delves into the intricate connection between the energy-intensive process of Bitcoin mining and its potential impact on the cryptocurrency's price. By analyzing the cost structure of Bitcoin mining operations and considering factors such as electricity expenses and mining difficulty adjustments, the article sheds light on how changes in mining costs may influence Bitcoin's market dynamics. In [18], Tran and Leirvik's study examines the efficiency of cryptocurrency markets by analyzing the speed and accuracy of price adjustments to new information. By utilizing a variety of criteria and approaches, the authors evaluate market efficiency and provide insight into whether or not exchange possibilities are present and the extent to which bitcoin prices accurately represent all available information. The authors [1] examine how sentiments about cryptocurrencies and the ensuing price fluctuations connect to news stories and social media posts. By employing sentiment analysis techniques, the study aims to uncover patterns and correlations that may inform predictive models for cryptocurrency prices. The significance of this work lies in its innovative approach to incorporating non-traditional data sources into financial analysis and prediction. Lamon et al. recognize the growing influence of news and social media on investor sentiment and market dynamics, highlighting the potential for sentiment analysis to offer valuable insights into cryptocurrency price movements.

2.1 A Systematic Review

In order to determine trending algorithms and their influence on pricing variations, we examined recent studies on Bitcoin price fluctuations. Our analysis uncovered a range of approaches, including machine learning strategies like Random Forest, SVM, and LSTM, which are well-liked right now for their high prediction accuracy. Furthermore, deep learning techniques and hybrid models like Deep Q-Network and BiLSTM have demonstrated great potential in enhancing prediction dependability. The Table 2.1 summarizes the purposes of studies on predicting Bitcoin price movements. The studies aim to Using the right techniques to avoid overfitting the data, Using social media data and an end-to- end approach, Comparing the performance of different models. The Table 2.2 description focuses on datasets used to predict Bitcoin prices, integrating historical data, technical indicators, and social media sentiment. Sources include Yahoo Finance, Kaggle, and various exchanges like GDAX and Binance, spanning from January 2012 to August 2023. The datasets range from daily to minute-by-minute intervals, capturing extensive trading data and social trends. These diverse data sources, featuring millions of records, are utilized to create robust predictive models, leveraging time series data to analyze market trends

Table 2.1: Key Purposes of The Reviewed Studies

Key Purpose	Analysis Type	Brief Description	Reference
Improve Bitcoin Forecasting	Classification	The right regularisation techniques and model evaluation methods to avoid overfitting	[41]
Improve Bitcoin Forecasting	Regression	Using hybrid technique to achieve best performance	[42]
Bitcoin Price Prediction	Regression	Using social media and an end-to-end approach	[34]
Compare Bitcoin Predictors	Regression	By comparing the performance of different models	[44]
Compare Bitcoin Algorithms	Classification	Hybrid model that combines traditional time series analysis with advanced techniques	[37]
Predict Bitcoin Fluctuations	Classification	Acquiring an accurate forecast is crucial, and reaching this accuracy	[39]
Compare Bitcoin Forecasters	Classification	Make more dependable and precise forecasts about the direction of the price of bitcoin using a Deep Q-Network (DQN) model.	[33]
Enhance Cryptocurrency Prediction	Regression	A pioneering methodology for time series prediction of Bitcoin	[38]
Bitcoin Price Prediction	Regression	Developing an efficient framework for predicting Bitcoin prices using various machine learning algorithms	[36]
Predict Cryptocurrency Movements	Classification	Machine learning approach using cryptocurrency market data to predict price changes, emphasizing the importance of feature selection, model updating, and efficient training methods	[35]
Compare Forecasting Models	Classification	The ability of various machine learning models to forecast bitcoin values and the possible uses of these models in trading methods	[31]
Enhance Cryptocurrency Prediction	Regression	Using the PELT algorithm to identify notable shifts in the price of cryptocurrencies	[30]
Predict Bitcoin Fluctuations	Classification	Investigating Bitcoin price fluctuation prediction problem	[14]

Table 2.2: A Synopsis of the Information Used in the Contextual Literature

Data Source	Data Volume	Literature
1. Historical Prices from Yahoo Finance 2. Technical Indicators 3. Daily tweets about Bitcoin & Google search trend data	1. Complete Dataset: Data from 17/09/2014 to 10/06/2021 2. Reduced Dataset: Data from 10/06/2020 to 10/06/2021 3. Validation Set: Data from 11/06/2021 to 09/08/2021	[41]
Bitcoin Historical Data from Kaggle competitions	January 2012 to September 2020 at 1 minute interval	[42]
1. Historical Prices from Yahoo Finance 2. Sentiments expressed on Twitter	1. Not Specified 2. Accounts that have more than 100k followers 25,64,350 rows, divided among three .csv files with approximately 8,54,783 rows each	[34]
Bitcoin Historical Dataset from Kaggle	From November 2014 to December 2021, or around 2700 samples	[44]
Bitcoin Historical Dataset from Kaggle	January 2012 to March 2021	[37]
1. Conventional statistics on pricing and volume 2. On-chain analytics that show the network activity that underpins Bitcoin 3. Social media metrics from Google Trends and Twitter	Collected on an hourly basis	[39]
Sentiment scores from social media platforms like Twitter and Google Trends	Collected on an hourly basis	[33]
Historical price data	Hourly and daily price data	[38]
Collected from multiple sources including trading platforms like Bitstamp	1 minute interval trading data over six years. Dataset 1: Bitcoin price data from 2014 to 2021. Dataset 2: Bitcoin price data from 2014 to 2022. Dataset 3, 4, and 5: Specific details about these datasets are not explicitly mentioned in the document sections reviewed, but they also pertain to Bitcoin's historical price data and potentially include similar features	[36]
GDAX exchange WebSocket API	Order flow data: 61,909,286 records, Ticker data: 128,593 data points, Level-2 data: 40,951,846 records	[35]
From Investing.com	The document does not specify the exact size of the dataset, but it mentions daily time-series data covering several cryptocurrencies over a substantial period, including the post-pandemic period (from January 1, 2020, to August 31, 2023)	[31]
From a reputable cryptocurrency exchange	Bitcoin price statistics every day between January 2020 and April 2023	[30]
Huobi, Coinbase, Binance, Bitstamp, and Bitfinex	August 2017 to May 2020, with an interval of 1 minute	[14]

and enhance the accuracy of Bitcoin price forecasting. The most recent developments (Table 2.3) in methodology used to forecast changes in Bitcoin price movement

Table 2.3: A Summary of the Algorithm Used in the Literature

Algorithm	**Evaluation Results**	**Reference**
LTSM, SVM, ANN, Random Forest	Accuracy (87.10%)	[41]
ARIMA, LSTM, FB-prophet, XG Boost, LSTM-GRU, LSTM-1D_CNN	RMSE (83.408) & MAE (9.140)	[42]
Logistic Regression, Ridge Regression, Elastic Net Decision, Tree Regression, Random Forest Regression, AdaBoost Regression, Gradient Boost Regression, XGBoost Regression	MAPE (8.49%)	[34]
Stochastic Gradient Descent (SGD) Regression, Ridge Regression, Elastic Net, Decision Tree Regression, Random Forest Regression, AdaBoost Regression, Gradient Boosting Regression, XGBoost Regression	Two algorithms LSTM and RNN should be used, (LR) had the highest mean absolute error (MAE) of 2476.9	[44]
Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Seasonal Autoregressive Integrated Moving Average (SARIMA)	Accuracy (98%)	[37]
Deep Q-Network (DQN)	F1 score (95%)	[39]
Deep Q-Network (DQN) algorithm with novel reward function	F1 score (95%)	[33]
Performer neural network and BiLSTM	Lower MSE & higher R-Square values	[38]
Linear Regression, Random Forest, AdaBoost, Decision Tree, K-Nearest Neighbors (KNN), Gradient Boosting, Constant Predictor, Neural Network, and Support Vector Machine (SVM)	Linear Regression was optimal for the first dataset, Random Forest for the second and fourth datasets, AdaBoost for the third dataset, and Linear Regression again for the fifth dataset	[36]
LSTM	F1 score (78%)	[35]
AdaBoost, LightGBM, Simple RNN, GRU, LSTM, ARIMA combined with MLP and LSTM	72.49% directional accuracy	[31]
Long Short-Term Memory (LSTM) network combined with the Pruned Exact Linear Time (PELT) algorithm	PELT algorithm outperformed the baseline LSTM model in terms of all evaluation metrics	[30]
Adaptive and Locally-Excited Network Model	Accuracy: 61.15% , Precision: 61.10%, Recall: 60.93%, F1 score: 61.01%	[14]

are presented in this report. Among the methods are SARIMA, K-Nearest Neighbors, Logistic Regression, Linear Regression, and sophisticated neural networks like RNN and LSTM. Combinations like LSTM with PELT and cutting-edge algorithms like Deep Q-Networks are examples of notable developments. Evaluation criteria including as accuracy, RMSE, MAE, and F1 scores demonstrate the effectiveness of these methods; some models reach up to 98% accuracy and achieve state-of-the-art performance. This review summarizes the many innovative technology approaches used in current studies to improve model resilience and prediction accuracy.

We verify the efficacy of our techniques by means of thorough backtesting, guaranteeing their resilience and dependability in actual situations. Our goal is to provide investors with actionable knowledge so they can make wise decisions and confidently adjust to changing market conditions. In order to help investors reach their goals in the constantly changing world of cryptocurrency trading, our strategy strategically aligns investment decisions with market sentiment and technical research. This approach aims to generate portfolio development and financial success.

Our study, which is especially designed for the volatile cryptocurrency market, combines sentiment analysis, backtesting, and technical indicators to solve a typical machine learning classification issue. Our goal is to maximize investor profits and minimize risk by utilizing these technologies to optimize investing strategies. Our method enables investors to buy, hold, or sell assets with confidence by continuously monitoring important variables and making well-informed decisions. This helps investors efficiently manage market swings. Our technique offers useful insights to improve portfolio performance and take advantage of profitable opportunities while limiting possible losses by methodically examining market patterns and sentiment.

Chapter 3

Methodology

The model architecture illustrated in Figure 3.1 offers a comprehensive and structured depiction of our research methodology, crucial for understanding the intricate steps involved in our study. Beginning with data collection, both historical and sentiment data are gathered to form the foundation of our analysis. Subsequently, the data undergoes meticulous preparation and feature extraction, incorporating technical indicators and rigorous data cleaning processes to ensure the integrity of the dataset. Feature selection techniques, including Maximum Relevance Minimum Redundancy, are then applied to refine the feature set, optimizing the model's predictive capabilities.

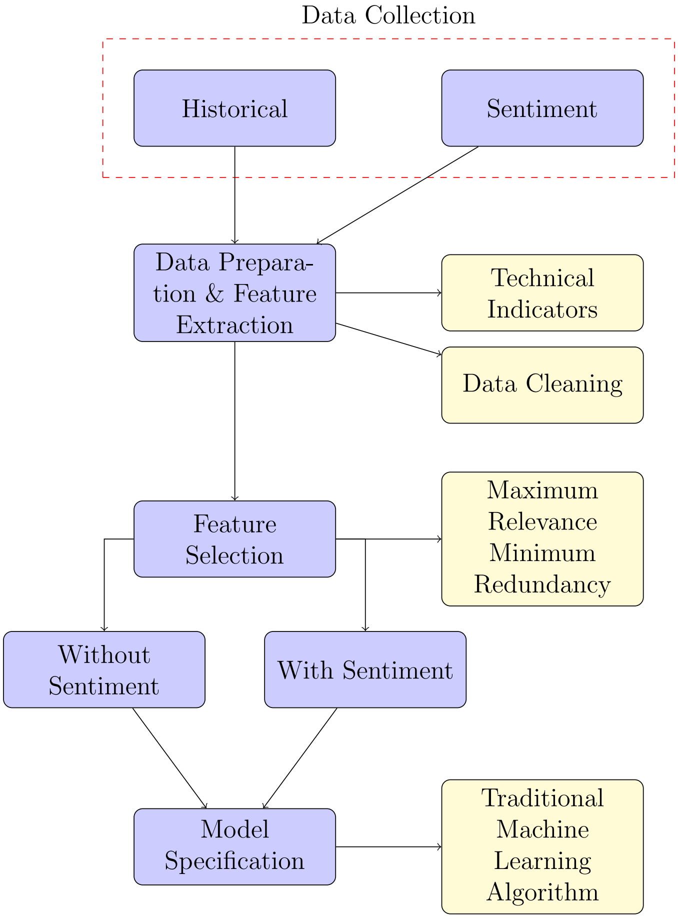

Figure 3.1: Top Level Overview of the Proposed Architecture

The model is further specified, distinguishing between scenarios with and without sentiment analysis integration, thereby accommodating different analytical approaches. Traditional machine learning algorithms are then deployed for model training and prediction, leveraging the insights gained from the refined dataset. Finally, the model's efficacy is evaluated through rigorous backtesting procedures using the Zipline library. This methodical approach ensures a robust analysis of financial data, while the integration of sentiment analysis adds an additional layer of insight to enhance predictive accuracy and decision-making capabilities.

3.1 Data Collection

Bitcoin (BTC) stands as the pioneering cryptocurrency, commanding significant attention and market dominance. In this study, BTC serves as the primary focus due to its widespread adoption, liquidity, and historical data availability. The methodology encompasses data collection and analysis from multiple sources to provide a comprehensive understanding of BTC price movements.

3.1.1 Historical Data

Historical (OHLCV) data for BTC are collected from the CoinMarketCap[19] Cryptocurrency Aggregate (CCCAGG) API.

Table 3.1: Historical Data

Features	**Description**
Date	The specific date associated with the historical data point. This could be in various formats such as YYYY-MM-DD or as a timestamp.
Open	The price of the financial instrument (e.g., stock, cryptocurrency) at the beginning of the time interval.
High	The highest price reached during the time interval.
Low	The lowest price reached during the time interval.
Close	The price of the financial instrument at the end of the time interval.
Volume	The total amount of the financial instrument traded during the time interval.

The data spans from January 1, 2021, to October 9, 2023, capturing various market conditions and trends over the selected timeframe.

3.1.2 Sentiment Data

Sentiment data from different authors and sources are collected to gauge market sentiment regarding BTC. The sentiment data cover various categories, including Markets, Finance, Technology, and others, to capture diverse perspectives and influences on cryptocurrency markets. Sentiment data were primarily collected from CoinDesk, a prominent cryptocurrency news platform, to ensure a comprehensive representation of sentiment across different categories. The sentiment data range from January 1, 2021, to October 9, 2023, aligning with the latest available sentiment analysis resources.

Table 3.2: CoinDesk Data

Features	Description
Date	The specific date associated with the CoinDesk data point. This could be in various formats such as YYYY-MM-DD or as a timestamp.
Category	CoinDesk's news platform encompasses a variety of content categories, including podcasts, market updates, and technology insights.
News Content	CoinDesk provides diverse news content covering the latest developments in the cryptocurrency and blockchain industry.
Authors	It provides a diverse roster of authors comprising seasoned journalists, industry insiders, and subject matter experts.

3.2 Data Preparation and Feature Extraction

Several technical indicators are used in this study to fully examine the price movements of cryptocurrencies. Technical indicators offer insights into market trends, momentum, volatility, and possible reversal points using mathematical computations based on past price, volume, or sentiment data. There are several technical indicators that each show distinct facets of price action and market momentum. Indictor selection is influenced by the trader's approach, the state of the market, and their level of risk tolerance. Volume-based indicators (accumulation/distribution Line), Oscillators (RSI, Stochastic Oscillator), Moving Averages (SMA, EMA), and Volatility indicators (Bollinger Bands) are examples of frequently used indicators. Trading techniques include technical indicators to offer objective standards for entering and exiting deals. Traders evaluate the performance of certain indicators and adjust settings by backtesting their techniques using past data. Traders may adjust to shifting market circumstances and improve their tactics over time by consistently checking their indicators. Even though technical indicators provide insightful information, they are not perfect and might provide erroneous signals, particularly in erratic or turbulent markets. Risk management strategies are crucial to preserve money and reduce losses, including sizing positions according to volatility and establishing stop-loss orders. Traders should use technical indicators as a component of an all-encompassing risk management strategy that is customized to meet their unique risk tolerance and trading goals.

3.2.1 Technical Indicators

Technical analysis relies heavily on indicators[29], which provide traders with important information about price movements, trends, momentum, and volatility in the financial markets. These instruments are available in several formats, such as indicators based on signals and indicators based on numerical values. Relative Strength Index (RSI) and Moving Average Convergence Divergence (MACD) Histogram are two examples of signal-based indicators that provide traders with precise buy and sell signals based on predetermined criteria, assisting them in identifying possible entry and exit positions. Conversely, traders may more accurately assess the intensity and direction of market moves thanks to numerical value-based indicators like Simple Moving Averages (SMA) and Exponential Moving Averages (EMA), which provide quantitative evaluations of price trends and momentum. For traders at any skill level, indicators are vital tools that support risk management, decision-making, and the creation of trading plans that are specific to each trader's preferences and the state of the market.For traders looking to efficiently pass through financial markets and take advantage of trading opportunities, indicators are essential tools, whether they are employed alone or in conjunction with other analytical approaches. Listed below is a selection of indicators.

1. SMA: The average price of an asset over a certain time period is determined by the SMA, usually using closing prices. By removing short-term swings, it helps to normalize pricing data and spot patterns. Higher SMA values imply an uptrend, while lower values show a downturn, as the SMA value indicates the average price for a certain period of time. SMA crossovers are a common tool used by traders to indicate bullish momentum when shorter-term SMAs cross above longer-term SMAs and vice versa.

2. EMA: The EMA, like the SMA, determines the average price of an asset over a given time frame. Nonetheless, it is more sensitive to the state of the market since it places greater weight on recent prices. Because of this, EMAs respond to price fluctuations faster than SMAs. EMA crossovers are used by traders, much like SMAs, to identify possible buy or sell opportunities. Bullish momentum is indicated when shorter-term EMAs cross above longer-term EMAs, and vice versa.

3. Chaikin Oscillator: The difference in the values of the Accumulation/Distribution Line over the short and long terms is measured by this momentum indicator. It shows how strong the market's purchasing and selling pressure is. Increased selling pressure is indicated by negative numbers, and increased purchasing pressure is suggested by positive ones. Depending on the trader's approach and the state of the market, crossings above zero may indicate possible purchase chances, while crosses below zero may indicate potential sell opportunities.

4. MACD: [29] The MACD The histogram shows how the signal line and the MACD line differ from one another. It offers information about a trend's momentum as well as possible trend reversals. Whereas negative histogram values imply bearish momentum, positive values indicate bullish momentum. In order to identify possible buy or sell opportunities, traders frequently search for histogram crosses above or below the zero line.

5. Williams Percent Range (%R): A momentum oscillator that gauges the current close in relation to the high-low range over a certain time period is the Williams Percent Range (%R). It fluctuates between -100 and 0, with readings below -80 indicating an oversold situation and values over -20 indicating an overbought one. To determine if an asset is potentially overbought or oversold, traders use %R readings. possible purchase opportunities may be indicated by crossings

above -50, while possible sell opportunities may be indicated by crosses below -50.

$$\%R = \frac{(H_{14} - C)}{(H_{14} - L_{14})} \times -100 \tag{3.1}$$

Where:

$$\begin{aligned} \%R &: \text{William Percent Range} \\ H_{14} &: \text{Highest price in the last 14 periods} \\ L_{14} &: \text{Lowest price in the last 14 periods} \\ C &: \text{Closing price} \end{aligned}$$

6. CCI: A momentum-based oscillator that gauges the relationship between an asset's price, moving average, and standard deviation is the Commodity Channel Index (CCI) [29]. It shows the strength of the trend and possible overbought or oversold circumstances. Bullish momentum is indicated by CCI values over 0, whilst bearish momentum is suggested by values below 0. When searching for possible buy or sell opportunities, traders frequently watch for crossovers above or below the zero line.

$$\text{CCI} = \frac{\text{TP} - \text{SMA}}{0.015 \times \text{Mean Deviation}} \tag{3.2}$$

where:

$$\begin{aligned} \text{TP} &= \text{Typical Price} \\ &= \frac{\text{High} + \text{Low} + \text{Close}}{3} \\ \text{SMA} &= \text{Simple Moving Average of TP} \\ \text{Mean Deviation} &= \frac{\sum_{i=1}^{n} |\text{TP}_i - \text{SMA}_n|}{n} \end{aligned}$$

7. Stochastic Oscillator: As a momentum indicator, the stochastic oscillator contrasts the closing price of the market today with the range of prices over a certain time frame. It is composed of two lines: %K and %D. %K shows the price position as of right now in relation to the price range, and %D is the moving average of %K. Stochastic oscillator signals, such %K crossings above or below %D, are used by traders to determine whether to buy or sell.

8. Keltner Channels: Three lines make up Keltner Channels, an upper and lower band based on Average True Range (ATR) and an Exponential Moving Average (EMA) in the center. Based on volatility, these channels seek to discover possible buy and sell signals. Breakouts below the lower band may indicate possible oversold conditions and point to a buy opportunity, while breakouts above the upper band might indicate possible overbought conditions and point to a sell opportunity.

9. Triple Exponential Average (TRIX): The Triple Exponential Average (TRIX) is a momentum oscillator that measures the percent rate of change of a triple exponentially smoothed moving average. It aims to filter out short-term fluctuations and identify long-term trends. TRIX values crossing above zero may signal potential buy opportunities, while crosses below zero may suggest potential sell opportunities, indicating shifts in momentum.

$$\text{TRIX} = \text{EMA}_n(\text{EMA}_n(\text{EMA}_n(\text{Close}))) \tag{3.3}$$

where:

$$\begin{aligned} \text{Close} &= \text{Closing Price} \\ \text{EMA}_n &= \text{Exponential Moving Average} \\ &\quad \text{with smoothing factor } \alpha = \frac{2}{n+1} \end{aligned}$$

10. Accumulation/Distribution: The Accumulation/Distribution [29] indicator measures the flow of money into or out of a security by analyzing price and volume data. It accumulates volume based on whether the close is higher or lower than the previous close and adjusts for the trading range. Increasing A/D values suggest buying pressure, while decreasing values suggest selling pressure. Crosses above or below its moving average can signal potential buy or sell opportunities.

11. Donchian Channels: Donchian Channels consist of upper and lower bands that represent the highest high and lowest low over a specified period. These channels aim to capture the price's trading range and identify potential buy and sell signals. Breakouts above the upper band may signal potential buy opportunities, while breakouts below the lower band may signal potential sell opportunities, indicating shifts in market momentum.

12. RSI: RSI is a momentum oscillator that measures the speed and change of price movements. It oscillates between 0 and 100 and indicates potential overbought or oversold conditions. RSI values above 70 indicate overbought conditions, while values below 30 suggest oversold conditions. Crosses above 70 or below 30 may signal potential sell or buy opportunities, respectively, depending on market conditions.

$$\text{RSI} = 100 - \frac{100}{1 + \text{RS}} \tag{3.4}$$

where:

$$\begin{aligned} \text{RS} &= \frac{\text{Average Gain}}{\text{Average Loss}} \\ \text{Average Gain} &= \text{Average of gains over a specified period} \\ \text{Average Loss} &= \text{Average of losses over a specified period} \end{aligned}$$

13. Bollinger Bands: Bollinger Bands are made up of three lines: upper and lower bands determined by standard deviations of price movements, and a center band that represents a SMA. The purpose of these bands is to record price fluctuations and spot possible buy and sell signals. Breakouts below the lower band may indicate possible oversold conditions and point to a buy opportunity, while breakouts above the upper band might indicate possible overbought conditions and point to a sell opportunity.

14. Average True Range (ATR): Calculating the average range between high and low prices over a certain period of time allows the Average True Range (ATR) to be used to quantify market volatility. It offers information about possible price reversals as well as the extent of price movement. Greater volatility and possible buy or sell opportunities are indicated by higher ATR readings. ATR level breakouts or breaks below them might indicate future shifts in the market's momentum.

$$\begin{aligned}\text{ATR} = \frac{1}{n}\sum_{i=1}^{n} \max\Big(& \\ & \text{high}_i - \text{low}_i, \\ & |\text{high}_i - \text{close}_{i-1}|, \\ & |\text{low}_i - \text{close}_{i-1}|\Big) \end{aligned} \tag{3.5}$$

Where:

$$\begin{aligned} \text{ATR} &= \text{Average True Range} \\ \text{high}_i &= \text{Highest price of the } i\text{th period} \\ \text{low}_i &= \text{Lowest price of the } i\text{th period} \\ \text{close}_{i-1} &= \text{Closing price of the previous period} \\ n &= \text{Number of periods} \end{aligned}$$

15. FSO: In order to find possible buy and sell signals, the FSO computes %K and %D based on previous price movements. A moving average of %K is represented by %D, and the current price position in relation to the price range is represented by %K. possible purchase opportunities may be indicated by a cross of %K above %D, while possible sell opportunities may be indicated by a cross below %D, which would indicate a change in momentum.

3.2.2 Signal-Based Indicators

The Stochastic Oscillator, Williams Percent R, and Moving Average Convergence Divergence (MACD) Histogram are examples of signal-based indicators that shed light on overbought and oversold situations as well as possible trend reversals. Traders utilize the indications produced by these indicators to guide their selections. For instance, a crossing over 80 on the Stochastic Oscillator suggests overbought conditions and may be a buy signal; a crossover below 20 suggests oversold conditions and may be a sell signal. Similar to this, crossings between the MACD

Table 3.3: Comparison of Indicators

Technical Indicators	**Parameters**	
	Signal	***Numercial Value***
SMA	period=20	period=20
EMA	period=20	period=20
Chaikin Oscillators	short period=3 long period=10	short period=3 long period=10
MACD	short period=12 long period=26 signal period=9	short period=12 long period=26 signal period=9
%R	period=30	period=14
CCI	period=14	period=14
Stochastic Oscillator	%k period=14 %d period=3	%k period=14 %d period=3
Keltner Channels	ema period=20 atr period=10 atr multiplier=2	ema period=20 atr period=20 atr multiplier=2
TRIX	ema period=15	ema period=15
(A/D)	period=50	n/a
Donchian Channels	lookback period=60	lookback period=20
RSI	period=14	period=14
Bollinger Bands	period=20 std multiplier=2	period=20 std multiplier=2
ATR	period=20	period=14
FSO	%k period=14 %d period=3	%k period=14 %d period=3

line and the signal line provide the basis for the buy and sell signals produced by the MACD Histogram. These indicators are useful for determining the sentiment of the market and spotting short-term trading opportunities.

3.2.3 Numerical Value-Based Indicators

The SMA, EMA, and RSI are examples of numerical value-based indicators that offer numerical values that indicate particular characteristics of the price movement or momentum of a securities. Quantitative insights into trends, momentum, and volatility are provided by these indicators. An upward trend, for example, is suggested by a rising SMA or EMA, whereas a downward trend is shown by a dropping SMA or EMA. The RSI, which has a range of 0 to 100, measures how strongly prices move and indicates whether an asset is overbought or oversold. circumstances that are overbought are indicated by a number above 70, and oversold circumstances are indicated by a rating below 30. These numerical figures are used by traders to evaluate the intensity and direction of trends, pinpoint possible entry and exit locations, and efficiently manage risk.

3.2.4 Data Cleaning

After computing technical indicators from historical data, the subsequent step involves the identification and removal of null values and outliers. Null values, often stemming from missing or incomplete data, can distort analysis results and compromise the integrity of the study. Likewise, outliers, which are data points significantly different from the rest of the dataset, can skew statistical measures and mislead analysis outcomes. By systematically filtering out null values and outliers, researchers ensure the dataset's reliability and enhance the accuracy of subsequent analyses. This preprocessing step is crucial for maintaining data quality and integrity, enabling researchers to derive meaningful insights into cryptocurrency price movements with greater confidence.

3.3 Feature Selection

In this study, OHLCV data, denoting Open, High, Low, Close, and Volume, is employed as the primary dataset for analysis. OHLCV data encapsulates essential information about financial instruments over a given time frame, including the opening and closing prices, highest and lowest prices reached, and the trading volume within that period. The Close price is selected as the primary feature for its significance in financial analysis, representing the final price at which a security was traded during the period.

Fifteen popular indicators are chosen for analysis, comprising various technical analysis tools widely used in financial markets. These indicators serve dual purposes: signal generation and numerical value computation. Signal generation refers to the identification of potential trading opportunities based on specific conditions derived from the indicators, while numerical value computation involves extracting quantitative information from these indicators to supplement the analysis. The chosen indicators include but are not limited to Moving Averages, RSI, Stochastic Oscillator, and MACD . Each indicator contributes one feature for signal generation, resulting in a total of 15 features. However, certain indicators, such as Keltner Channels, Bollinger Bands, and Donchian Channels, introduce variations that lead to an expanded feature set for numerical value computation. For instance, Keltner Channels consist of an upper and lower channel, while Bollinger Bands include upper and lower bands, and Donchian Channels also encompass upper and lower channels. Hence, the total number of features for numerical value computation amounts to 18. Additionally, the feature set includes the 3-day percentage change of the Close price, which provides insight into short-term price dynamics and market volatility. The target variable is defined based on the percentage change of the Close price over the specified period. If the percentage change exceeds 2.5%, a buy signal is generated (labeled as 1), indicating a potential bullish trend. Conversely, if the percentage change falls below -2.5%, a sell signal is generated (labeled as -1), suggesting a potential bearish trend. If the percentage change falls within the range of -2.5% to 2.5%, a hold signal is assigned (labeled as 0), indicating a neutral or uncertain market condition.

Upon completing the feature calculations, we amalgamated various elements to form a comprehensive dataset. This amalgamation included incorporating the closing price and percentage data alongside 15 distinct features derived from signal generation processes, complemented by an additional 18 features representing numerical values. Consequently, the dataset was enriched with a total of 35 features, excluding the target variables. This meticulous consolidation process ensured that the dataset encapsulated diverse aspects relevant to our analysis, laying a solid foundation for subsequent investigations and modeling endeavors.

3.3.1 MRMR Classifier

The Highest Significance A feature selection technique used in machine learning for classification applications is the Minimum Redundancy (MRMR) classifier. In order to minimize redundancy among the selected characteristics, it seeks to identify a subset of features from a broader collection of features that are most pertinent to the target variable. Features with the highest relevance to the target variable are chosen by MRMR. Stated differently, it gives priority to characteristics that are most useful in differentiating across groups or classifications. MRMR not only maximizes relevance but also reduces duplication among the chosen characteristics.This implies that instead of choosing several characteristics that communicate the same information, it aims to incorporate elements that offer distinct and complimentary information. With a vast number of features, high-dimensional data sets may be handled well by MRMR. It lessens the chance of overfitting and increases classification process efficiency by choosing a small subset of pertinent characteristics. Support vector machines (SVM), neural networks, decision trees, and other classification methods may all be used with the versatile feature selection technique known as MRMR. It is adaptable to many machine learning models and is not dependent on any particular classifier.

The MRMR classifier operates in two main steps: Feature Ranking - In the first step, MRMR computes a relevance score for each feature based on its correlation or mutual information with the target variable. Features are ranked in descending order of relevance. In the second step, MRMR selects a subset of features that maximize relevance while minimizing redundancy. This is achieved through iterative selection of features that offer the highest marginal relevance (i.e., the most informative) while considering the redundancy with previously selected features.

To select features using MRMR principles, we typically compute relevance (R) and redundancy (D) scores for each feature and then select features that maximize relevance while minimizing redundancy. One way to do this is through the use of mutual information.

The relevance score for feature i with respect to the target variable Y can be computed as:

$$R_i = I(X_i; Y) \tag{3.6}$$

Where $I(X_i; Y)$ represents the mutual information between feature X_i and the target variable Y.

The redundancy score for feature i with respect to the previously selected features S can be computed as:

$$D_i = \frac{1}{|S|} \sum_{j \in S} I(X_i; X_j) \tag{3.7}$$

Where $|S|$ represents the number of previously selected features, and $I(X_i; X_j)$ represents the mutual information between feature X_i and feature X_j.
Finally, we select features based on a criterion that balances relevance and redundancy, such as:

$$\text{MRMR score}_i = R_i - \lambda \cdot D_i \tag{3.8}$$

Where λ is a parameter that controls the trade-off between relevance and redundancy. Features with higher MRMR scores are preferred for selection.

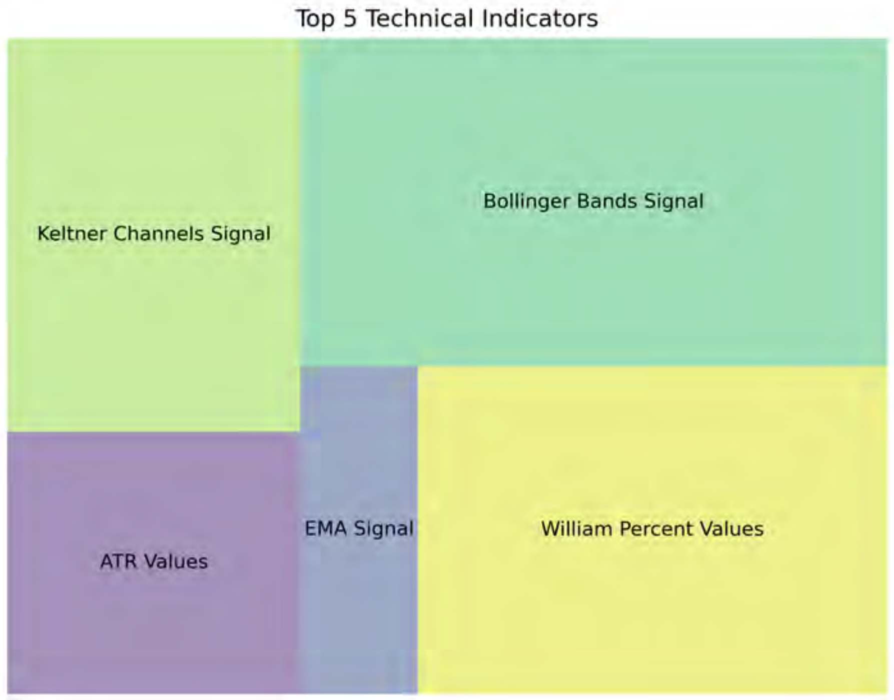

Figure 3.2: Treemap of Top 5 Technical Indicators by MRMR Feature Selection Method

Following the initial feature selection process, a total of 35 features were identified, excluding the target feature. However, through meticulous data cleaning procedures involving the removal of anomalies and null values, the dataset was refined to 1901 entries. Employing MRMR (Minimum Redundancy Maximum Relevance) classifier, the feature set was further reduced to 5, excluding the target variable. Within this refined set, 3 features were attributed to signal generation, while 2 features represented numerical values. The remaining features encompassed data related to closing prices and percentage changes. This systematic approach not only facilitated the reduction of irrelevant features but also ensured the retention of those most pertinent to the analysis, thereby enhancing the robustness and efficiency of subsequent modeling and predictive tasks.

3.4 Sentiment Analysis

CoinDesk stands as a premier platform in the cryptocurrency space, renowned for its authoritative coverage, insightful analysis, and up-to-date news regarding Bitcoin and other cryptocurrencies. It is a leading source of blockchain and crypto news, meticulously curated sentiment data from January 1st, 2021, to April 4th, 2023. This comprehensive dataset provides a daily reflection of authors' statements regarding Bitcoin across various categories, including podcasts, market updates, and technological insights. Authored by journalists, industry insiders, and subject matter experts, these statements offer invaluable insights into Bitcoin's evolving narrative. Across this timeframe, we diligently collected data reflecting the sentiments expressed by different authors within the specified categories. Each day brought forth a diverse array of opinions, analyses, and forecasts, reflecting the dynamic nature of the Bitcoin ecosystem. From bullish market projections to critical technological developments, the daily statements encapsulate the multifaceted discussions surrounding Bitcoin.

3.4.1 Sentiment Analysis Using VADER

A key component of this research is sentiment analysis[43], which uses the VADER tool to classify the sentiment of news stories taken from CoinDesk. For our research, we used Vader's Sentiment analysis, a robust tool for gauging sentiment in textual data. This analysis allowed us for the calculation of a compound score, indicating the overall sentiment conveyed by each statement. Furthermore, each statement was categorized as neutral, positive, or negative based on its compound score. Examining the sentiment data over the specified timeframe reveals intriguing patterns and trends. Bullish sentiments may surge following significant technological advancements or positive regulatory developments. Conversely, negative sentiments may emerge in response to market volatility or regulatory uncertainty. Furthermore, the diversity of authors contributing to this sentiment data adds richness to the analysis. Journalists provide objective reporting, industry insiders offer insider perspectives, while subject matter experts provide deep insights into the technological intricacies of Bitcoin.

Bullish sentiments may surge following significant technological advancements or positive regulatory developments. Conversely, negative sentiments may emerge in response to market volatility or regulatory uncertainty. Furthermore, the diversity of authors contributing to this sentiment data adds richness to the analysis. Journalists provide objective reporting, industry insiders offer insider perspectives, while subject matter experts provide deep insights into the technological intricacies of Bitcoin. However, amid this wealth of data, instances of missing information arose due to the absence of specific statements from authors on certain days. To address this challenge, we adopted a different approach. Any missing data points were treated as neutral, acknowledging the lack of sentiment conveyed in the absence of statements. A compound score of zero was assigned to these instances, ensuring a balanced interpretation of Bitcoin's sentiment landscape.

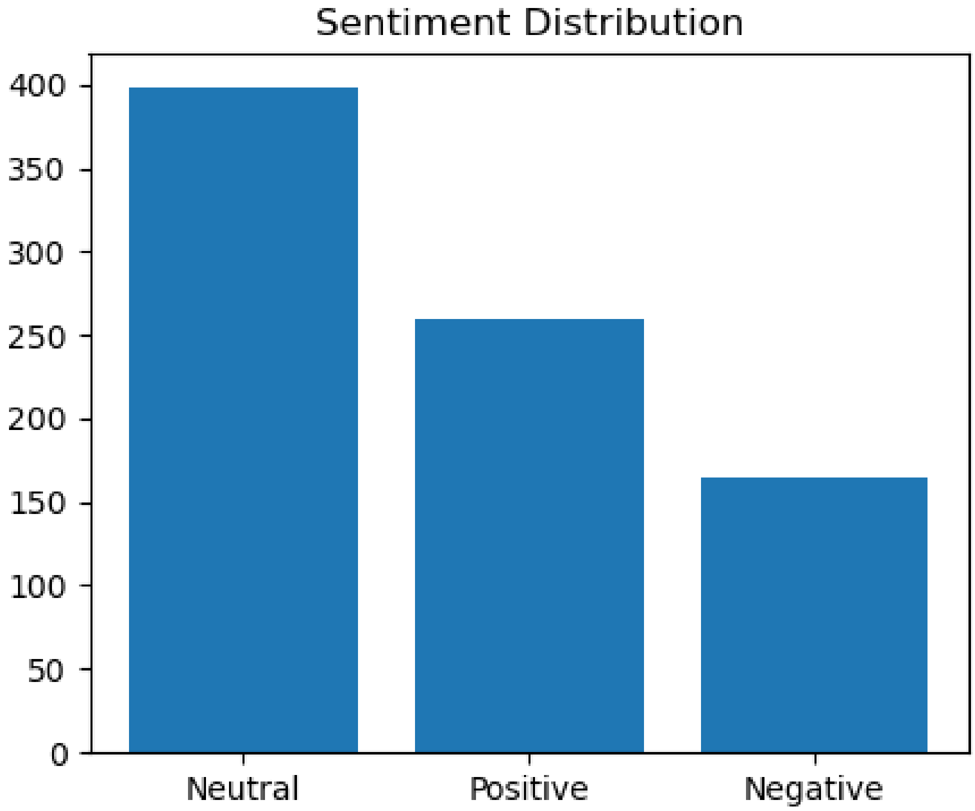

Figure 3.3: Distribution of Sentiment Categories Identified by VADER Sentiment Analysis

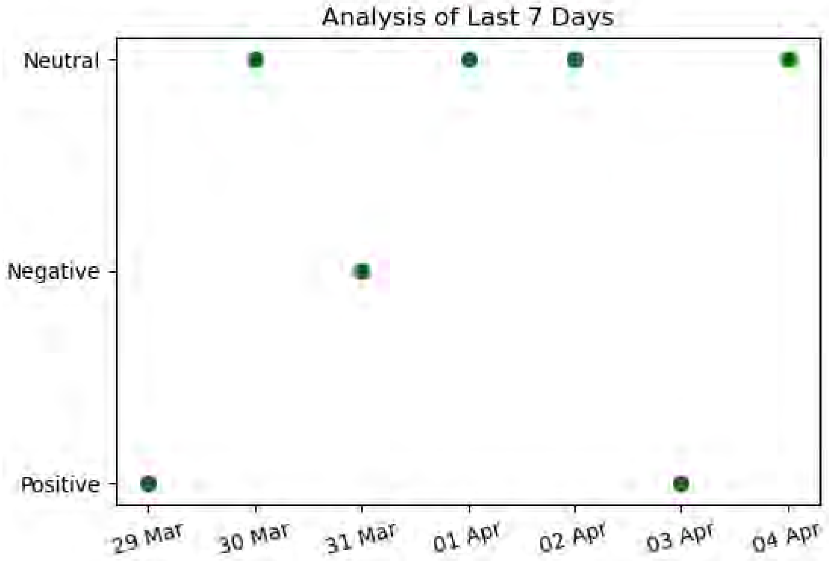

Figure 3.4: Exploring Bitcoin's Sentiment Trends from March to April 2023

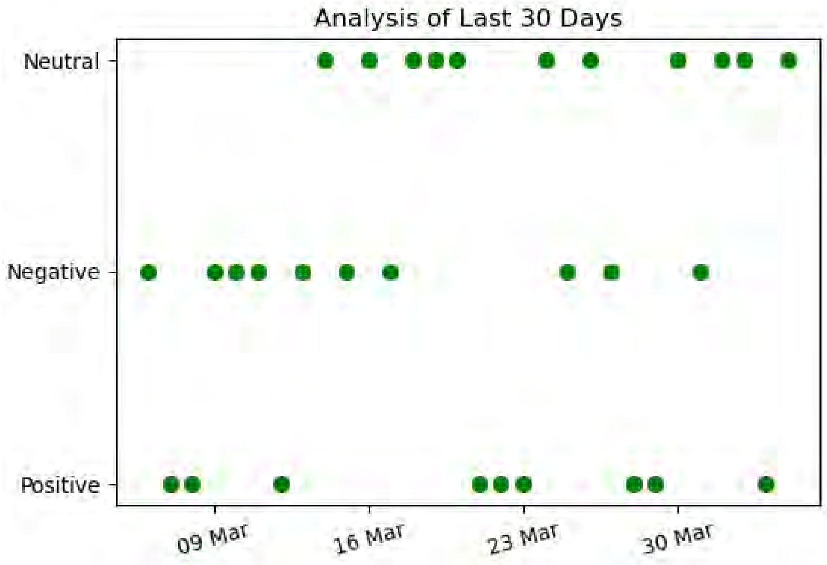

Figure 3.5: Unveiling Bitcoin's Sentiment Trends Leading into March 2023

3.4.2 Data Fusion with Historical Data

The sentiment features derived from VADER analysis are seamlessly integrated with OHLCV (Open, High, Low, Close, Volume) data based on matched publication dates (Table 3.4). This integration enables a holistic analysis framework that incorporates both market data and sentiment signals, empowering predictive modeling and analysis of Bitcoin price movements in relation to sentiment dynamics. By integrating sentiment data with OHLCV data, researchers can explore the intricate interplay between market sentiment, news events, and Bitcoin price dynamics. Whether it's examining the impact of sentiment spikes on short-term price volatility or identifying long-term sentiment trends correlating with market trends, this integrated approach offers a comprehensive understanding of the relationship between sentiment and Bitcoin price movements. This comprehensive methodology ensures a robust and nuanced analysis of Bitcoin sentiment, drawing upon data from CoinDesk, sophisticated sentiment analysis techniques, and integration with market data to offer valuable insights into the complex dynamics shaping the cryptocurrency landscape.

3.4.3 Impact of Sentiment on Price Movements

From Table 3.5, The correlation research over a six-month period shows interesting connections between sentiment measures and market dynamics. First, there are large negative correlations between the sentiment counts—both positive and negative—and the closing price, suggesting that the closure price tends to decline as sentiment rises. On the other hand, a different trend may be seen in the 3-day closing price's % change. This statistic has a positive correlation with positive sentiment counts, suggesting that higher levels of positive sentiment frequently correspond with larger percentage changes in the closing price over a three-day period. Its lesser link with negative sentiment counts, however, points to a less significant influence on short-term price fluctuations. A clear association between sentiment and the target variable is evident from the strikingly perfect positive correlation that the target

Table 3.4: Final Feature List

Features	Description
Williams Percent R (Values)	Technical Indicator
ATR (Values)	Technical Indicator
Bollinger Bands (Signal)	Technical Indicator
EMA (Signal)	Technical Indicator
Keltner Channels (Signal)	Technical Indicator
Close Price	The price of the financial instrument at the end of the time interval
Percentage Change	3 days percentage change of closing price
Compound Score	This is a single normalized score that represents the overall sentiment of the news. It ranges from -1 (extremely negative) to 1 (extremely positive)
Positive Score	The proportion of the news that falls into the positive sentiment category. If positive then 1 else 0
Negative Score	The proportion of the news that falls into the negative sentiment category. If negative then 1 else 0
Neutral Score	The proportion of the news that is considered neutral. If neutral then 1 else 0
Target	A buy signal (1) indicates a bullish trend, a sell signal (-1) indicates a bearish trend, and a hold signal (0) indicates a neutral market condition.

variable (weekly) shows with sentiment counts.

From Table 3.6, Correlation research over a six-month period reveals interesting correlations between sentiment compound scores and market trends. First, there is a significant negative correlation (-0.975096) between the close price and the positive compound score, indicating that the close price tends to decline as the positive sentiment score rises. On the other hand, the negative compound score and the closing price have a substantial positive connection (0.980856), suggesting that higher close prices are associated with more negative sentiment. When analyzing the 3-day closing price percentage change, the negative compound score shows a somewhat negative association (-0.743289), which suggests that the percentage change tends to drop as negative sentiment increases. Conversely, the positive compound score has a positive correlation with the percentage change (0.831411), indicating that greater percentage movements in the closing price over a three-day period correspond with more positive sentiment.

Table 3.5: Analysis of Sentiment Volume and Correlation with Market Prices

Correlation	Period	Positive Sentiment	Negative Sentiment
Close Price	6 Month	-0.949954	-0.986723
Percentage Change of 3 Days Closing Price	6 Month	0.949011	-0.509276
Target	Weekly	1	1

Table 3.6: Correlation of Sentiment Compound Scores with Price Dynamics

Correlation	Period	Positive Compound Score	Negative Compound Score
Close Price	6 Month	-0.975096	0.980856
Percentage Change of 3 Days Closing Price	6 Month	-0.743289	0.831411

These findings emphasize the intricate correlation between emotion compound scores and market dynamics, underscoring the need of taking into account both positive and negative attitudes when examining price fluctuations. Gaining an understanding of these relationships can help you anticipate market movements and make wise financial choices.

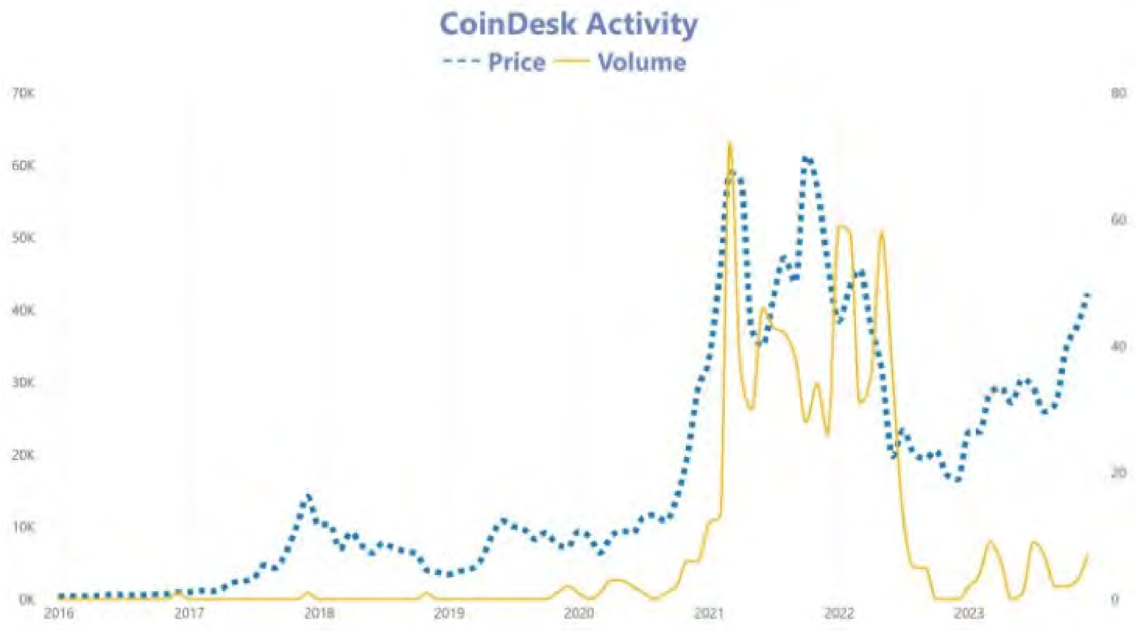

Figure 3.6: Exploring the Bitcoin Price Dynamics Through Sentiment Analysis Over Years

3.5 Model Specification

In this section, we delve into the process of model building and analysis for predicting Bitcoin price movement. Our study focuses on investigating the significance of sentiment features extracted from news articles in enhancing the predictive performance of machine learning models. We explore the effectiveness of various classifiers and their parameter configurations to discern the impact of sentiment on model accuracy and robustness. The dataset utilized in our analysis comprises 647 records, each characterized by 12 features excluding the target feature, obtained through the integration of sentiment data with OHLCV (Open, High, Low, Close, Volume) data. These features encapsulate various aspects relevant to Bitcoin market dynamics, providing a comprehensive basis for predictive modeling. We employ a range of machine learning classifiers to build predictive models for Bitcoin price movement. Specifically, we consider five distinct models: Support Vector Machine (SVM), Random Forest, Logistic Regression, Naive Bayes, and Decision Tree. To optimize model performance, we tune the hyperparameters of each classifier using cross-validated grid search.

3.5.1 Support Vector Classification (SVC)

A supervised learning approach used for classification problems is called Support Vector Classification, or SVC [28]. It is a member of the Support Vector Machine (SVM) family of machine learning models, which are strong and adaptable and able to handle decision boundaries that are either linear or nonlinear. The main goal of SVC is to identify the hyperplane in the feature space that best divides the various classes. In order to maximize the margin—that is, the distance between the hyperplane and the closest data points from each class—also referred to as support vectors, this hyperplane was selected. In order to facilitate the identification of a separate hyperplane, SVC operates by converting the input data into a higher-dimensional space.In order to maximize the margin—that is, the distance between the hyperplane and the closest data points from each class—also referred to as support vectors, this hyperplane was selected. In order to facilitate the identification of a separate hyperplane, SVC operates by converting the input data into a higher-dimensional space. A kernel function is used to carry out this transformation; it efficiently computes the inner products of data points in the higher-dimensional space without the need to explicitly compute the transformation. In SVC, the choice of kernel function is essential since it shapes the decision boundary and, consequently, the model's capacity to represent intricate connections in the data.In order to maximize the margin—that is, the distance between the hyperplane and the closest data points from each class—also referred to as support vectors, this hyperplane was selected. In order to facilitate the identification of a separate hyperplane, SVC operates by converting the input data into a higher-dimensional space. A kernel function is used to carry out this transformation; it efficiently computes the inner products of data points in the higher-dimensional space without the need to explicitly compute the transformation.In SVC, the choice of kernel function is essential since it shapes the decision boundary and, consequently, the model's capacity to represent intricate connections in the data. Sigmoid, polynomial, linear, and radial basis function (RBF) kernels are frequently utilized kernel functions. In actuality, binary classification problems involving small to moderately big datasets are very well-suited for SVC. It has several benefits, including as resistance to overfitting, handling high-dimensional data, and efficiency in capturing nonlinear correlations. The regularization parameter C, which regulates the trade-off between maximizing the margin and reducing the classification error, is one of the key hyperparameters to modify while training an SVC model.The model's performance may also be impacted by the kernel selection and the parameters that go along with it (such as the kernel width for an RBF kernel). All things considered, SVC is a strong and adaptable classification algorithm that can be used to solve a variety of issues, making it an invaluable tool in the toolbox of a machine learning practitioner. The ideal hyperplane that divides the classes in the feature space is the goal of support vector classification. The issue may be expressed mathematically as follows:

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{3.9}$$

$$\text{subject to} \tag{3.10}$$

$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \tag{3.11}$$

$$\xi_i \geq 0, \quad i = 1,\ldots,n \tag{3.12}$$

In this formulation:

- $\mathbf{w}$ represents the weight vector defining the hyperplane.
- C is the regularization parameter controlling the trade-off between maximizing the margin and minimizing the classification error.
- ξ_i are slack variables representing the classification error for each data point.
- y_i represents the class label of the i-th data point ($y_i = \{-1, 1\}$).

- $\mathbf{x}_i$ represents the feature vector of the i-th data point.
- b is the bias term.

The objective function seeks to minimize the norm of the weight vector $\mathbf{w}$ while penalizing misclassifications with the term $C\sum_{i=1}^{n}\xi_i$. The constraints ensure that each data point is correctly classified or lies within a margin of at least 1 from the decision boundary, with slack variables ξ_i allowing for some degree of misclassification. This formulation captures the essence of the SVC algorithm, providing a clear mathematical representation of its optimization problem for classifying data points into distinct categories.

3.5.2 Random Forest

A popular and adaptable machine learning approach for both classification and regression applications is the Random Forest classifier. It is a member of the ensemble learning family, which enhances resilience and predictive performance by combining many independent models. The fundamental principle of Random Forest is to build a large number of decision trees in the training stage and then aggregate their predictions using an averaging (for regression) or voting mechanism (for classification). By using a random subset of the training data and features, each decision tree is trained separately; this technique is referred to as bootstrap aggregating or bagging. Comparing Random Forest to individual decision trees reveals several benefits, such as: Decreased overfitting: Random Forest increases its resilience to noisy or sparse data by reducing the likelihood of overfitting by averaging forecasts from several trees. Enhanced stability: When compared to a single decision tree, Random Forest is often less sensitive to changes in the training data and more stable. By offering a measure of feature relevance, Random Forest enables users to pinpoint the characteristics that have the most influence on the classification process.A Random Forest model's hyperparameters, such as the number of trees in the forest (n_estimators), the maximum depth of each tree (max_depth), and the amount of features taken into consideration for each split (max_features), can be adjusted to maximize performance. Random Forest is extensively utilized in many different fields and applications, including as image classification, natural language processing, finance, and healthcare. When looking for high-performance classification models, many machine learning practitioners choose it because of its robustness, scalability, and user-friendliness. To increase predictive performance, the Random Forest classifier integrates the predictions of many decision trees. The Random Forest classifier's mathematical prediction is represented as follows:
Let $T_1, T_2, \ldots, T_n$ denote the individual decision trees in the forest.
For classification tasks, the predicted class $\hat{y}$ for a given input vector $\mathbf{x}$ is determined by a majority vote of the predictions of the individual trees:

$$\hat{y} = \text{mode}\left(T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_n(\mathbf{x})\right) \tag{3.13}$$

where mode represents the most frequent class prediction among the trees.
For regression tasks, the predicted value $\hat{y}$ for a given input vector $\mathbf{x}$ is determined by averaging the predictions of the individual trees:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} T_i(\mathbf{x}) \tag{3.14}$$

where n is the total number of trees in the forest.
In the training phase, each decision tree T_i is trained independently on a random subset of the training data and features, a process known as bootstrap aggregating (bagging). Additionally, each node in the tree is split based on the best feature among a random subset of features, adding further randomness and reducing overfitting. The hyperparameters of the Random Forest classifier, such as the number of trees in the forest (n_estimators) and the maximum depth of each tree (max_depth), can be tuned to optimize predictive performance. The Random Forest classifier offers several advantages, including reduced overfitting, increased stability, and the ability to measure feature importance, making it a popular choice for various classification and regression tasks.

3.5.3 Logistic Regression

Logistic Regression stands as a cornerstone in the domain of machine learning and statistical modeling, providing a powerful and interpretable framework for binary classification tasks. Despite its name, Logistic Regression is not a regression algorithm but rather a classification method, adept at predicting binary outcomes based on input features. Unlike linear regression, which predicts continuous values, Logistic Regression outputs probabilities bounded between 0 and 1, making it well-suited for scenarios where discrete outcomes are of interest. At its heart, Logistic Regression operates on the principle of modeling the probability that a given sample belongs to a particular class. This is achieved through a mathematical function known as the sigmoid or logistic function. The sigmoid function transforms the linear combination of input features and associated weights into a probability score, effectively mapping the input space onto the interval [0, 1]. The decision boundary separating the two classes is typically set at 0.5, with samples above the threshold being classified as positive and those below as negative. One of the key strengths of Logistic Regression lies in its interpretability. Unlike some black-box machine learning algorithms, Logistic Regression provides transparent insights into the relationship between input features and the target variable. The coefficients associated with each feature indicate the direction and magnitude of their impact on the log-odds of the outcome. This interpretability makes Logistic Regression particularly valuable in applications where understanding the underlying factors driving classification decisions is paramount, such as in healthcare, finance, and social sciences. Moreover, Logistic Regression is known for its robustness and efficiency. It can handle large datasets and high-dimensional feature spaces with relative ease, making it suitable for real-world applications with varying scales of data. Despite its simplicity, Logistic Regression tends to perform well in practice, especially when the relationship between predictors and the response variable is approximately linear. However, like any modeling technique, Logistic Regression has its limitations and assumptions. One notable assumption is the linearity between the log-odds of the outcome and the input features. While this assumption holds true in many cases, it may not capture complex nonlinear relationships present in the data. Additionally, Logistic Regression may struggle with imbalanced datasets, where one class is significantly more prevalent than the other, leading to biased predictions. Despite these limitations, Logistic Regression finds widespread application across diverse domains. In healthcare, it is used for predicting patient outcomes and diagnosing diseases based on medical test results. In finance, Logistic Regression aids in credit scoring and fraud detection by assessing the risk associated with loan applicants or financial transactions. In marketing, Logistic Regression informs customer segmentation and targeting strategies by predicting the likelihood of customer response to marketing campaigns. Logistic Regression stands as a pillar of classification modeling, offering a blend of interpret-ability, robustness, and efficiency. Its simplicity, coupled with its ability to provide actionable insights, makes it an invaluable tool in the data scientist's toolkit. By leveraging Logistic Regression effectively, practitioners can unlock the potential to make informed decisions and drive impact outcomes in a wide range of applications. Logistic Regression is a probabilistic model used for binary classification tasks. Its mathematical formulation can be expressed as follows: Given a set of input features $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the probability that a sample belongs to class $y = 1$ is modeled using the sigmoid function:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{x}-b}} \tag{3.15}$$

where $\mathbf{w}$ represents the weight vector, b denotes the bias term, and e is the base of the natural logarithm.
Similarly, the probability that the sample belongs to class $y = 0$ is given by:

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) \tag{3.16}$$

These probabilities are then used to make binary predictions based on a chosen threshold (e.g., 0.5), where samples with predicted probabilities above the threshold are classified as positive (class 1) and those below as negative (class 0). The parameters $\mathbf{w}$ and b of the Logistic Regression model are typically learned through optimization techniques such as gradient descent or Newton's method. The objective is to maximize the likelihood of the observed data given the model parameters, a process known as maximum likelihood estimation. Logistic Regression offers several advantages, including interpretability, robustness, and efficiency. However, it also has limitations, such as the assumption of linear relationship between features and log-odds of the outcome.

3.5.4 Naive Bayes Gaussian

A probabilistic classifier based on the Bayes theorem and supposing feature independence is called a naive Bayes Gaussian. In spite of its ease of use, the Naive Bayes Gaussian algorithm is a potent and effective tool for classification tasks, especially those involving sentiment analysis, spam filtering, and text categorization. Fundamentally, the Naive Bayes Gaussian algorithm uses Bayes' theorem to determine the conditional probability of a class given a collection of characteristics. In its name, "Gaussian" denotes the presumption that the characteristics' probability distribution is Gaussian, or normal.This assumption increases the algorithm's computing efficiency and makes calculating probabilities easier, particularly when working with continuous-valued data. Estimating the Gaussian distribution's parameters for every feature in every class is the fundamental principle of the Naive Bayes Gaussian algorithm. The mean and standard deviation of the feature values within each class are commonly included in these parameters. Based on the observed feature values and estimated parameters, the classifier can forecast the likelihood that a sample will belong to each class. The feature independence assumption of the Naive Bayes Gaussian is one of its distinguishing traits.This ”naive” assumption suggests that the likelihood of one feature being or not doesn't change depending on the existence or nonexistence of another characteristic. Naive Bayes Gaussian frequently works remarkably well in reality, especially when the features are conditionally independent given the class label, albeit this assumption may not hold true in all real-world cases. The interpretability and ease of use of the Naive Bayes Gaussian algorithm is another benefit. Even those with no background in machine learning may use the classifier since it is simple to comprehend and use. Additionally, Naive Bayes Gaussian resists overfitting, which makes it a good match for noisy or tiny datasets when more intricate models would not perform well. The Naive Bayes Gaussian has drawbacks despite its advantages.In datasets where features are heavily linked, the assumption of feature independence may result in inferior performance. Furthermore, non-Gaussian features may cause difficulties for the algorithm, while this is typically avoidable with the use of feature engineering or other preprocessing methods. Naive Bayes Gaussian is a widely used statistical method in many fields. For example, it is applied in text classification to categorize documents into groups like spam or ham; in sentiment analysis, it is used to predict the sentiment of textual data; and in medical diagnosis, it helps identify diseases based on patient symptoms.The Naive Bayes Gaussian classifier is a straightforward yet efficient algorithm that utilizes the Gaussian distribution and the Bayes theorem to provide probabilistic predictions. Naive Bayes Gaussian is still a useful tool in the machine learning toolbox because it strikes a mix between simplicity, efficiency, and interpretability for classification tasks—even if its assumptions might not always hold true in practice. A probabilistic classifier based on the Bayes theorem and supposing feature independence is called a naive Bayes Gaussian. The following is how it may be stated mathematically:

Given a set of input features $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and a class label y, the conditional probability of observing the feature vector $\mathbf{x}$ given the class label y is modeled as a multivariate Gaussian distribution:

$$P(\mathbf{x}|y) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_y|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T \mathbf{\Sigma}_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)\right) \tag{3.17}$$

where: - $\boldsymbol{\mu}_y$ is the mean vector of the feature values for class y. - $\mathbf{\Sigma}_y$ is the covariance matrix of the feature values for class y. - d is the dimensionality of the feature space. The prior probability $P(y)$ of class y is estimated from the training data as the proportion of samples belonging to class y.

Given a new sample $\mathbf{x}^*$, the posterior probability of class y given $\mathbf{x}^*$ is calculated using Bayes' theorem:

$$P(y|\mathbf{x}^*) = \frac{P(\mathbf{x}^*|y)P(y)}{P(\mathbf{x}^*)} \tag{3.18}$$

where $P(\mathbf{x}^*)$ is the evidence probability, obtained by summing the probabilities of $\mathbf{x}^*$ across all classes. The class label for $\mathbf{x}^*$ is then assigned based on the class with the highest posterior probability. Naive Bayes Gaussian classifier assumes that the features are conditionally independent given the class label, allowing for efficient parameter estimation and classification.

Decision trees are a popular and versatile machine learning algorithm that can be used for both classification and regression tasks. They offer a transparent and intuitive approach to modeling complex relationships in data, making them widely applicable across various domains.

Chapter 4

Results and Discussion

Because algorithms can execute trades quickly and with little human involvement, they are becoming more and more common in financial markets. But before being implemented, algorithmic trading techniques must undergo extensive testing and assessment in order to be successful. Evaluating the feasibility and efficacy of these techniques heavily relies on backtesting, the practice of modeling transactions using past market data. Because of their strong backtesting capabilities, platforms such as Zipline have become more and more popular among researchers and traders in recent years. The importance of Zipline backtesting in improving algorithmic trading techniques is examined in this research.Quantopian created Zipline, an open-source Python backtesting tool that offers a comprehensive and adaptable framework for assessing trading algorithms. Fundamentally, Zipline uses historical market data to replicate transactions, giving customers the ability to evaluate how well their strategies work in different market scenarios. The framework is appropriate for a variety of trading techniques since it covers a multitude of asset classes, such as futures, cryptocurrencies, and stocks. In addition to a comprehensive comparison analysis, the outcomes that our proposed model generated provide valuable insights into the patterns and deviations.

4.1 Performance Matrices

The performance scores for our models that were run were assessed in terms of F1 Score, Accuracy, Precision, Recall, Sharpe Ratio, Cumulative Return, Annualized Return and Maximum Drawdown. In addition, we performed multiple runs of each model to determine the variance in the performance ratings. We utilized the zipline Python library for backtesting [40] the predicted signals and subsequently analyzed the results based on key performance metrics to measure the effectiveness of our approach.

1. One typical statistic used to assess how well categorization models function is accuracy. It may be defined as the proportion of accurate forecasts to all of the model's predictions. Accuracy in mathematics (Acc) can be expressed as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4.1}$$

2. In classification, precision is a statistic used to assess how well a model predicts

positive outcomes. It calculates the percentage of accurately predicted positive cases, or true positive predictions, out of all the instances the model predicts as positive.

Mathematically, precision is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.2}$$

where:

- The cases that the model accurately predicts as positive are known as true positives.
- False Positives are situations that the model predicts as positive while, in reality, they are negative.

3. In classification, recall—also referred to as sensitivity or true positive rate—is a statistic that assesses a model's accuracy in identifying positive cases. It calculates the ratio of all real positive cases to genuine positive forecasts, or positively anticipated instances.

   Mathematically, recall is defined as:

   $$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.3}$$

   where:

   - True Positives are the instances that are correctly predicted as positive by the model.
   - False Negatives are the instances that are incorrectly predicted as negative by the model (they are actually positive).

4. A statistic called the F1 score is used in classification to integrate recall and accuracy into a single number. It offers a balance between the two metrics and is the harmonic mean of accuracy and recall.

   Mathematically, the F1 score is defined as:

   $$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

   where:

   - Precision is the proportion of true positive predictions among all instances predicted as positive.
   - Recall is the proportion of true positive predictions among all actual positive instances.

5. An investment or trading strategy's risk-adjusted return is measured by the Sharpe ratio. It is frequently employed to assess how well investment portfolios are performing. The ratio of the investment's (or portfolio's) excess return above the risk-free rate to the excess return's standard deviation is known as the Sharpe ratio.

   Mathematically, the Sharpe ratio is defined as:

   $$\text{Sharpe ratio} = \frac{R_p - R_f}{\sigma_p} \tag{4.5}$$

   where:

   - R_p is the average return of the investment or portfolio.
   - R_f is the risk-free rate of return.
   - σ_p is the standard deviation of the excess return of the investment or portfolio.

6. An investment's success over a certain time period is assessed using the Cumulative Return. It shows the whole percentage change in the investment's value during the specified time period, taking into account any dividends or interest earned in addition to capital appreciation (or depreciation). The cumulative return (CR) may be computed mathematically as follows:

   $$CR = \left(\frac{P_t}{P_0} - 1\right) \times 100\% \tag{4.6}$$

   where:

   - P_0 is the initial price of the investment.
   - P_t is the price of the investment at time t.

7. The Annualized Return measures the average annual percentage return of the investment over the backtesting period. Mathematically, the annualized return (AR) can be calculated as:

   $$AR = \left(\frac{P_t}{P_0}\right)^{\frac{1}{n}} - 1 \tag{4.7}$$

   where:

   - P_0 is the initial portfolio value.
   - P_t is the portfolio value at the end of the backtesting period.
   - n is the number of years in the backtesting period.

8. The Maximum Drawdown measures the largest drop in the value of the investment from a peak to a trough during a specific period of time. It provides insight into the risk of an investment by quantifying the largest loss experienced. Mathematically, the maximum drawdown (MD) can be calculated as:

$$MD = \max_{i,j:j>i} \left(\frac{P_i - P_j}{P_i} \right) \tag{4.8}$$

where:

- P_i is the portfolio value at time i.
- P_j is the portfolio value at time j.
- The maximum drawdown is the maximum of all relative drawdowns observed between time i and time j, where $j > i$.

4.2 Result Analysis

In the realm of machine learning-based financial modeling, selecting the right algorithms can significantly impact the effectiveness of investment strategies. In our study, we sought to identify the most reliable models for predicting financial outcomes by evaluating several popular algorithms, including Random Forest and Support Vector Machine (SVM)[20]. These models have consistently demonstrated robust performance across various domains, making them compelling choices for our analysis. To ensure a comprehensive comparison, we conducted a comparative analysis involving other notable multivariate models, such as Naive Bayes Gaussian and Logistic Regression[26]. Our objective was to identify the models that excel in different scenarios, particularly concerning the presence or absence of sentiment analysis. The results of our analysis, as summarized in Table 4.1, revealed interesting insights into the performance of different models.

Support Vector Machine emerged as the top performer in scenarios involving sentiment, exhibiting superior test accuracy compared to other models. On the other hand, Random Forest demonstrated exceptional accuracy in sentiment-absent situations, outperforming its counterparts in this context. Beyond accuracy, we also evaluated various performance metrics to gain a holistic understanding of each model's capabilities. These metrics encompassed measures such as precision, recall, and F1-score providing valuable insights into the models' predictive power and robustness. Encouraged by the promising results of our comparative analysis, we proceeded to backtest the two top-performing models—Random Forest and Support Vector Machine—using their predicted data. Backtesting serves as a crucial step in assessing the real-world applicability of machine learning models in financial decision-making. Table 4.2 presents the backtesting findings, showcasing the performance of the two distinct models applied to financial data, with and without sentiment analysis integration. Let's delve deeper into the implications of these results and the factors contributing to the observed performance disparities.

When utilizing the Random Forest model without sentiment analysis integration, our strategy yielded promising annual returns of approximately 3.59%. These returns translated into cumulative gains of 4.726%, indicating the model's effectiveness in generating consistent profits over time. Despite the commendable returns, it's essential to assess the risk associated with the strategy. The Sharpe Ratio, a widely-used measure of risk-adjusted return, stood at 1.95, reflecting a favorable balance between returns and volatility (Figure 4.1).

Table 4.1: Model Performance

Metric	**Support Vector Machine**	**Logistic Regression**	**Random Forest**	**Naive Bayes Gaussian**
Train Accuracy	0.989879	0.959514	1	0.6417
Train Accuracy (with Sentiment)	0.989879	0.959514	1	0.6417
Test Accuracy	0.963636	0.948485	0.978788	0.730303
Test Accuracy (with Sentiment)	0.972727	0.945455	0.89697	0.730303
Precision	0.965412	0.953726	0.97932	0.777284
Precision (with Sentiment)	0.972789	0.949912	0.905383	0.779327
Recall	0.963636	0.948485	0.978788	0.730303
Recall (with Sentiment)	0.972727	0.945455	0.89697	0.730303
F1 Score	0.963562	0.94848	0.978701	0.719489
F1 Score (with Sentiment)	0.97271	0.945383	0.897602	0.718739

Table 4.2: Back Testing Result

Backtest	**Model**	**Annual Returns (%)**	**Cumulative Returns (%)**	**Sharpe Ratio**	**Maximum Draw down (%)**
Without Sentiment	Random Forest	3.59	4.726	1.95	-0.704
With Sentiment	Support Vector Machine	10.112	13.445	2.81	-1.003

However, it's worth noting that the strategy encountered a maximum drawdown of -0.704%, suggesting periods of downturns where significant losses were incurred. The robust performance of the Random Forest model underscores its suitability for generating stable returns in sentiment-neutral market conditions. By leveraging ensemble learning techniques and aggregating predictions from multiple decision trees, Random Forest effectively captures complex patterns in financial data, leading to reliable predictions. In contrast, employing the Support Vector Machine model with sentiment analysis integration resulted in substantially higher annual returns of about 10.112% (Figure 4.2).

The integration of sentiment analysis enhanced the model's predictive capabilities, enabling it to capitalize on market sentiment trends and make more informed investment decisions. The cumulative returns generated by the SVM model reached an impressive 13.445% (Figure 4.3), highlighting the efficacy of incorporating sentiment analysis into financial modeling. By considering not only numerical data but also qualitative information related to market sentiment, SVM demonstrated superior predictive accuracy and profit potential. Moreover, the strategy based on the SVM model exhibited an improved Sharpe Ratio of 2.81, indicating superior risk-adjusted returns compared to the Random Forest approach.

The lower maximum drawdown of -1.003% (Figure 4.4) further reinforces the re-

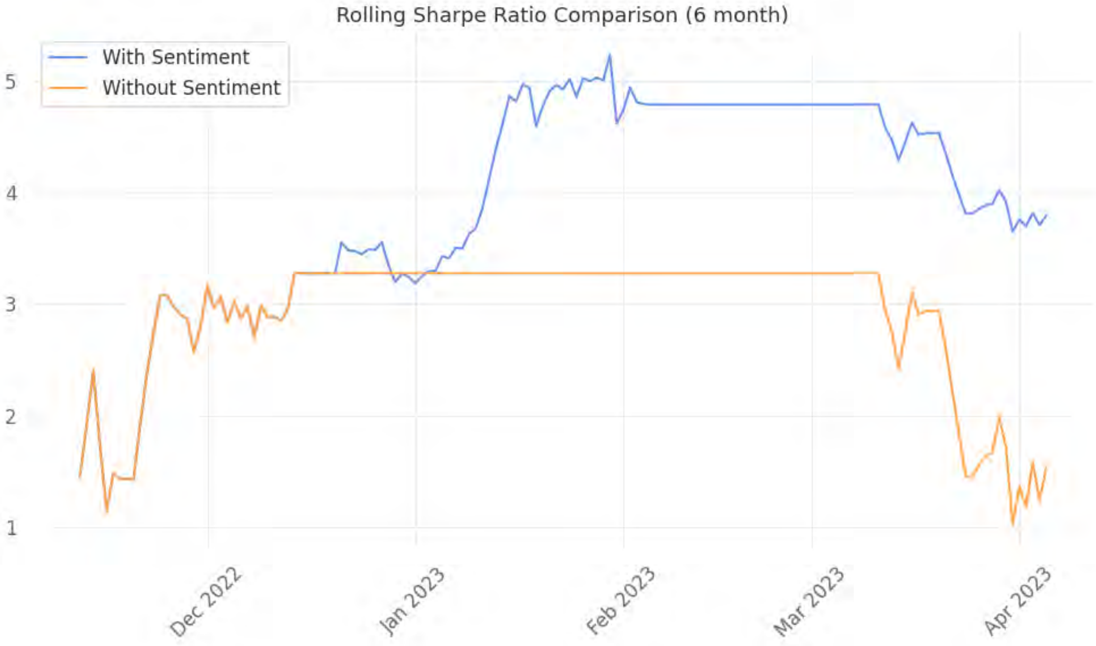

Figure 4.1: Sharpe Ratio Comparison between Strategies with and without Sentiment Analysis

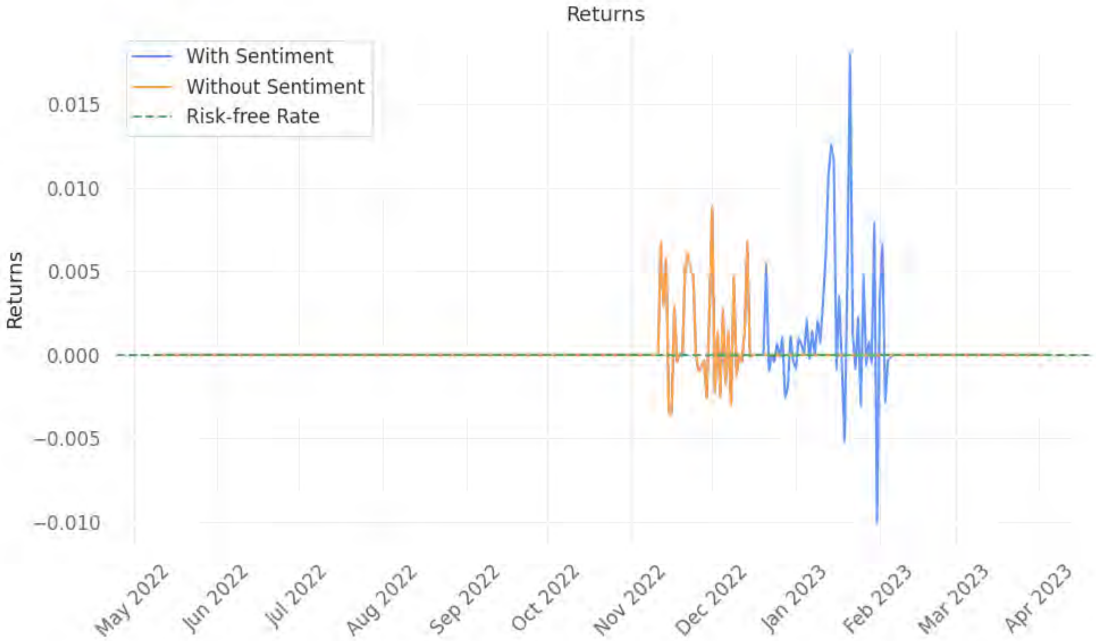

Figure 4.2: Comparative Analysis of Portfolio Returns

silience of the SVM model in mitigating downside risk, thereby enhancing portfolio stability. The significant performance disparities between the Random Forest and SVM models underscore the importance of considering contextual factors, such as market sentiment, in financial modeling. While Random Forest excels in sentiment-neutral environments, SVM leverages sentiment analysis to gain a competitive edge in sentiment-driven markets. The findings from our backtesting exercise highlight

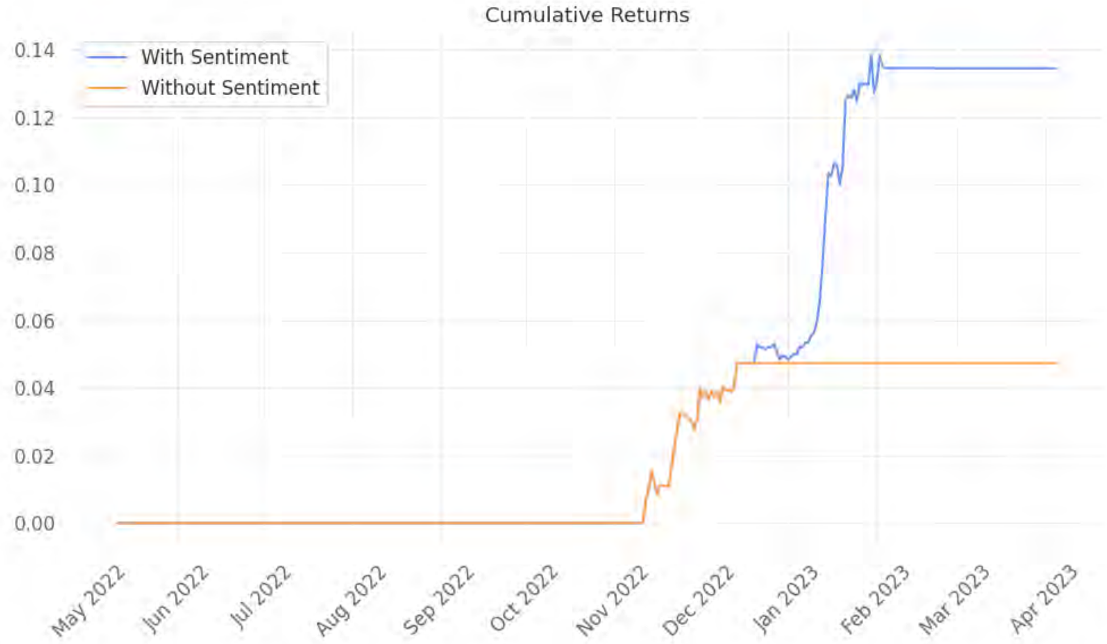

Figure 4.3: Comparison of Cumulative Returns

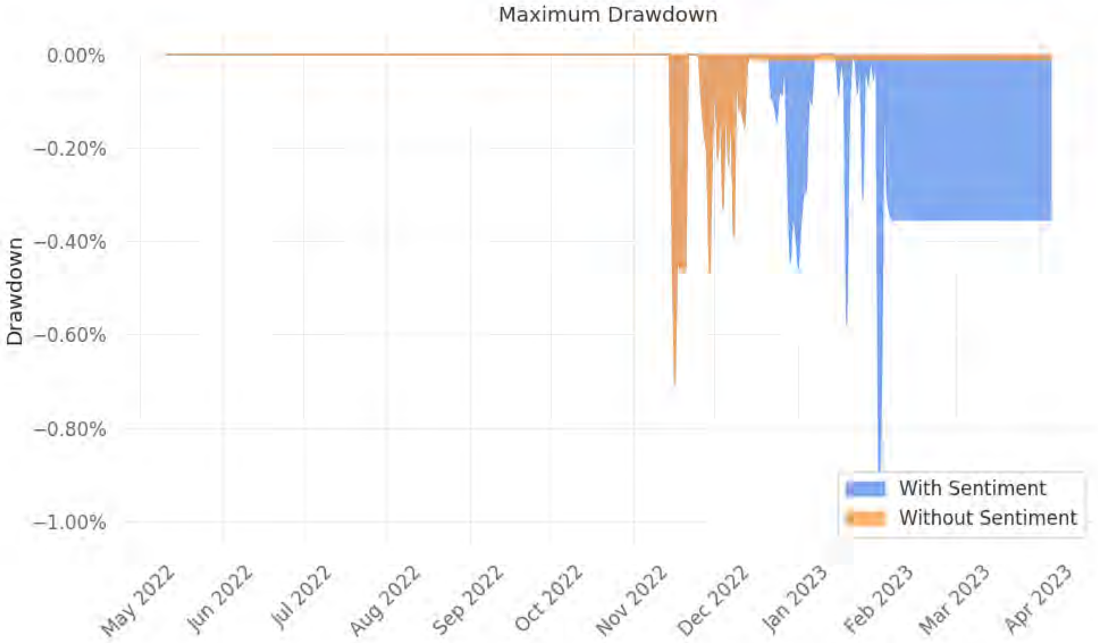

Figure 4.4: Comparative Maximum Drawdown Analysis with Sentiment Analysis Influence

the transformative impact of sentiment analysis on financial modeling and investment strategies. By incorporating sentiment data into machine learning models, investors can gain valuable insights into market sentiment trends and adjust their strategies accordingly. One of the key advantages of sentiment analysis integration is its ability to capture qualitative information that traditional financial metrics may overlook. Market sentiment, driven by factors such as news sentiment, social media

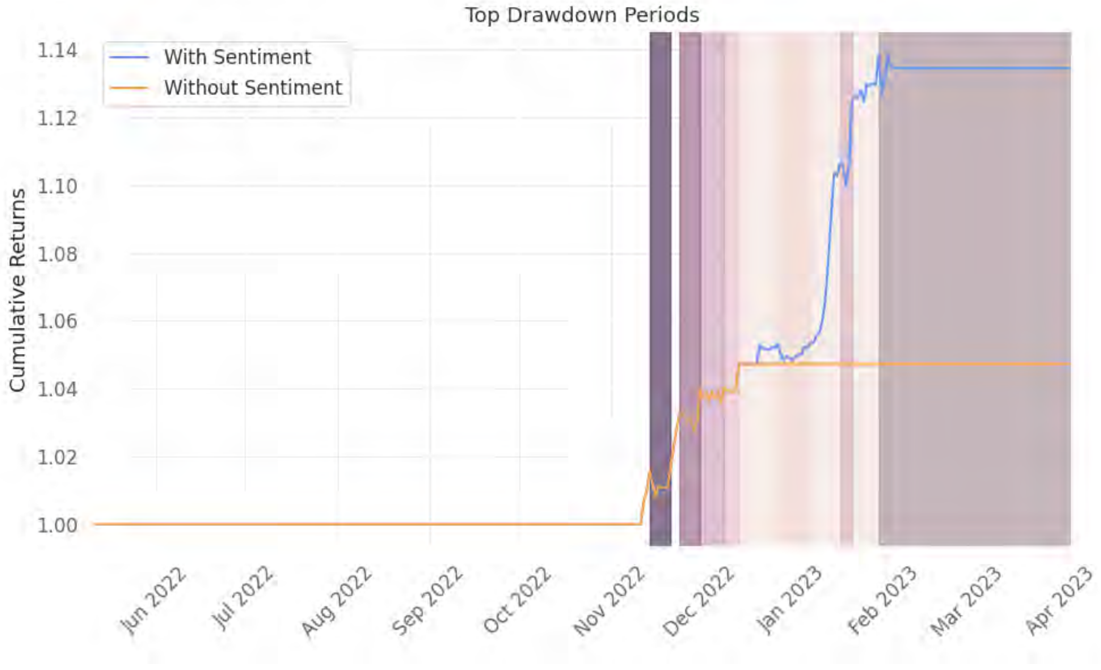

Figure 4.5: Top Drawdown Periods in Backtesting

activity, and investor sentiment, can exert a significant influence on asset prices and market dynamics. By leveraging advanced natural language processing (NLP) techniques, sentiment analysis algorithms can analyze vast amounts of textual data from news articles, social media posts, and financial reports to gauge market sentiment accurately. This nuanced understanding of sentiment enables machine learning models to make more informed predictions and identify profitable trading opportunities. Furthermore, the integration of sentiment analysis enhances risk management practices by providing early warning signals for market downturns and sentiment-driven volatility. By incorporating sentiment-based indicators into risk models, investors can better anticipate market movements and implement proactive risk mitigation strategies.

4.3 Discussion

Table 4.1 shows that the SVM model outperformed the other models in terms of sentiment. All performance indicators, including recall, accuracy, precision, and F1 score, were 97%. On the other hand, Random Forest outperformed other models in the absence of sentiment. Additionally, it received 97 on all criteria, including recall, accuracy, precision, and f1 score. While logistic regression did well among the other two models, it did not perform as well as SVM and RF. Nonetheless, it outperforms the naive Bayes model. The results of Naive Bayes were not very good. Its accuracy of 73% is extremely low when compared to other models. We extracted the anticipated data from these models and backtested it for further analysis. According to Table 4.2, when sentiment was present, the maximum drawdown decreased by 0.2% (Figure 4.4), the annual return rose to 6.5%, the cumulative return was 8.7% (Figure 4.3) and the sharpe ratio was 0.86%. The only indicator that underperformed was maximum drawdown; in terms of emotion, other metrics such as the sharpe ratio, cumulative return, and annual return did well. As a result, these indicators helps in the decision-making of investors. In Figure 4.1, The six-month rolling sharpe ratio rose significantly in terms of sentiment. Figure 4.5 illustrates the highest drawdown periods, which are from January 2023 to February 2023 and from November 2022 to December 2022.

Chapter 5

Conclusion

Our research has demonstrated the effectiveness of utilizing machine learning models, particularly Random Forest and Support Vector Machine (SVM), in predicting financial market movements. Through a comprehensive comparative analysis, we found that SVM performs best when sentiment analysis is incorporated, while Random Forest excels in situations where sentiment data is absent. These models, when backtested, have shown promising results in terms of annual returns and cumulative returns. Specifically, employing the Random Forest model without sentiment analysis yielded notable annual returns of approximately 3.59%, with cumulative returns reaching 4.726%. Despite experiencing a moderate Sharpe Ratio, the strategy demonstrated favorable risk-adjusted returns, albeit with occasional downturns. On the other hand, integrating sentiment analysis with the Support Vector Machine model significantly improved performance metrics. This approach generated substantially higher annual returns of about 10.112%, resulting in cumulative returns reaching 13.445%. Moreover, the strategy exhibited an improved Sharpe Ratio and a mitigated maximum drawdown, indicating superior risk-adjusted returns and lower downside risk. These findings underscore the importance of incorporating sentiment analysis techniques into financial modeling, as evidenced by the substantial enhancements observed in both return metrics and risk management. The integration of sentiment analysis with machine learning models holds significant promise for bolstering investment strategies and optimizing portfolio performance. For future work, several avenues present themselves for further exploration and refinement. Expanding the scope of sentiment analysis to include a broader range of sources, such as social media platforms, news articles, and market sentiment indices, could provide deeper insights into market dynamics. Additionally, exploring advanced machine learning techniques, such as deep learning algorithms, ensemble methods, or reinforcement learning, may further enhance predictive accuracy and robustness. Furthermore, conducting robustness tests and sensitivity analyses across different market conditions and time periods would provide a more comprehensive understanding of model performance. Incorporating real-time data feeds and implementing dynamic model updating mechanisms could improve adaptability to changing market conditions and enhance decision-making capabilities. Overall, continued research in this direction holds promise for advancing the field of financial modeling and improving investment outcomes.

Bibliography

[1] C. Lamon, E. Nielsen, and E. Redondo, “Cryptocurrency price prediction using news and social media sentiment,” *SMU Data Sci. Rev.*, vol. 1, no. 3, pp. 1–22, 2017.

[2] A. Radityo, Q. Munajat, and I. Budi, “Prediction of bitcoin exchange rate to american dollar using artificial neural network methods,” in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, 2017, pp. 433–438. DOI: 10.1109/ICACSIS.2017.8355070.

[3] S. Khuntia and J. K. Pattanayak, “Adaptive market hypothesis and evolving predictability of bitcoin,” *Econ. Lett.*, vol. 167, pp. 26–28, Jun. 2018.

[4] T. Phaladisailoed and T. Numnonda, “Machine learning models comparison for bitcoin price prediction,” in *Proc. 10th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE)*, 2018, pp. 506–511.

[5] M. Wimalagunaratne and G. Poravi, “A predictive model for the global cryptocurrency market: A holistic approach to predicting cryptocurrency prices,” in *Proc. 8th Int. Conf. Intell. Syst., Modelling Simulation (ISMS)*, 2018, pp. 78–83.

[6] I. A. Hashish, F. Forni, G. Andreotti, T. Facchinetti, and S. Darjani, “A hybrid model for bitcoin prices prediction using hidden markov models and optimized lstm networks,” in *Proc. 24th IEEE Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, 2019, pp. 721–728.

[7] A. Inamdar and et al., “Predicting cryptocurrency value using sentiment analysis,” *IEEE Xplore*, May 2019.

[8] R. Sharma, *Do bitcoin mining energy costs influence its price?* Accessed: 2019, 2019.

[9] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, “Price movement prediction of cryptocurrencies using sentiment analysis and machine learning,” *Entropy*, vol. 21, no. 6, p. 589, Jun. 2019.

[10] W. Yiying and Z. Yeze, “Cryptocurrency price analysis with artificial intelligence,” *IEEE Xplore*, 2019. DOI: 10.1109/INFO-MAN.2019.8714700.

[11] S. Dipple, A. Choudhary, J. Flamino, B. K. Szymanski, and G. Korniss, “Using correlated stochastic differential equations to forecast cryptocurrency rates and social media activities,” *Applied Network Science*, 2020. DOI: 10.1007/s41109-020-00259-1.

[12] P. Jay, V. Kalariya, P. Parmar, S. Tanwar, N. Kumar, and M. Alazab, “Stochastic neural networks for cryptocurrency price prediction,” *IEEE Access*, vol. 8, pp. 82 804–82 818, 2020.

[13] Y. Li, Z. Zheng, and H.-N. Dai, "Enhancing bitcoin price fluctuation prediction using attentive lstm and embedding network," *Appl. Sci.*, vol. 10, no. 14, p. 4872, 2020. DOI: 10.3390/app10144872.

[14] Y. Li, Z. Zheng, and H.-N. Dai, "Enhancing bitcoin price fluctuation prediction using attentive lstm and embedding network," *Applied Sciences*, vol. Year 2020, Jul. 2020.

[15] S. Mohapatra, N. Ahmed, and P. Alencar, "Kryptooracle: A real-time cryptocurrency price prediction platform using twitter sentiments," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2020, pp. 5544–5551.

[16] M. M. Patel, S. Tanwar, R. Gupta, and N. Kumar, "A deep learning-based cryptocurrency price prediction scheme for financial institutions," *J. Inf. Secur. Appl.*, vol. 55, Dec. 2020.

[17] G. Serafini, P. Yi, Q. Zhang, *et al.*, "Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2020, pp. 1–8.

[18] V. L. Tran and T. Leirvik, "Efficiency in the markets of crypto-currencies," *Finance Res. Lett.*, vol. 35, p. 101 382, Jul. 2020.

[19] CoinMarketCap, *Today's cryptocurrency prices by market cap*, Accessed: 2021, 2021. [Online]. Available: https://coinmarketcap.com/.

[20] M. A. Ganaie, M. Tanveer, and for the Alzheimer's Disease Neuroimaging Initiative, "Fuzzy least squares projection twin support vector machines for class imbalance learning," *Applied Soft Computing*, vol. 113, p. 107 933, Dec. 2021, Version of Record 22 October 2021. DOI: 10.1016/j.asoc.2021.107933.

[21] M. J. Hamayel and A. Y. Owda, "A novel cryptocurrency price prediction model using gru, lstm and bi-lstm machine learning algorithms," *AI*, vol. 2, no. 4, pp. 477–496, 2021.

[22] J. Kim, S. Kim, H. Wimmer, and H. Liu, "A cryptocurrency prediction model using lstm and gru algorithms," in *Proc. IEEE/ACIS 6th Int. Conf. Big Data, Cloud Comput., Data Sci. (BCD)*, 2021, pp. 37–44.

[23] I. E. Livieris and et al., "An advanced cnn-lstm model for cryptocurrency forecasting," *Electronics*, vol. 10, no. 3, p. 287, 2021. DOI: 10.3390/electronics10030287.

[24] A. Politis, K. Doka, and N. Koziris, "Ether price prediction using advanced deep learning models," in *Proc. IEEE Int. Conf. Blockchain Cryptocurrency (ICBC)*, 2021, pp. 1–3.

[25] S. Tanwar, N. P. Patel, S. N. Patel, J. R. Patel, G. Sharma, and I. E. Davidson, "Deep learning-based cryptocurrency price prediction scheme with inter-dependent relations," *IEEE Access*, vol. 9, pp. 138 633–138 646, 2021.

[26] M. Timilsina, A. Figueroa, M. d'Aquin, and H. Yang, "Semi-supervised regression using diffusion on graphs," *Applied Soft Computing*, vol. 104, p. 107 188, Jun. 2021, Version of Record 23 February 2021. DOI: 10.1016/j.asoc.2021.107188.

[27] A. Gaspar, D. Oliva, S. Hinojosa, I. Aranguren, and D. Zaldivar, “An optimized kernel extreme learning machine for the classification of the autism spectrum disorder by using gaze tracking images,” *Applied Soft Computing*, vol. 120, p. 108 654, May 2022, Version of Record 9 March 2022. DOI: 10.1016/j.asoc.2022.108654.

[28] W. Hilal, S. A. Gadsden, and J. Yawney, “Financial fraud: A review of anomaly detection techniques and recent advances,” *Expert Syst. Appl.*, vol. 193, pp. 9–34, May 2022.

[29] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, “Machine learning techniques and data for stock market forecasting: A literature review,” *Expert Syst. Appl.*, vol. 197, pp. 1–41, Jul. 2022.

[30] V. Akila, N. M. V. S., P. I., S. R. M., and A. K. G., “A cryptocurrency price prediction model using deep learning,” *E3S Web of Conferences*, vol. 391, p. 01 112, 2023. DOI: 10.1051/e3sconf/202339101112.

[31] A. Bouteska, M. Z. Abedin, P. Hajek, and K. Yuan, “Cryptocurrency price forecasting – a comparative analysis of ensemble learning and deep learning methods,” *International Review of Financial Analysis*, 2023. [Online]. Available: https://www.elsevier.com/locate/irfa.

[32] B. Gülmez, “Stock price prediction with optimized deep lstm network with artificial rabbits optimization algorithm,” *Expert Syst. Appl.*, vol. 227, pp. 1–16, Oct. 2023.

[33] O. M. Ahmed and A. M. Abdulazeez, “Comparative analysis of machine learning and deep learning models for bitcoin price prediction,” *Indonesian Journal of Computer Science*, 2024, ISSN: 2549-7286. [Online]. Available: http://ijcs.stmikindonesia.ac.id.

[34] H. Bute, A. Singh, S. Nandurbarkar, S. A. Wagle, and P. Pareek, “Bitcoin price prediction using twitter sentiment analysis,” *International Journal of Intelligent Systems and Applications in Engineering*, 2024, ISSN: 2147-6799. [Online]. Available: http://www.ijisae.org.

[35] F. Fang, W. Chung, C. Ventre, *et al.*, “Ascertaining price formation in cryptocurrency markets with machine learning,” *The European Journal of Finance*, vol. 30, no. 1, pp. 78–100, 2024. DOI: 10.1080/1351847X.2021.1908390. [Online]. Available: https://doi.org/10.1080/1351847X.2021.1908390.

[36] M. Farouk, N. S. Ragab, D. Salam, *et al.*, “Bitcoin_ml: An efficient framework for bitcoin price prediction using machine learning,” *Journal of Computing and Communication*, vol. 3, no. 1, pp. 70–87, 2024.

[37] G. Kaur, P. Agrawal, L. Pinjarkar, *et al.*, “Predictive modeling of bitcoin prices using machine learning techniques,” *International Journal of Intelligent Systems and Applications in Engineering*, 2024, ISSN: 2147-6799. [Online]. Available: http://www.ijisae.org.

[38] M. A. L. Khaniki and M. Manthouri, “Enhancing price prediction in cryptocurrency using transformer neural network and technical indicators,” *Quantitative Finance*, 2024. DOI: 10.48550/arXiv.2403.03606. [Online]. Available: https://doi.org/10.48550/arXiv.2403.03606.

[39] A. Muminov, O. Sattarov, and D. Na, “Enhanced bitcoin price direction forecasting with dqn,” *IEEE Access*, vol. Year 2024, p. 3 367 719, Feb. 2024. DOI: 10.1109/ACCESS.2024.3367719.

[40] M. Parente, L. Rizzuti, and M. Trerotola, “A profitable trading algorithm for cryptocurrencies using a neural network model,” *Expert Syst. Appl.*, vol. 238, pp. 1–13, Mar. 2024.

[41] H. M. Tanrikulu and H. Pabuccu, “The effect of data types’ on the performance of machine learning algorithms for financial prediction,” *arXiv*, vol. 2404.19324, 2024. DOI: 10.48550/arXiv.2404.19324. [Online]. Available: https://arxiv.org/abs/2404.19324.

[42] N. Tripathy, S. Hota, D. Mishra, P. Satapathy, and S. Nayak, “Empirical forecasting analysis of bitcoin prices: A comparison of machine learning, deep learning, and ensemble learning models,” vol. Vol. 15, pp. 21–29, Jan. 2024.

[43] K. Ueda, H. Suwa, M. Yamada, *et al.*, “Sscdv: Social media document embedding with sentiment and topics for financial market forecasting,” *Expert Syst. Appl.*, vol. 245, pp. 1–17, Jul. 2024.

[44] C. Zhao, “Research on prediction of bitcoin price based on machine learning methods,” *SHS Web of Conferences*, vol. 181, p. 02 006, 2024. DOI: 10.1051/shsconf/202418102006.

[45] I. Hashish, F. Forni, G. Andreotti, T. Facchinetti, and S. Darjani, “A hybrid model for bitcoin prices prediction using hidden markov models and optimized lstm networks.”