Bachelor of Science in Computer Science
October 2023

Sentiment-Driven Cryptocurrency Price Prediction

A Comparative Analysis of AI Models

Jammithri Kotapati
Suma Vendrapu

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science. The thesis is equivalent to 10 weeks of full-time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Jammithri Kotapati
E-mail: jako22@student.bth.se

Suma Vendrapu
E-mail: suve22@student.bth.se

University advisor:
Senior Lecturer Lawrence Henesey
Department of Computer Science

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Background:
In the last few years, there has been rapid growth in the use of cryptocurrency, as it is a form of digital currency and was developed using blockchain technology, so it is almost impossible to counterfeit cryptocurrency. Due to these features, it has attracted a lot of popularity and attention in the market. There has been a research gap in predicting accurate cryptocurrency prices by using sentiment analysis. This study will use Artificial Intelligence-based methods and sentiment analysis to develop a model for predicting cryptocurrency prices. By using the mentioned methods in this thesis, the developed model will provide precise results.
Objective:
The objective of the thesis is to compare artificial intelligence models for cryptocurrency price prediction and analyse the importance of sentiment analysis by understanding the public pulse in cryptocurrencies and how it affects price fluctuations, analyzing the correlation within news articles, social media posts, and price fluctuations, as well as evaluating the model performance by employing metrics like RSME, MSE, MAE, and R^2 error.
Methods:
The thesis follows the use of a systematic literature review along with an experimental model for comparing artificial intelligence models. Sentiment analysis played a crucial role in understanding market dynamics. By using linear regression, random forest, and gradient boosting algorithms artificial intelligence models are built to predict cryptocurrency prices using sentiment analysis. The developed models are then compared using performance metrics. This research has analyzed and evaluated each model's performance in predicting cryptocurrency prices.
Results:
The results of the systematic literature review indicated that market sentiment affects cryptocurrency prices. Prices have increased when market sentiment has been positive, whereas prices dropped when sentiment has been negative. The correlation between cryptocurrency values and market mood, however, is complicated as it depends on a variety of factors. Based on the evaluation measures, the random forest artificial intelligence model is the most accurate in predicting cryptocurrency prices after evaluating the three artificial intelligence models.
Conclusions:
This study utilized sentiment analysis and artificial intelligence to forecast cryptocurrency prices. It highlighted the significance of sentiment analysis as a tool for predicting the short-term price of cryptocurrencies by demonstrating how negative sentiment is correlated with decreases in price compared to positive sentiment with price increases. However, it recognized that it was necessary to take into consideration the complexity and broad range of effects on cryptocurrency markets. Research in the future will examine comprehensive sentiment analysis methods and broadening data sources.

Keywords: Artificial Intelligence, Cryptocurrency Prices, News Articles, Sentiment Analysis, Social Media Posts.

Acknowledgments

We are grateful to our supervisor, Lawrence Henesey for his support and encouragement in helping us complete the bachelor thesis, as well as to our parents and friends who encouraged us and gave us the support we needed. We would like to express our gratitude to everyone who helped us finish this project.

Jammithri Kotapati
Suma vendrapu

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence.

BART Binary Auto-Regressive Tree.

BERT Bidirectional Encoder Representations from Transformers.

LSTM Long Short-Term Memory.

MAE Mean Absolute Error.

ML Machine Learning.

MSE Mean Square Error.

NLP Natural Language Processing.

RMSE Root Mean Square Error.

Chapter 1

Introduction

Cryptocurrency is a form of digital money; it has no physical counterpart. As it doesn't have any physical form it is stored in digital wallets only. Cryptocurrency is not issued by the central government. Blockchain technology is used to create a cryptocurrency [35]. Blockchain refers to a network of interconnected blocks that stores data on an online ledger. Every block will contain a set of transactions and will be verified by the validator on the network. Each node must verify the newly created block once it has been generated before it can be authenticated. Therefore, it will be difficult to hack the system or change it in any way that is possible as it is more secure.

Bitcoin was the first cryptocurrency [14]. The launch of bitcoin took place in January 2009. It was created by Satoshi Nakamoto, who is said to be a pseudonym because their real identity is still unknown. The goal of developing and designing bitcoin was to utilize it as a payment method. Despite the fact that there are many cryptocurrencies in the market, bitcoin is the most widely used one. In the cryptocurrency market, bitcoin had been valued at more than 450 billion dollars as of January 2023. Similar to bitcoin, litecoin was developed by Charlie Lie in 2011. It is well known for facilitating transactions quickly and affordably. After bitcoin, ethereum is one of the most widely used cryptocurrencies in 2015. It is a blockchain network that uses the digital currency ETH (Ether) [47].

There are 420 million users of cryptocurrencies, thus there will be a wide range of views on them [44]. This inspired us to investigate the impact of market sentiment on cryptocurrency pricing. External elements like news articles, social media posts, and current affairs have an impact on cryptocurrencies. There is a connection between news articles and the cost of cryptocurrencies [30]. Cryptocurrency is impacted by all of these factors in a positive or negative way. In September 2021, Walmart announced cooperation with litecoin.

Following the announcement, the price of the cryptocurrency increased up to 30 %, but once it became clear that the report was false, the price of the cryptocurrency dropped back down [38]. Being able to grasp the market and make investments with little risk will be quite helpful for a person. We are therefore doing this research and developing an artificial intelligence model to predict future cryptocurrency prices. Cryptocurrency price prediction involves the collection of historical data on cryptocurrency prices and a dataset containing tweets for sentiment analysis. The col-

lection of data is the primary step. For predicting cryptocurrency prices, we have collected historical cryptocurrency price data from investing.com [22]. For sentiment analysis, we have collected the cryptocurrency tweets dataset from kaggle.com [24].

Pre-processing the data is the next step. We eliminated unwanted data from the tweets dataset during this phase [17]. In order to achieve more accurate outcomes, the data is sorted. During data pre-processing, the emojis, URLs, and misspelled words are removed and eliminated from the tweets collected from news articles, and social media posts. Pre-processing the data will help us achieve greater accuracy by reducing overfitting and reducing the training period. To create sentiment scores for tweets, BERT model was used. Then the data will be classified into three categories: positive, negative, or neutral [18].

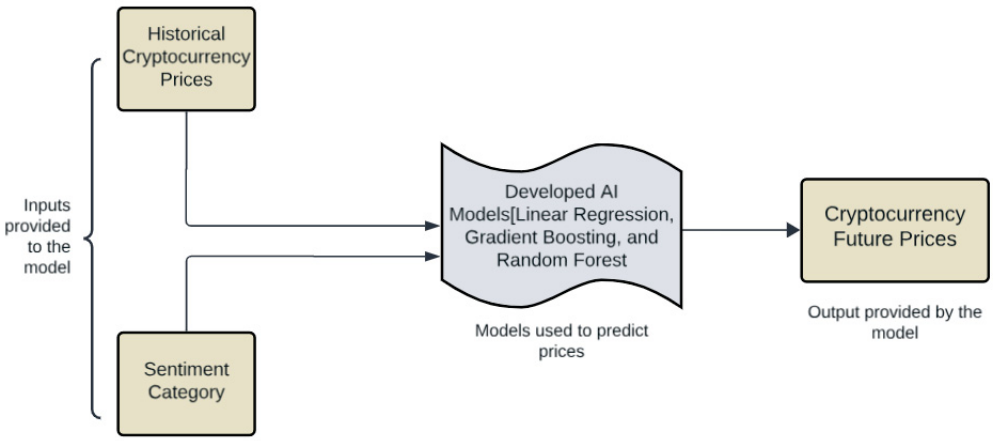


Figure 1.1: The developed model input and output parameters are displayed in the form of a flow chart

After classifying, the tweets dataset will be merged with the cryptocurrency price historical dataset to predict the prices as shown in fig 1.1. This dataset will be trained and tested with a linear regression algorithm, random forest algorithm, and gradient boosting algorithm. The developed models will then predict cryptocurrency prices based on market sentiment. The performance of the model will then be evaluated by the root of mean square error, mean square error, mean absolute error, and R^2 error. The model with the lowest MSE, RMSE, MAE, and highest R^2 error value is therefore considered to be the most effective method for predicting cryptocurrency prices using sentiment analysis. Evaluation metrics are used to compare and identify the most effective model in predicting cryptocurrency prices using sentiment analysis.

1.1 Research Gap

When performing a literature review in the field of sentiment analysis for price prediction, we observed a noticeable gap in this field of research. Mainly, our review

found that numerous studies have used a variety of sentiment analysis methodologies to predict price changes in financial markets. However, there were actually very few studies using the BERT model to categorize tweets into positive, negative, and neutral sentiments.

1.2 Aim and Objectives

1.2.1 Aim

The thesis aims to investigate how sentiment analysis contributes to understanding cryptocurrency-market sentiment and to identify the artificial intelligence model that performs best in predicting cryptocurrency prices when integrated with sentiment analysis.

1.2.2 Objectives

Objective 1: To conduct a comprehensive literature review on sentiment analysis and its relationship with cryptocurrency prices, exploring existing studies, methodologies, and findings in the field.

Objective 2: To gather relevant data from news articles and social media platforms to obtain sentiment-related information, alongside cryptocurrency data for preprocessing purposes.

Objective 3: To classify the collected data into positive, negative, and neutral sentiments, allowing for sentiment analysis of cryptocurrency-related content.

Objective 4: To evaluate and identify the most effective algorithms in conjunction with sentiment analysis and other selected features for improved prediction.

Objective 5: To compare and assess the performance of the developed model with the actual cryptocurrency prices and by using appropriate performance metrics, such as the root of mean square error, mean square error, mean absolute error, and R^2 error, to measure the effectiveness of sentiment-based cryptocurrency price predictions.

1.3 Research Questions

RQ1: How does sentiment analysis contribute to effective trading strategies in the cryptocurrency market?

RQ2: When integrated with sentiment analysis, which artificial intelligence model offers the most accurate predictions for cryptocurrency prices when assessed using metrics like MSE, RMSE, MAE, and R^2 error?

1.4 Ethical, Societal and Sustainability Aspects

As data is collected from news articles and social media posts, the information gathered for sentiment analysis is available to the public. The identity of the user will be anonymous. Investors will have a significant impact by the cryptocurrency price prediction thus, transparency will be maintained. The data will be trained fairly and without any form of bias. Using sentiment analysis, the price of cryptocurrencies is predicted, and it will be utilized ethically by avoiding manipulation and misleading investors. It is essential to take into consideration both the sustainability of the cryptocurrency ecosystem and the environmental impact that mining has on the environment.

1.5 Scope of the thesis

The scope of the thesis investigates the usage of sentiment analysis in the cryptocurrency market, concentrating on short-term price forecasts. Development and comparisons of the model will be the primary focus of this study, along with a literature review, an analysis of its limitations, as well as suggestions for further study. This study will offer guidance to traders and investors in the field of cryptocurrencies.

1.6 Outline

- The Chapter 1, introduction to the basic objective of this research, which involves the prediction of cryptocurrency prices using sentiment analysis.
- The Chapter 2, provides context, that involves a brief overview of the methods with concepts used in the thesis to introduce the concepts to the viewers.
- The Chapter 3, the following part aims to explain the background to the extent of the current research on the topic and presents major results, methods, and concepts taken from previous research.
- The Chapter 4, describes briefly the methodology, research approach, methods to collect data, and processes used to achieve its goals.
- The Chapter 5, of the thesis provides actual results in the form of tables and graphs, that will be assessed in terms of the current concepts and research questions.
- The Chapter 6, works with an exhaustive review of the results, searching for their effects, relevance, and impact on the research questions.
- The Chapter 7, generated several results from the initial research, and we also provided additional components for our research.

Chapter 2

Background

2.1 Cryptocurrency Prediction

Cryptocurrency is a digital or virtual form of currency that is secured by digital keys. To forecast cryptocurrency, numerous elements must be considered, including market trends, technology progress, changes in laws, and investor state of mind. However, it is hard to foresee the crypto market with total certainty. Technical analysis, which involves studying previous price charts and patterns to detect trends, is one method for predicting cryptocurrency. Fundamental analysis is another method for forecasting bitcoin prices. This involves investigating the cryptocurrency's core fundamentals, such as its technology, adoption rate, and regulatory environment. Lastly, by using pre-trained models, these models use artificial intelligence algorithms to analyze historical data and identify patterns that can be used to predict future prices [15].

The integration of sentiment analysis into cryptocurrency trading strategies represents a significant leap forward in understanding market dynamics. By analyzing social media posts, news articles, and other online sources, sentiment analysis can identify the overall positive or negative sentiment towards a particular cryptocurrency. This information can then be used to make more informed predictions about its future price movements [15].

2.2 Sentiment Analysis

Sentiment analysis will focus on analyzing the emotions of the text provided in the dataset. The text will be labeled as positive, negative, or neutral after analysis. NLP and ML are two different AI techniques that are used in sentiment analysis.

NLP converts human language into a structured form that computers can understand. NLP processes and analyzes the text using syntactic methods [2]. The syntactic methods are tokenization, lemmatization, and part-of-speech tagging. The text will then be given to machine learning for text classification after it has been analyzed by NLP. The classified text is labeled as positive, negative, and neutral with scores +1, -1, and 0 respectively to analyze the text sentiment.

There are 4 different types of sentiment analysis models:

1. Fine-grained sentiment analysis: Using this model, we can determine how consumers feel about a specific product [20]. The tweets will be divided up into individual phrases or words. This facilitates a more effective review analysis. Positive, negative, and neutral emotions are all recognized in fine grain. Some examples are:

a. Product Reviews: Analyzing product reviews to determine whether a product is appreciated or not.

b. News analysis: In order to understand how the public feels about the current topic at the present time.

c. Political Analysis: To analyze how the public reacted to a particular policy, rule, or speech given by a politician.

2. Intent-based sentiment analysis: It will help to analyze the difference between the facts and the opinions of the person. In the intent-based model not only the tone can be analyzed but also the intentions of the person from the text [43]. It will understand if the person is raising a complaint, a query or a feedback. Some examples are:

a. Human Resources: The opinions of employees will be collected and analyzed. It will help in enhancing the areas in which they have received negative feedback.

b. Health Sector: Patient input will be collected in order to gain insight into the patient's health and monitor the patient based on the feedback gathered.

c. Marketing Industry: By understanding the intentions of the public the marketing strategies applied will be changed to make them more effective.

3. Aspect-based sentiment analysis: Aspect-based sentiment analysis: In the aspect-based model, the emotions are extracted based on a specific aspect. Understanding public opinion helps in improving things. Some examples are:

a. Health sector: There are specific areas in the health industry, such as patient prescriptions, treatment schedules, and health monitoring, where feedback is gathered for improved patient care.

b. Finance: Understanding the industry's risk, growth, and returns will help the investor make a better choice.

c. E-commerce: Helps in analyzing specific product characteristics, such as the design, pricing, and quality, to determine if the customer will purchase it or not.

4. Emotion-based sentiment analysis: The person's emotional state will be assessed through the text in the emotion-based method. Emotions such as happiness, sadness, fear, and rage will be detected to comprehend the person's emotional state. Some examples are:

a. Feedback from customers: The emotion of the text can be understood by analyzing customer behavior patterns. It is beneficial to understand the customer's perspective on the product.

b. Social media posts: Analysing the texts and comments on social media will help you understand the feelings of the general public.

2.3 Artificial Intelligence

AI has been developed in such a way that systems will be able to carry out tasks that humans can carry out in daily life. The key principles on which AI operates are generalized learning, reasoning, and problem-solving abilities. It has the ability to analyze the situation, adapt to any surroundings, and find solutions for the assigned task. Deep learning and machine learning are subdivisions of artificial intelligence [1]. However, machine learning is a subset of deep learning. To perform the tasks, ML will be trained using a wide variety of datasets. DL will understand the data and its patterns, making it better to analyze the data.

Predicting cryptocurrency is a tremendous challenge for traders, investors, and enthusiasts. In recent years, AI has emerged as a powerful tool for decoding the complex patterns and factors influencing cryptocurrency markets. Techniques that are used in AI to automate predicting cryptocurrency are Data Analysis and Pattern Recognition, Sentiment Analysis, Market Indicators and Data Fusion, Real-time Monitoring, and Alerts [46]. These techniques allow AI to analyze vast amounts of data, such as historical price movements, social media sentiment, and market trends, to identify patterns and make predictions about future cryptocurrency prices. By leveraging AI's ability to process and analyze data quickly and accurately, traders and investors can gain valuable insights that can inform their decision-making and potentially increase their chances of success in the volatile cryptocurrency market.

2.4 Machine Learning

One of the most popular technologies in the field of computer science is machine learning. Machine learning is the ability of machines to learn without being explicitly programmed [33]. Machine learning is a subset of the broader field of artificial intelligence. It is mainly used in sentiment analysis, fraud detection, healthcare, and price prediction. Supervised learning, unsupervised learning, and reinforcement learning are the three categories of machine learning techniques.

Labeled data is used to train supervised learning to make predictions. The labeled data consists of input labels that provide the features of the specific data, and the output labels represent the model's output. Regression and classification are the two types of algorithms used in supervised learning.
In regression algorithm values are predicted based on the trained dataset provided as the input to the model. It is most commonly used in price predictions. The most popular application is in price forecasts. Some of the algorithms used in regression learning include neural networks, decision trees, and linear regression. In the classification algorithm, the data is classified into groups to make predictions. It is used in classifying spam messages, and emails from the received data. Linear regression, random forest, KNN, and SVM are some of the supervised learning algorithms.

Unsupervised learning is trained with unlabelled datasets. It is provided with input labels but not the output of the model. The 2 different types of input that are pro-

vided to the unsupervised model are unstructured and unlabelled. There are two categories of unsupervised learning: clustering and association. In order to categorize the data according to the input given, the clustering method is mainly used. The association algorithm is used to understand the relationship between different parameters. It helps to analyze how one parameter is associated with another parameter. K- means clustering, DBSCAN, BIRCH, and Hierarchical clustering are some algorithms of unsupervised learning.

2.5 NLP

One of the branches of AI is NLP. NLP has gained a lot of popularity because of its ability to convert speech into text that a computer understands. NLP applies computational methods to process and evaluate the speech [10]. Sentiment analysis, detection of spam, and smart voice assistants are where NLP is most frequently utilized.

Tokenization, part-of-speech tagging, named entity recognition, and text classification are NLP approaches. In tokenization, the words are divided up into smaller fragments of text or phrases. Named entity recognition is the process of identifying and sorting the entities based on the events in the text. The words will be labeled with the linguistic components of speech in parts-of-speech recognition. Text classification is the process of categorizing the text into defined topics

In speech-to-text, NLP will translate the speech into text. In sentiment analysis, NLP will classify the text into positive, negative, and neutral. It is also used in computers to interact with humans in the form of chatbots and voice assistants. It will also be useful in machine translation.

2.6 BERT

BERT was introduced by Jakob Devlin in 2018. BERT is a model developed based on the transformer architecture [26]. BERT is the language model that can learn specific tasks when it comes to language. The main advantage of BERT is it can analyze how words are related to each other. BERT has 2 main steps: pre-training and fine-tuning.

In the masked language model, some percent of words from the sentence will be masked and will be given as input to the model. The task of the BERT model is to predict the masked words based on the sentence. Two sentences are provided to BERT and trained for the next sentence prediction model. The goal of BERT is to determine whether or not sentences 1 and 2 are related to one another. The BERT model will be pre-trained in this method.

The developer Jakob defines fine-tuning as a straightforward process. After pre-training, data will be given as input to the BERT model. The pre-trained model

will then tune the dataset. Then it will classify the data into positive, negative, and neutral. If fine-tuning is performed on the BERT it can perform well on different types of language tasks.

2.7 Gradient Boosting Regression

The most popular boosting regression method is gradient boosting regression. The algorithm's primary goal is to continually improve the model while taking into account errors that occurred in earlier iterations of the model [9]. In order to train the new model that is being developed, the error of the prior model is used as a label. With the help of this approach, weak and strong learners are combined, and a new model is created that is trained to minimize the loss function.

To reduce the loss, this technique will update the prior model's negative value to the one that has just been created. The cross-entropy or mean squared error serves as the loss function. A variety of methods, including decision trees and linear models, are employed as the foundation for this technique.

The main feature of this approach is shrinkage. After each predicted amount has been multiplied by learning rates 0 and 1, it is shrunk. By implementing a gradient-boosting approach, bias error will be minimized. This algorithm's key benefits are the model's accuracy and less time taken for prediction. The formula for the gradient boosting algorithm is shown below [36]:
To initialize the model we use the formula:

$$F_o(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{2.1}$$

The formula used after minimizing the errors:

$$F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} 1(x \in R_{jm}) \tag{2.2}$$

where,
L = loss
$argmin_\gamma$ = new value
m = index
γ_{jm} = residual
F_{m-1} = previous step
J = leaves
V = learning speed

2.8 Random Forest Regression

Leo Briman and Adele Cutler invented random forest regression [29]. The term "Random forest regression" refers to the use of several decision trees to perform both

classification and regression tasks [28]. In this algorithm, an ensemble technique is being used. The method used in random forest regression is bagging. Bagging is often referred to as bootstrapping aggregation.

A part of the data is extracted from a large dataset in the bagging process. This data will be used for training and developing the model. Similarly, many subsets of the data are collected and trained. After the training, they will generate a number of models. The ultimate output will be the model that has the majority of the outcomes after all the datasets are trained and developed.

When working with large datasets that have numerous attributes, this algorithm works best. This approach will take care of any missing values, noisy data, and the issue of overfitting [13]. The health industry, finance, and the classification of spam messages are its main uses.

2.9 Linear Regression

Linear regression is used to determine the relationship between a dependent variable and an independent variable. The best-fit line explains the dynamics of their relationship. The linear regression method is one of the straightforward regressions used for analysis and predictions. The best-fit line's goal is to minimize the difference between the predicted and actual values. It is most frequently used in the banking industry to forecast pricing. The formula to calculate the best-fit line is [19]:

$$Y_i = \beta_o + \beta_1 X_i \tag{2.3}$$

where,
Y_i = dependent variable
β_o = constant
β_i = slope
X_i = independent variable

2.9.1 Cost Function

The cost function is used to discover the optimal values for the constant and the slope. It helps in estimating the difference between the projected values and the actual values. The best-fit line is obtained when the residual sum of squares is minimized.The formula is [32]:

$$cost = \frac{1}{N}\sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_o))^2 \tag{2.4}$$

By using the formula we can find the values of the constant and the slope.

2.9.2 Gradient Descent

For determining the best solution, the cost function will be reduced. In order to achieve an ideal solution, the constant and slope values will be modified iteratively. The formula for gradient descent is [27]:

$$\beta_1 = \beta_1 - \alpha \frac{2}{n} \sum_{i=1}^{n} (X_i - y_i).x_i \tag{2.5}$$

2.10 Evaluation Metrics

2.10.1 Root of mean square error(RMSE):

The RMSE calculates the average frequency of the errors in prediction. The accuracy of the model is increased by a lower RMSE value, which indicates that model predictions are more accurate. The formula is [42]:

$$RSME = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2} \tag{2.6}$$

where,
n = observations,
y_i = actual cryptocurrency price at time i,
$\hat{y_i}$ = predicted cryptocurrency

2.10.2 Mean Squared Error(MSE):

The average squared error between the forecasts and the real prices is computed. Lower MSE values represent better accuracy, similar to RMSE.The formula to calculate is: [12]

$$MSE = \frac{1}{n} \sum_{1=1}^{n} (y_i - \hat{y_i})^2 \tag{2.7}$$

2.10.3 Mean absolute error(MAE):

MAE calculates the median absolute variances between values as predicted and actual values. The formula for calculating MAE is: [12]

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}| \tag{2.8}$$

2.10.4 R^2 Error:

R^2 analyzes the relationship between the model's forecasts and the price variation of cryptocurrencies. The formula is: [12]

$$R^2 = \frac{SSR}{SST}\sqrt{\frac{\sum(\hat{y_i} - \vec{y})^2}{\sum(Y_i - \vec{y})^2}} \tag{2.9}$$

SSR = Sum of squared error of regression line,
SST = sum of squared error in mean time.

Chapter 3

Related Work

Lokesh Vaddi [16] conducted a study on predicting cryptocurrency prices. This researcher used different machine learning algorithms and deep learning techniques such as Linear regression, Bayesian regression, Logistic regression, and Support Vector Machine, and they proceeded with the different models called long short-term memory LSTM and recurrent neural networks (RNN) and they used python tensor flow in the code. As a result of which they obtained a 93 percentage of accuracy on the support vector machine for predicting the bitcoin price.

Vasily Derbentsev [45] conducted research on forecasting cryptocurrency prices time series using a machine learning approach. In this research, the researcher used a binary auto-regressive tree BART that is adapted from a standard model of regression trees and the data of the time series. BART combines the classic algorithm and classification of regression trees and autoregressive models(ARIMA). They used these algorithms to make short-term predictions ranging from 5 to 30 days on three cryptocurrencies: bitcoin, ethereum, and ripple. BART outperformed ARIMA in terms of accuracy.

Ioannis E. Livieris [31] did research on ensemble deep learning models for cryptocurrency time-series forecasting. To estimate bitcoin prices, the author used machine learning models such as long short-term memory (LSTM), Bi-directional LSTM, and convolutional layers, as well as regression models. So far, the author has decided to forecast bitcoin prices for one hour. They saw spikes and declines in bitcoin prices over the next hour and used autocorrelation to detect inaccuracies. So far, regression models for predicting bitcoin prices have shown to be more accurate.

Zheshi Chen [11] conducted research on bitcoin price prediction using machine learning: An approach to sample dimension engineering, in this paper author, has chosen different features such as property and network, trading and market, attention and gold spot price, statistical methods including logistic regression and another algorithm have achieved 66 percent of accuracy and then benchmark results for daily cryptocurrency prediction and machine learning models 63 percent accuracy finally, machine learning models outperform statistical methods in terms of accuracy, reaching 67.2 percent.

Sean McNally [34] conducted a study on predicting the price of bitcoin using machine learning. In this work, the author discovered the accuracy of the USD based on

price data using a price data index. The author employed various methods, including a Bayesian-optimised recurrent neural network (RNN) and a long short term memory (LSTM) network. After classification, LSTM had a high accuracy of 52 percent and RMSE had an accuracy of 8 percent. Finally, when both deep learning models were benchmarked on both CPU and GPU with training time, GPU outperformed CPU by 67.7 percent.

Patel Vijay [23]conducted research on Stochastic Neural Networks for Cryptocurrency Price Prediction, and in this study, the author selects a stochastic neural network model for cryptocurrency price prediction. They employed a walkway model for identifying wide cryptomarkets for modeling stock markets. They used long-short term memory (LSTM) and Multi-layer perceptron (MLP) for Ethereum, Bitcoin, and Litecoin price prediction. The results outperform deterministic models.

Ahmed.M [25] conducted research on the forecast of cryptocurrency prices using classic statistics and machine-learning techniques: A survey, the author chooses a standard statistical model such as Bayesian regression, Logistic regression, Linear regression, Support Vector Machine, Artificial Neural Network, Deep Learning, and Reinforcement Learning in this thesis. The author picked different cryptocurrency data from 2010 to 2020, and after predicting the cryptocurrency price, linear regression, and support vector machine performed well when compared to other models, with 98 percent for linear regression and 98.5 percent for support vector machine.

Jethin Abraham [7]did research on the prediction of cryptocurrency prices using tweet volumes and sentiment analysis. In this thesis, the author gathered tweet data from Twitter and Google trends data for Ethereum and Bitcoin during the last two years to forecast cryptocurrency prices, which can be useful for anyone interested in purchasing such cryptocurrencies in the future. The author employed various methods for this thesis, including linear regression, support vector machine, LSTM, and others. When compared to other models, the naive bayes fared best, with 98 percent accuracy.

IliasMaglogiannis [37]investigated the problem of cryptocurrency price prediction using a deep learning approach. The author employed deep learning models such as CNN, ANN, LSTM, and BiLSTM in this thesis. They have cryptocurrency data for bitcoin, Ethereum, and Litecoin dating back 10 months, and they have applied a deep learning model to this data to create the best model for predicting cryptocurrency price predictions. Following this ANN, LSTM outperformed the remaining algorithms with 98 and 95 percent accuracy.

Joseph Bamidele Awotunde [8] has done research on Machine Learning Algorithms for Cryptocurrency Price Prediction. The author conducted research on bitcoin price prediction utilizing deep learning models such as LSTM and BiLSTM, among others, for this thesis. For the past year, they have collected cryptocurrency pricing data for Ethereum, bitcoin, and Litecoin. After completing this thesis, LSTM fared best when compared to other models, achieving 67 percent accuracy.

Zeinab Shahbazi [40] investigated Improving Cryptocurrency Price Prediction Performance Using Reinforcement Learning.The author gathered real-time price data for this thesis in order to predict cryptocurrency prices using reinforcement learning and machine learning models such as linear regression, support vector machine, and naive bayes, and soon for developing the best model for cryptocurrencies such as bitcoin, ethereum, and litecoin. After finishing the project, LSTM received the highest score when compared to the remaining models, with an accuracy of 80 percent.

Mohammad J. Hamayal [21] did a study on A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM, and bi-LSTM Machine Learning Algorithms. This author has employed three recurrent neural network models, including gated recurrent unit (GRU) and long-short-term memory (LSTM), as well as bidirectional LSTM (Bi-LSTM) models, for cryptocurrencies such as Bitcoin (BTC), Litecoin (LTC), and Ethereum (ETC). The models with the lowest mean absolute percentage error (MAPE) have the greatest score. Although GRU receives the highest score when compared to the other models, MAPE scores for LTC, ETC, and BTC are 0.2454,0.8267, and 0.2116 percent, respectively. For the bi-LSTM algorithm, the MAPE score for cryptocurrency is 5.990, 6.85, and 2.332.

Chapter 4

Method

To predict cryptocurrency prices using sentiment analysis qualitative and quantitative approaches have been used. The qualitative and quantitative approaches are beneficial for understanding how social media posts and articles related to cryptocurrencies may affect the price of cryptocurrencies.

4.1 Literature Review

To address research question 1, a qualitative approach has been opted by performing a systematic literature review. The systematic literature review aims to provide a comprehensive overview of the state of the art in cryptocurrency price prediction, with a specific focus on the integration of sentiment analysis and artificial intelligence. We have analyzed various papers and done research on how sentiment analysis affects the factors of the cryptocurrency market.
To conduct this literature review, we accessed various academic databases and online repositories, including IEEE Xplore, Google Scholar, and arXiv. These sources contain a wealth of research papers, articles, and studies related to cryptocurrency price prediction and sentiment analysis. The scope of this literature review encompasses all the key areas. They are as follows:
1. Information Retrieval Strategies: We have identified main keywords or phrases like "cryptocurrency price prediction," "artificial intelligence," "trading strategies," and "sentiment analysis," to search for the publications and research papers over the internet through selected databases that result in a huge list of sources.
2. Reviewing Search Results: The search results were carefully examined after being obtained, with special attention being placed on the titles and abstracts. Publications that did not have a direct connection to sentiment analysis in cryptocurrency trading techniques were eliminated, resulting in a select group of study results.
3. Summarizing Key Findings: The selected sources were carefully reviewed with a focus on identifying major results, techniques, and contributions. Important findings were extracted and cataloged for future reference.
Categorizing the literature: The papers, reviewed by us were systematically categorized in the following manner:
A. Sentiment Analysis in Cryptocurrency Markets:

(i) An investigation of sentiment analysis techniques used in cryptocurrency marketplaces.

(ii) Evaluation of the precision and efficacy of sentiment analysis in predicting market sentiment.

B.Trading Strategies in the Cryptocurrency Market:

(i) A summary of the numerous trading techniques used by bitcoin traders

(ii) Discussion done about the difficulties and possibilities of trading cryptocurrencies.

C. Incorporation of Sentiment Analysis into Trading Strategies:

(i) An examination of how sentiment analysis might help traders to make trading decisions.

(ii) Analysis of the profitability impact of sentiment-based trading techniques

D. Case Evaluations and The Study Results:

(i) Research on sentiment-driven trading methods is presented empirically and through case studies.

(ii) The real-world significance of utilizing sentiment analysis in crypto-trading is discussed.

4. Trends and Methodologies: The study of the literature highlighted trends in the use of natural language processing (NLP) and deep learning models, such as BERT, for sentiment analysis in the cryptocurrency market. In addition, technologies such as reinforcement learning and ensemble techniques are becoming increasingly popular in sentiment-driven trading strategies.

5. Challenges: We discovered that the application of sentiment analysis in cryptocurrency trading has experienced various challenges, including data quality, model accuracy, and the ever-changing nature of social media sentiment, in this study.

4.2 Resources Used in Our Thesis

To address research question 2, a quantitative approach has been opted by performing an experiment. In our thesis, we utilized a wide range of tools to conduct our experiment. These tools included an editor such as visual studio, which allowed us to write and edit our code efficiently. Additionally, we used python libraries such as numpy, pandas, matplotlib, nltk, tensorflow, scikit-learn, and seaborn. These tools helped us gather and interpret the necessary data for our study, ensuring a comprehensive and thorough analysis. Apart from this, we downloaded the tweets dataset from kaggle and the bitcoin cryptocurrency dataset from investing.com. This dataset provided us with a large amount of real-world data to work with, allowing us to validate our findings and draw meaningful conclusions. Our thesis made use of both software and hardware tools, which we have separated into two categories.

4.2.1 Hardware Tools

In the hardware category, we utilised devices such as high-performance computer. These tools enabled us to run our algorithms and models efficiently, ensuring accurate results and reducing processing time. The configurations of our system are outlined in the table below.

Operating System	Windows
Processor	12th Gen i5
RAM (Memory)	8GB
Storage	600 GB

Table 4.1: Hardware Tools

4.2.2 Software Tools

In the software category, we employed various programming languages and software packages to develop and implement our algorithms. These tools allowed us to manipulate and analyze the data, perform statistical calculations, and visualize the results. Additionally, we utilized machine learning libraries to train our models and optimize their performance. The specific software tools we utilized are listed in the table provided below.

1. **Numpy :** Numpy is a widely used library in python for numerical computing and array manipulation. It provided us with efficient data structures and functions to handle large datasets and perform mathematical operations. In our editor, we used the most recent numpy version 1.19.3 to handle data and mathematical calculations [1].

2. **Pandas :** Pandas is another popular library in python that we utilized for data manipulation and analysis. It allowed us to easily handle and analyze structured data, such as CSV files. We used the latest version of pandas, 1.1.4, in our project to efficiently process and manipulate our datasets [2].

3. **Matplotlib :** Matplotlib is a powerful library in python that we employed for data visualization. It provided us with various plotting options to create insightful graphs and charts from our datasets. We utilized the latest version of matplotlib, 3.3.3, to generate high-quality visual representations of our data [3].

4. **NLTK :** NLTK, which stands for Natural Language Toolkit, is a widely-used library in python for natural language processing tasks. It offers a range of tools and resources for tasks such as tokenization, stemming, and part-of-speech tagging. We incorporated NLTK into our project to analyze and process textual data in a meaningful way [4].

[1] https://numpy.org/
[2] https://pandas.pydata.org/
[3] https://matplotlib.org/
[4] https://www.nltk.org/

5. **Tensor flow :** TensorFlow is an open-source machine-learning framework developed by Google. It provides a comprehensive set of tools and libraries for building and training various types of deep learning models. We leveraged TensorFlow in our project to develop and train our neural network models for tasks such as natural language processing [5].

6. **Scikit-learn :** Scikit-learn is a popular machine-learning library in python that offers a wide range of algorithms and tools for tasks such as classification, regression, clustering, and dimensionality reduction. We utilized scikit-learn in our project to implement and evaluate different machine learning models for data analysis and prediction [6].

7. **Seaborn :** Seaborn is a data visualization library in Python that provides a high-level interface for creating informative and visually appealing statistical graphics. We incorporated seaborn in our project to effectively visualize and analyze the results of our machine learning models [7].

4.2.3 Kaggle

Kaggle is a popular online platform for data science competitions and hosting datasets [8]. We used kaggle to access and explore various datasets related to our project. From this platform, we downloaded the datasets for Bitcoin tweets [9]. These datasets provided valuable insights and information that helped us understand the trends and patterns in Bitcoin data.

4.3 Experiment

To tackle RQ2, we needed to employ a statistical approach. We decided to use the experimental research method in the statistical approach to determine which algorithm in linear regression, random forest regression, and gradient boosting to forecast the cryptocurrency is best suited for forecasting bitcoin cryptocurrency. To conduct the experiment, we performed the steps given below.

1. The datasets related to cryptocurrency tweets, social media posts, and news articles are gathered. This dataset is used for sentiment analysis. In addition, we also gathered a different dataset from investing.com that consists of historical bitcoin cryptocurrency prices.

2. Preprocess the given datasets according to the necessary format to train and predict the model, and use the BERT model to do sentiment analysis on the tweet dataset and save it.

[5]https://www.tensorflow.org/
[6]https://scikit-learn.org/
[7]https://seaborn.pydata.org/
[8]https://www.kaggle.com/
[9]https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets

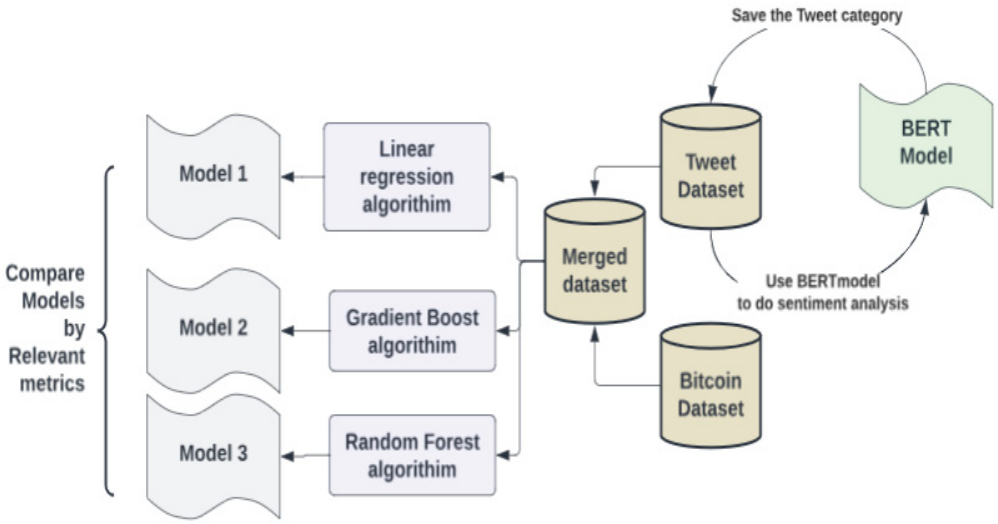


Figure 4.1: The subsequent steps followed to build a robust AI model to predict cryptocurrency prices and to compare the developed models.

3. Create a model using the linear regression algorithm using the Tweet dataset and the bitcoin dataset and save it.
4. Comparable to step 3, but in this step, we built and saved it using the gradient boost algorithm.
5. Comparable to step 3, but in this step, we built and saved it using the random forest regression.
6. Lastly, we compared the three saved model models with the valid evaluation metrics.

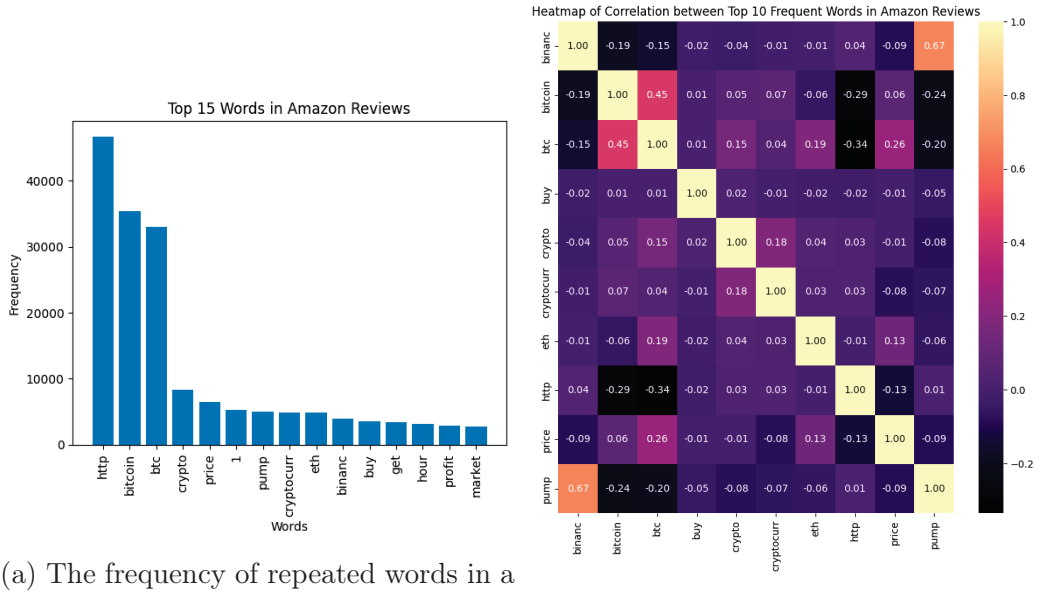
4.4 Data Set Overview

To develop a robust AI model, it is important to have a thorough understanding of the dataset. We had begun searching for datasets from kaggle. After exploring various datasets on kaggle, we were able to find the perfect datasets for our project on Bitcoin tweets and daily transactions of bitcoin. The reason these datasets were chosen was their relevance to our project goals. *Bitcoin tweets* dataset provided us with real-time sentiment analysis of the cryptocurrency, allowing us to evaluate public opinion and market sentiment. On the other hand, *Bitcoin Daily Price dataset* gave us a comprehensive view of the transaction volume and patterns within the Bitcoin network. By combining these two datasets, we were able to gain a holistic understanding of Bitcoin's market dynamics and make informed decisions for our AI model development.

We combined both datasets after doing sentiment analysis using the BERT model. The combined dataset was named *Merge dataset*. This dataset is fed to the linear

regression, random forest regression, and gradient-boosting algorithms, which were trained to forecast the cryptocurrency prices.

To visualize this *tweet_ dataset*, we used data visualization techniques such as plotting bar graphs and heatmaps. To do so, we used the matplotlib library to generate visualizations.



(a) The frequency of repeated words in a tweet and their corresponding word are displayed in a bar graph.

(b) Heatmaps are used to show patterns and relationships between tweets.

Figure 4.2: Overview of *tweet_ dataset*

This *Bitcoin Daily Price* dataset includes bitcoin price history from 2018 to 2023. The information contains daily price data as well as volume, market capitalization, and other relevant variables. We used a line graph to visualize the Bitcoin Daily Price dataset.

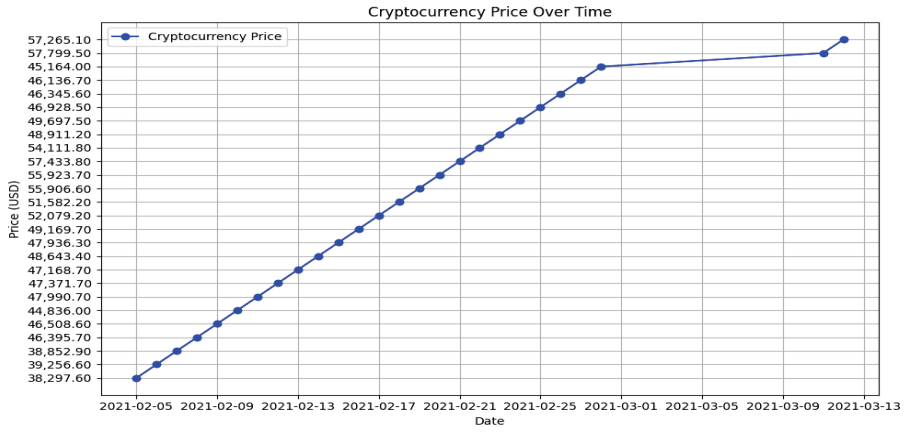
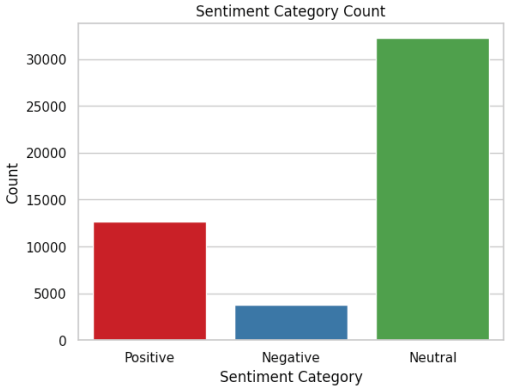


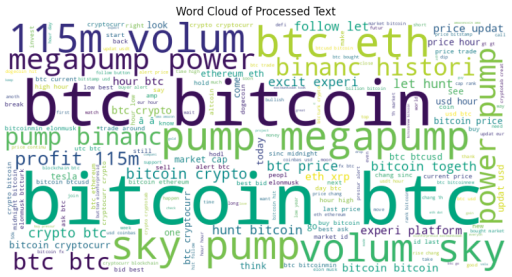
Figure 4.3: In the *Bitcoin_ daily_ price dataset*, a line chart of the bitcoin price from '2021-02 -05' to '2021 -03 -13' is plotted.

We referred to the *Merge dataset* after categorizing sentiment analysis of tweets and merging it with bitcoin cryptocurrency history data. We used bar graphs, a word cloud, and a box plot to visualize our findings.



(a) A Simple bar graph is plotted for combined datasets.

(b) Concurrently, a histogram representation of the same parameters in a bar graph.



(c) After pre-processing a *merged_data_set*, a word cloud is engendered.

Figure 4.4: Overview of *Merged_data_set*

4.5 Data Preprocessing

4.5.1 Column renaming:

In the tweets dataset, the column "text" has been changed to "tweet_text." The dataset will appear more uniform and simpler to use, making it simple to locate the tweets column. The text column will be changed to the tweet_text column by using the line rename(column=’text’: ’tweets_text’).

4.5.2 Converting to lowercase

There are an enormous number of tweets in the dataset that use both uppercase and lowercase letters. All uppercase letters will be changed to lowercase in order to maintain consistency throughout the dataset. All capital words will be converted to lowercase using the str.lower() function.

4.5.3 Tokenization

Tokenization is the process of dividing the sentences into smaller phrases. The sentence's breakdown into words will facilitate a more thorough analysis of the data. The "punkt" package for tokenization has been downloaded from the nlkt resources. The sentence is divided into words by applying apply(word_tokenize).

4.5.4 Removing non-alphanumeric tokens:

Non-alphanumeric tokens will be eliminated from the dataset, including all punctuation, symbols, and numerals. To make the dataset easier to review, it will clean it up by eliminating all non-alphanumeric tokens. All non-alphanumeric tokens are eliminated using the 'token.isalnum' function.

4.5.5 Removing Stopwords

The stopword package is downloaded from nltk resources. The words that are most regularly used are removed from the sentence. To concentrate more on the main words like 'a', 'the', 'is', 'and' are removed from the dataset. It helps to reduce noise and improves efficiency in the dataset to classify the data into categories. By using stopwords.words("english") all the words that are commonly used will be removed.

4.5.6 Removing stemwords:

Words that sound similar to one another will be broken down into their root term. It will improve the accuracy while classifying the data into sentiment categories. The similar phrases are reduced down to their simplest forms using PorterStemmer().

4.5.7 Removal of HTML Tags

HTML tags will be deleted from datasets because they don't help categorize the data and are therefore unnecessary. The data will be cleaned and made easier to process by removing the HTML tags. The HTML tags in the text will be eliminated using removal_html(text).

4.6 Generate Sentiment Analysis by using BERT model

After preprocessing the *Bitcoin Tweets* dataset into the desired form, we use the pretrained BERT (Bidirectional Encoder Representations from Transformers) model for sentiment analysis. BERT is a state-of-the-art language model that can understand the context and meaning of words in a sentence. By utilizing BERT, we are able to accurately analyze the sentiment of each tweet in our dataset, providing valuable insights into the overall market sentiment as shown in fig 4.5.

In order to do sentiment analysis for the preprocessed tweet dataset, we followed the below steps:

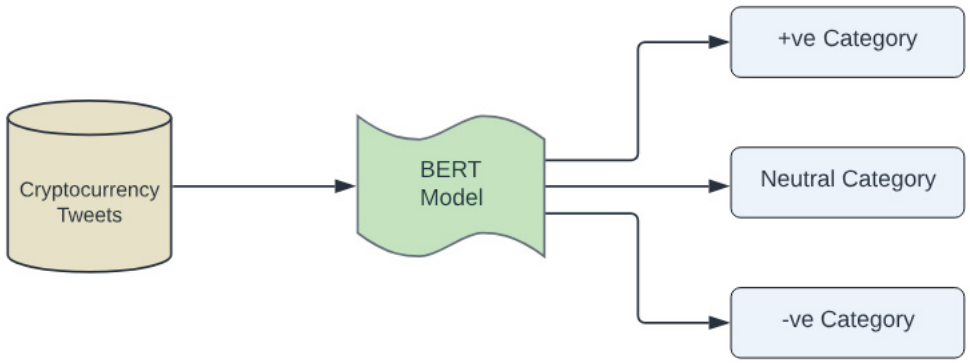


Figure 4.5: The BERT model is used to classify cryptocurrency tweets into three sentiment categories.

1. **Model Construction :** We built the model with TensorFlowHub because, as previously stated, the BERT model is huge and takes some time to load. We can avoid having to load the model every time we wish to utilize it if we build it first.

```
# model construction (we construct model first inorder to use bert_layer's tokenizer)
bert_base = "https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/1"
model = nlp_model(bert_base)
model.summary()
```

Figure 4.6: A code snippet from the Visual Studio editor describes the model setup.

2. **Splitting into Training and Testing Data :**

   a) This step is used to generate training and testing data for a sentiment analysis task that employs BERT. The first four lines of code generate training data by translating the text of the tweets into features that the BERT model can use. The tokenizer divides the text into tokens, and the MAX_SEQ_LEN constant specifies the maximum length of the sequences.

   b) In the same way, the next four lines of code generate the testing data.

   c) The last two lines of code turn the tweet sentiment categories into one-hot encoded vectors. Because the BERT model needs the labels to be in this format, this is required.

3. **Fine Tuning :** The final step was to fine-tune the model using Adam Optimizer, with a learning rate of $2e^{-5}$, a batch size of 10, and a categorical_crossentropy loss function.

```
# create training data and testing data
x_train = convert_sentences_to_features(tweets_data['processed_text'][:1000], tokenizer, MAX_SEQ_LEN)
x_valid = convert_sentences_to_features(tweets_data['processed_text'][1000:1500], tokenizer, MAX_SEQ_LEN)
x_test = convert_sentences_to_features(tweets_data['processed_text'][1500:2000], tokenizer, MAX_SEQ_LEN)
tweets_data['sentiment_category'].replace('positive', 1., inplace=True)
tweets_data['sentiment_category'].replace('negative', 0., inplace=True)
one_hot_encoded = to_categorical(tweets_data['sentiment_category'].values)
y_train = one_hot_encoded[:1000]
y_valid = one_hot_encoded[1000:1500]
y_test =  one_hot_encoded[1500:2000]
```

Figure 4.7: A code snippet from the Visual Studio editor describes the Fine Tunning Process.

4.7 Training and validating Merged data using AI algorithms

After categorizing tweets, we merged both datasets into a single dataset and named *merged_ dataset*. This dataset is trained with AI models using random forest regression, gradient boost algorithm, and linear regression algorithm. This dataset of the first five rows would turn out like this after preprocessing.

	user_followers	user_friends	user_favourites	processed_text	sentiment_category	Price	user_verified_True
0	8534	7605	4838	blue ridg bank share halt nyse bitcoin atm ann...	3	44836.0	0
1	6769	1532	25483	today thursday take 2 friend leowandersleb btc...	3	44836.0	0
2	128	332	924	guy even read articl btc would like share http...	3	44836.0	0
3	625	129	14	btc big chanc billion price bitcoin fx btc crypto	3	44836.0	0
4	1249	1472	10482	network secur 9 508 node today soon biggest be...	3	44836.0	0

Figure 4.8: A code snippet from the Visual Studio editor describes the data frame of a new dataset called *merged_ dataset*.

After merging the dataset, this data set is used for building an AI model with these algorithms. To build a robust AI model, we first split the dataset into 80:20 parts for training and testing. In the training, we again divided training and validation into a 70:10 ratio. The training dataset is used to train the AI model using algorithms such as linear regression, gradient boosting, and random forest algorithms. The validation dataset is then used to fine-tune the model's hyperparameters and ensure its performance. Finally, the testing dataset is used to evaluate the model's accuracy and generalisation ability before deploying it in real-world applications.

4.7.1 Linear Regression Algorithm

The linear regression algorithm aims to find the best-fit line that minimizes the difference between the predicted and actual values. To implement this algorithm in our environment, we performed the below steps :

1. We import the linear regression model from the scikit-learn library for machine learning.
2. Created an instance and selected 'processed text' and 'sentiment_category' features.
3. And we divided the dataset into training and testing with an 80:20 training-to-testing ratio by train_test_split() method from sklearn.model_selection module.
4. Finally, we fitted the model by model.fit() method by x_train and y_train.

4.7.2 Gradient Boosting Algorithm

The Gradient Boosting algorithm is a powerful ensemble method that combines multiple weak learners to create a strong predictive model. We took the following actions to implement this algorithm in our setting:

1. We imported the XGB regressor from the XGBboost module and created an instance.
2. Selected the same features as stated in the linear regression algorithm.
3. And, using the train_test_split() method from sklearn.model_selection module, we split the dataset into training and testing with an 80:20 training-to-testing ratio.
4. Finally, we fitted the model by model.fit() method by using training data.

4.7.3 Random Forest Algorithim

The Random Forest algorithm is a popular ensemble learning method that combines multiple decision trees to make predictions. To use this algorithm in our environment, we did the following:

1. We imported RandomForest Regressor from sklearn.ensemble module and created an instance.
2. Select the same features as stated in both algorithms.
3. And, using the train_test_split() method from sklearn.model_selection module, we split the dataset into training and testing with an 80:20 training-to-testing ratio.
4. Finally, we fitted the model by model.fit() method by using training data.

4.8 Model Evaluation

The evaluation metrics will enable a quantitative assessment of the performance of the models constructed for predicting prices using sentiment analysis. RMSE, MSE, MAE, and R^2 Error are the evaluation metrics used for model comparison. As the predictive model influences investment decisions, it is important to analyze the model. The measurements will help in determining the degree to which the real and predicted values correspond. The cryptocurrency markets tend to be highly unpredictable and are influenced by a variety of parameters. Developing a model that is precise will be beneficial to investors. Analyzing the model's strengths and weaknesses will improve its ability to perform. Based on the evaluation measures, three regression models are compared: linear regression, random forest regression, and gradient boosting regression.

These metrics provide us with different perspectives on the performance of our model. The mean square error measures the average squared difference between the predicted and actual values, while the root mean square error provides a more interpretable measure by taking the square root of the mean square error. The mean absolute error calculates the average absolute difference between the predicted and actual values, and the R^2 value indicates how well our model fits the data compared to a baseline model. By considering these metrics, we can assess the accuracy of the model.

There is a python package called sklearn.metrics that includes ready-to-use functions for calculating the mean square error, root mean square error, mean absolute error, and R^2 value metrics in our model. These functions are simple to integrate into our code; for example, mean square error has a built-in function named *mean_ sqaure_ error()*. We used the numpy library to calculate the square root of MSE to get RMSE, and for MAE, there is a predefined method called *mean_ absolute_ error()*, and for R^2, there is a predefined function called *r2_ scrore()* from sklearn.metrics.

Chapter 5

Results and Analysis

We used a structured research approach to conduct qualitative research for RQ1. The results are summarized in the table. For RQ2, which required comparing algorithms, we used a statistical approach that followed an experimental research methodology. This allowed us to make better comparisons in terms of numbers.

5.1 Literature Review Results

S.No	Title	Findings
1	Sentiment and trading decisions in an ambiguous environment: A study on cryptocurrency traders [2022] [5]	This paper focuses on how sentiment analysis will influence investors' decision-making. The findings showed positive emotion stimulated active trading while negative sentiment led to money withdrawals. The results also demonstrate that emotions can affect credibility, investing decisions, and trading behavior.
2	Sentiment, Google queries, and explosivity in the cryptocurrency market [2022] [6]	This study concentrated on the utilization of statistical and sentiment analysis approaches to identify speculative bubbles in cryptocurrency pricing. The results demonstrate that sentiment analysis will provide a forewarning of price changes.

3	Sentiment-Driven Price Prediction of the Bitcoin based on Statistical and Deep Learning Approaches [2020] [39]	The comparison of ARIMAX and LSTM-based RNN models for predicting cryptocurrency values using sentiment analysis was the main objective of this study. It focused on how sentiment analysis will play an important role in generating useful outcomes. The research in this paper illustrates that sentiment analysis is an important factor in evaluating the cryptocurrency market. It demonstrates that when cryptocurrency prices are predicted using sentiment analysis, the ARIMAX model produces accurate results.
4	Does investor sentiment on social media provide robust information for Bitcoin returns predictability? [2021] [3]	The purpose of this paper was to gain further insight into whether market sentiment influences predictions of bitcoin prices. This research came to the conclusion that there is a statistical connection between market sentiment as well as price volatility. According to their findings, sentiment analysis has a short-term impact on bitcoin prices but has no long-term impact.

5	On-chain analytics for sentiment-driven statistical causality in cryptocurrencies [2022] [4].	Through an understanding of market sentiments and prices, this research paper concentrated on the changing trends of the cryptocurrency market. This thesis explains the significance of classifying sentiment about markets using polarity-specific sentiments. The results of this study demonstrated that market sentiment had an impact on cryptocurrency pricing.
6	Sentiment Analysis Trading Indicators [2020] [41].	This study concentrated on formulating an NLP model using trading regulations and sentiment about the market. The primary objective of the paper is to recognize several elements that affect the bitcoin market. The findings of this study demonstrate that there is a significant connection between price forecasting tweet volume and market emotion for cryptocurrency. This thesis findings show some of the factors that influence cryptocurrency prices are investor feelings, headlines, and legislative modifications.

The literature evaluation showed the impact of sentiment analysis on cryptocurrency pricing. The positive tweets have increased in price while the negative tweets have decreased in price. It also highlighted that fewer research studies classified tweets using the BERT model. The literature review clearly demonstrated that the models linear regression, gradient boosting, and random forest have not been collectively utilized in any research to forecast cryptocurrency prices.

5.2 Experimental Results

We experimented with three different algorithms, namely linear regression, gradient boosting, and random forest, on the *merged_ dataset*. The results are depicted by graphs. The quantities that are being measured in the graph are the accuracy of the predicted prices based on the sentiment of the market. We used mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and R^2 scores to evaluate the performance of these algorithms. The graphs below represent the cryptocurrency actual price along with the predicted prices by the three algorithms.

5.2.1 Linear Regression Model

The actual and predicted prices of cryptocurrencies from May 1st, 2023 to May 7th, 2023 are shown in the graph below.

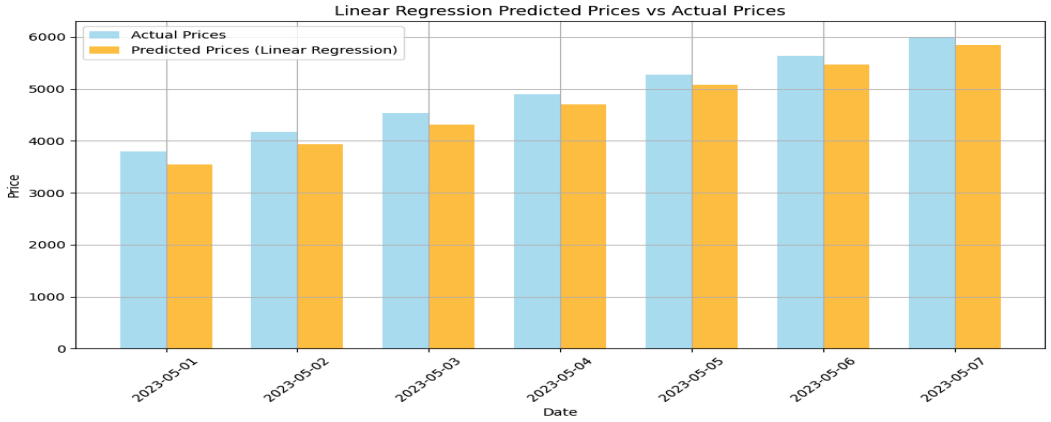


Figure 5.1: Graph displaying the actual prices and the predicted prices by the linear regression model

5.2.2 Gradient Boosting Model

The actual and predicted prices of cryptocurrencies from May 1st, 2023 to May 7th, 2023 are shown in the graph below.

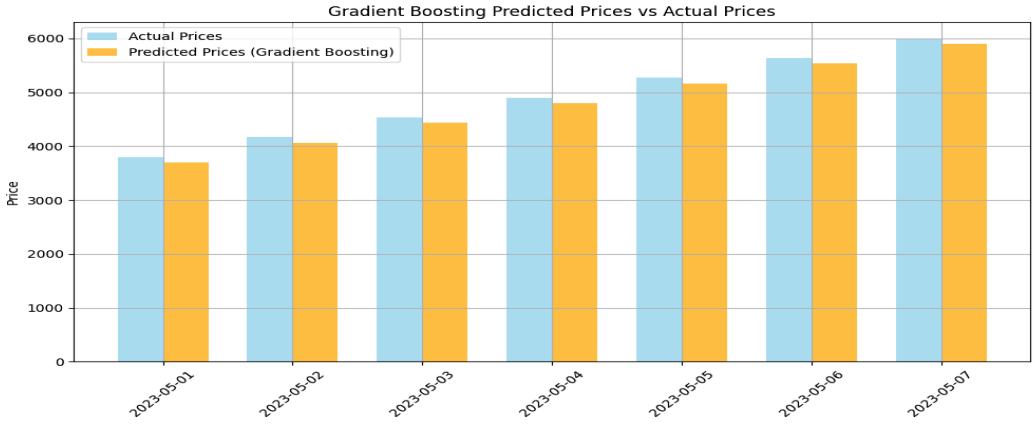


Figure 5.2: Graph displaying the actual prices and the predicted prices by the gradient boosting model

5.2.3 Random Forest Model

The actual and predicted prices of cryptocurrencies from May 1st, 2023 to May 7th, 2023 are shown in the graph below.

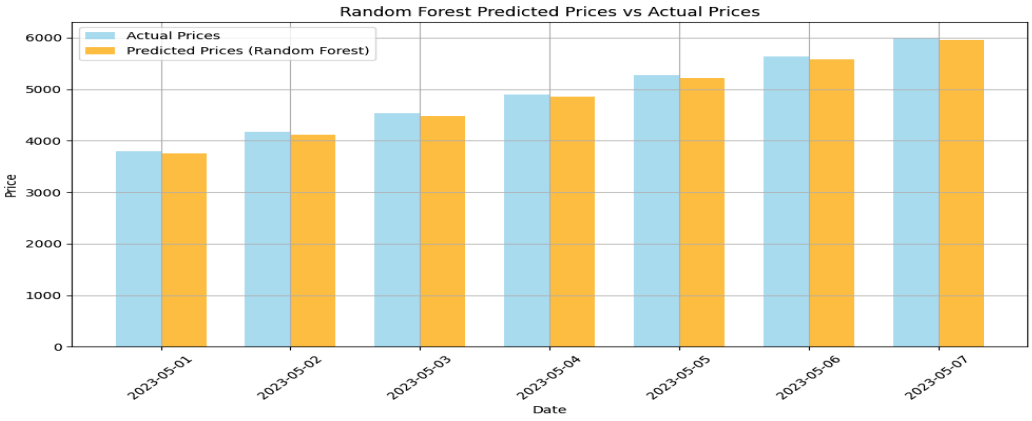


Figure 5.3: Graph displaying the actual prices and the predicted prices by the random forest model

5.2.4 Comparison of the three models

The graph below shows the predicted prices for each of the three models. In comparison to the other two models, the price predictions made by the linear regression model are less accurate. The predictions made by the gradient boosting model performed better than the linear regression model. Compared to the other two models, the random forest model predicts cryptocurrency prices more accurately. Therefore, using sentiment analysis to predict cryptocurrency prices, the random forest model outperformed the others.

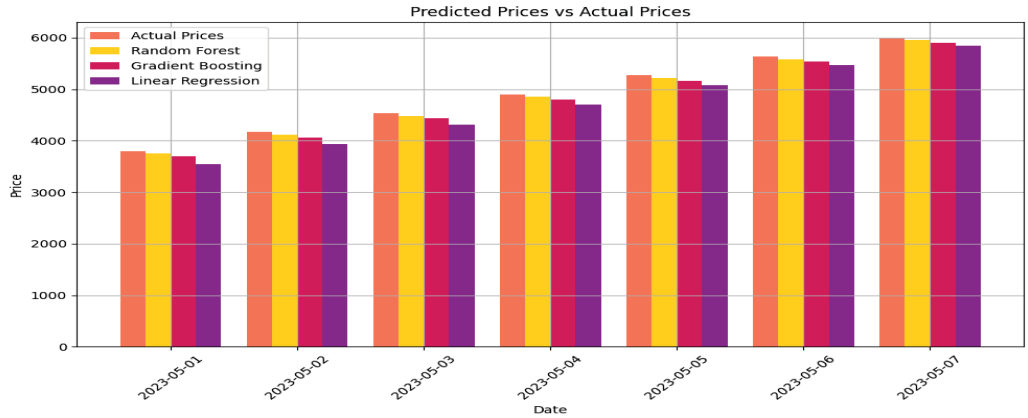


Figure 5.4: Comparison of actual prices and predicted prices by the three models

5.2.5 Comparing three models using the Mean Square Error (MSE) metric

The graph below shows the accuracy of the three models in predicting cryptocurrency prices using MSE. By comparing three models using the mean square error metric. The algorithm with the least MSE value will perform better in predicting cryptocurrency prices. It was noticed that the linear regression algorithm has high values when compared to random forest and gradient boosting algorithms. This suggests that the linear regression algorithm had a lower accuracy in predicting the target variable compared to the random forest and gradient boosting algorithms. The gradient boosting algorithm performed better when compared to the linear regression algorithm but was not accurate in predicting the prices as it has a greater value than the random forest algorithm. The random forest algorithm has the least value when compared to the other two algorithms. So, the random forest model is accurate at predicting cryptocurrency prices.

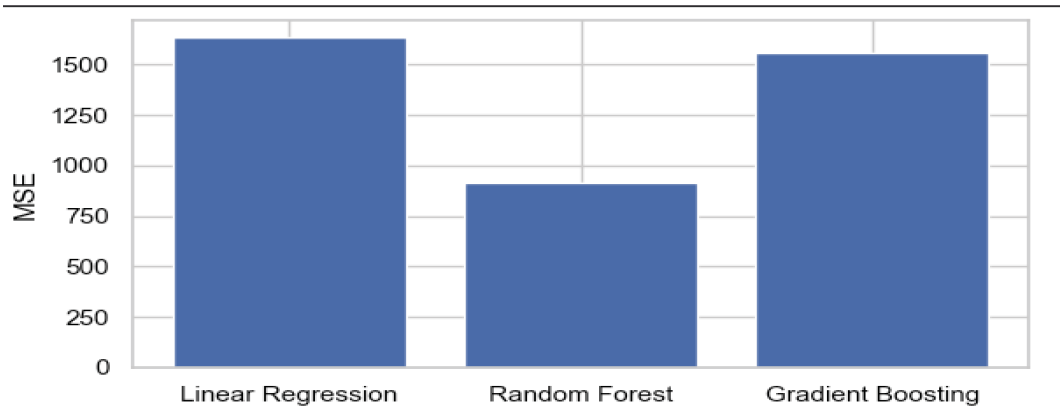


Figure 5.5: Using a bar graph, three different algorithms model accuracy is compared using MSE.

5.2.6 Comparing three models using the Root Mean Square Error(RMSE) metric

The graph below shows the accuracy of the three models in predicting cryptocurrency prices using RMSE By comparing the three models using the root mean square error metric. The algorithm with the least RMSE value will perform better in predicting cryptocurrency prices. It was observed that the linear regression algorithm still had the highest overall error. This indicates that the linear regression algorithm consistently provided less accurate predictions than both the random forest and gradient boosting algorithms. The gradient boosting algorithm has less value compared to linear regression but has greater value when compared to the random forest algorithm. The random forest model has the least value compared to the other two algorithms. Thus, the random forest algorithm is accurate at predicting the prices.

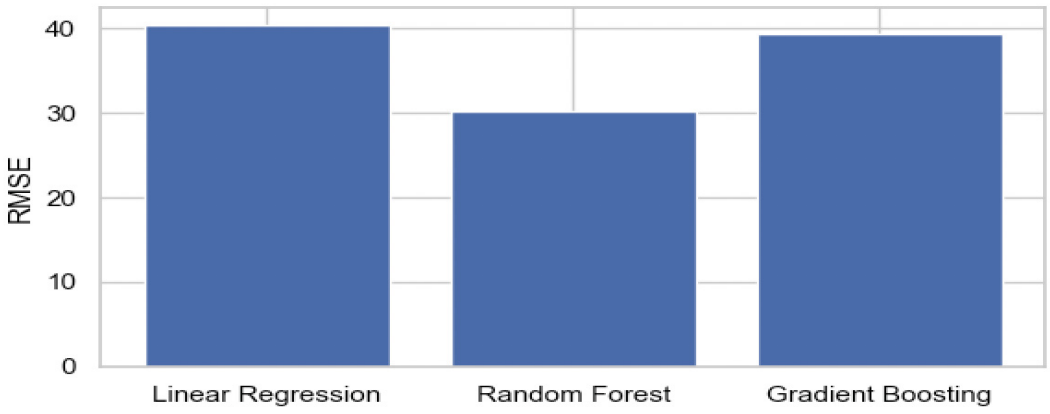


Figure 5.6: Using a bar graph, three different algorithms RMSE scores are compared.

5.2.7 Comparing three models using the Mean Absolute Error (MAE) metric

The graph below shows the accuracy of the three models in predicting cryptocurrency prices using MAE. The algorithm with the least MAE value is effective in minimizing the absolute difference between the predicted and actual values. It is similar to the root mean square graph, with the linear regression algorithm having the highest MAE score. This indicates that the linear regression algorithm is less effective at minimizing the absolute differences between predicted and actual values, making it a more unreliable choice for predicting the prices. The gradient boosting algorithm has less value compared to the linear regression algorithm but has greater value when compared to the random forest algorithm. The random forest algorithm has the least value and it performs best at minimizing the absolute difference between the predicted and actual values.

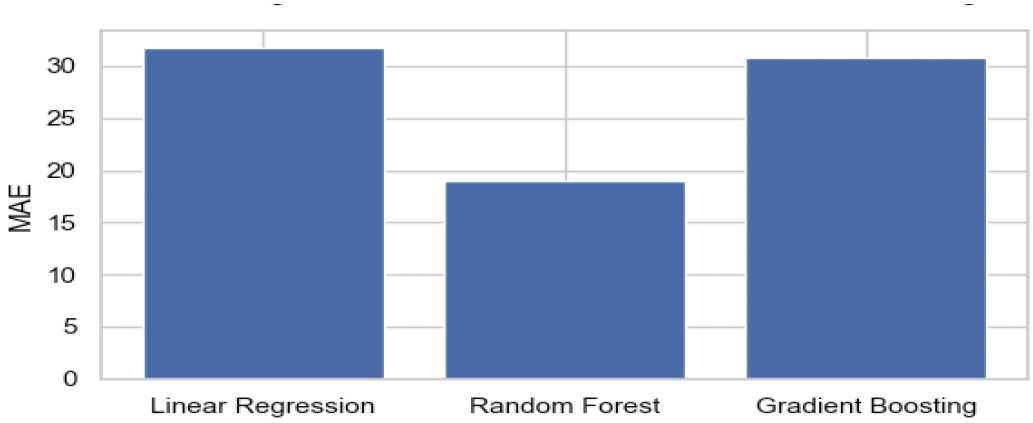


Figure 5.7: Using a bar graph, three different algorithms model accuracy is compared using MAE.

5.2.8 Comparing three models using the R^2 metric

The graph below shows the accuracy of the three models in predicting cryptocurrency prices using R^2 error. The algorithm with the highest R^2 score is accurate at predicting cryptocurrency prices. Comparing three models using the R^2 metric, the linear regression algorithm has the lowest R^2 score. This suggests that the linear regression algorithm performs the least in predicting the prices accurately. The gradient boosting algorithm has a greater value when compared to the linear regression algorithm but less value when compared to the random forest algorithm. The random forest algorithm has the highest R^2 score. So, the random forest model is accurate in predicting cryptocurrency prices.

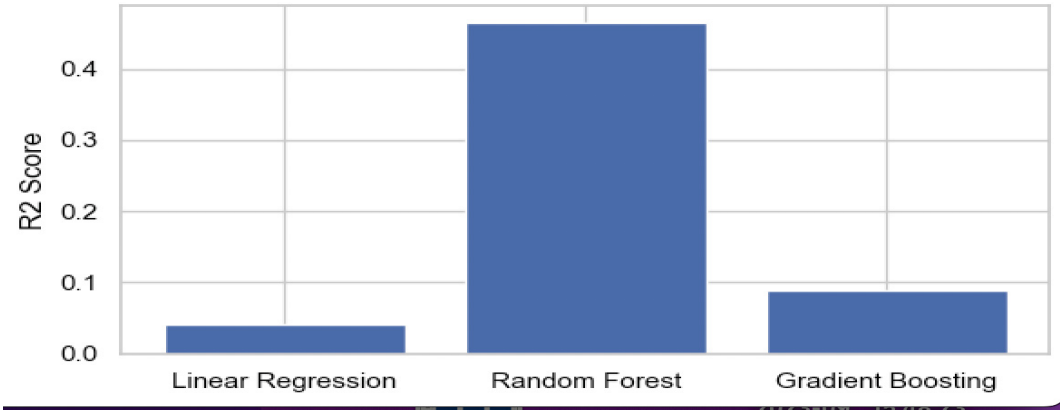


Figure 5.8: Using a bar graph, three different algorithms accuracy is compared using R^2 metric.

5.3 Analysis

The results of the experiment show that the linear regression algorithm performed the worst out of the three algorithms. It had the highest MSE, RMSE, and MAE scores, and the lowest R^2 score. This suggests that the linear regression algorithm is

not a good fit for this dataset.

The random forest algorithm performed the best out of the three algorithms. It had the lowest MSE, RMSE, and MAE scores, and the highest R^2 score. This suggests that the random forest algorithm is a good fit for this dataset.

The gradient boosting algorithm has performed in between the linear regression algorithm and the random forest algorithm. It had MSE, RMSE, and MAE scores that were similar scores to the linear regression algorithm.

Here are some possible reasons why the random forest algorithm performed the best:

1. The random forest algorithm is a non-parametric algorithm, which means that it does not make any assumptions about the distribution of the data. This makes it more robust to non-linear relationships and outliers.
2. The random forest algorithm is a bagging algorithm, which means that it creates multiple models and averages their predictions. This helps to reduce the variance of the predictions.
3. The random forest algorithm is a tree-based algorithm, which means that it can learn complex relationships in the data.

Chapter 6

Discussions

6.1 Research Questions

6.1.1 Research Question 1

How does sentiment analysis contribute to effective trading strategies in the cryptocurrency market?

In cryptocurrency price prediction sentiment analysis plays a crucial role. It will classify the emotions of the tweets into different categories based on the polarity score. It helps in effective trading for the investors. As cryptocurrency is a volatile market sentiment analysis will help in making informed decisions. It will analyze and observe how the public pulse is affecting the cryptocurrency market. As the cryptocurrency market will be changing continuously sentiment analysis is important for trading.

Sentiment analysis will help traders to analyze the factors that affect the market conditions. Some factors that affect the market conditions are government rules, social media sentiment, the supply and demand. As sentiment analysis is not a foolproof strategy it is necessary to be cautious while analyzing the market sentiment. The cryptocurrency markets are highly volatile so sentiment analysis works best for short-term predictions rather than long-term predictions. Traders should not completely depend on sentiment analysis and also have to consider other factor factors along with market sentiment.

6.1.2 Research Question 2

What is the quantitative relationship between news articles and social media comments and the price fluctuations of cryptocurrencies?

To know the quantitative relationship between news articles, social media comments, and variations in cryptocurrency prices. We intend to implement an AI-based model with the optimal algorithm. To determine the optimal algorithm, we chose an experimental methodology. In this methodology, we collected a large dataset of news articles and social media comments related to cryptocurrencies, along with corresponding price data. We then analyzed the data using various algorithms and evaluated their performance in predicting price fluctuations. This approach will pro-

vide valuable insights into the relationship between news, social media sentiment, and cryptocurrency prices, enabling traders and investors to make more informed decisions in this volatile industry.

The algorithms we selected were based on artificial intelligence techniques such as random forest, linear regression, and gradient boosting algorithms. These algorithms were trained on the dataset to identify patterns and correlations between news sentiment, social media comments, and cryptocurrency price movements. By comparing their performance, we can determine which algorithm is most effective in predicting price fluctuations and potentially improving trading strategies in the cryptocurrency market.

Here is the finding of model performance when utilizing the linear regression algorithm by using relevant metrics such as mean square error, root mean square error, mean absolute error, and R^2 metrics scores of 1750, 40, 35, and 0.002. Based on the metrics scores, it appears that the linear regression algorithm has a relatively high mean square error, root mean square error, and mean absolute error, and the R^2 score of 0.002 suggests that the algorithm may not be accurately predicting price fluctuations in the cryptocurrency market. Further analysis and comparison with other algorithms are needed to determine the most effective prediction model.

For the Gradient Boosting algorithm, the scores are 1600, 38, 32, and 0.008. The scores for the Gradient Boosting algorithm indicate that it performs better than linear regression in terms of mean square error, root mean square error, and mean absolute error. However, the R^2 score of 0.008 still suggests that there may be room for improvement in accurately predicting price fluctuations in the cryptocurrency market. Additional evaluation and comparison with other algorithms are necessary to identify the most effective prediction model.

Finally, the algorithm we will investigate in our thesis is the random forest algorithm. The scores of the random forest algorithm in terms of mean square error, root mean square error, mean absolute error and R^2 scores are 850, 30, 18, and 0.475. These scores indicate that the random forest algorithm performs reasonably well in predicting price fluctuations in the cryptocurrency market.

When compared to the scores of these three algorithms, the random forest algorithm outperforms both the linear regression and decision tree algorithms. This suggests that the random forest algorithm may be a more accurate and reliable model for predicting price fluctuations in the cryptocurrency market. Furthermore, the high R^2 score of 0.475 indicates that the random forest algorithm has a high level of precision in its predictions. This suggests that it can provide valuable insights for investors and traders to make informed decisions in the cryptocurrency market.

6.1.3 Reflections

The highlighted research gap is addressed with significant implications by utilizing the BERT model to categorize cryptocurrency-related tweets into positive, negative,

and neutral sentiments, and then using these categorized tweets to predict future cryptocurrency prices. It highlights the tangible value of data-driven decision-making for investors and traders by potentially improving sentiment-based price prediction's forecasting accuracy.

Chapter 7

Conclusions and Future Work

This research concentrated on the exciting field of predicting cryptocurrency prices using sentiment analysis. The cryptocurrency market's significant expansion or instability caused a major concern in developing novel approaches for forecasting changes in price. This study is intended to give significant insights into consumer sentiment and its possible impact on cryptocurrency prices using collecting and evaluating sentiment from an expansive variety of textual data sources, like news articles, social media, and websites. This research proved that sentiment analysis provides useful extra data to identify short-term prices for the cryptocurrency. The negative sentiment was frequently connected with price decreases, and the positive sentiment was usually associated with price increases. By gathering and analyzing sentiment from various textual data sources such as news articles, social media, and websites, researchers were able to study the impact of consumer sentiment on cryptocurrency prices. The findings of this research demonstrated that sentiment analysis offers valuable additional information in determining short-term price movements for cryptocurrencies. Negative sentiment was often correlated with price declines, while positive sentiment was commonly linked to price increase.

In this thesis, we developed a robust AI model that incorporates a random forest algorithm and was trained with both text and price datasets simultaneously. This model allows us to accurately predict short-term price movements based on sentiment analysis of news and social media. By incorporating both textual data and real-time price information, our model can capture the complex dynamics of the cryptocurrency market and provide reliable predictions. Additionally, the random forest algorithm ensures robustness and adaptability to changing market conditions, making it a valuable tool for traders and investors in this volatile industry.

In the future, we would like to develop improved sentiment analysis methods that extend more fundamental direction evaluation. Aspect-based sentiment analysis, for example, could contain deeper data by analyzing sentiment regarding particular features of cryptocurrencies and markets. We would like to use text-based sentiment analysis for various types of data from social networks and newspapers, such as videos and images. So that it gives a more accurate depiction of the public approach. Adding cross-asset sentiment analysis might identify whether sentiments and behaviors within standard markets can impact cryptocurrency price changes, allowing for an overall approach to predicting cryptocurrency price volatility. Simultaneously, the creation of facile applications using AI and sentiment analysis for

specific trading and techniques for risk management has the potential to increase the usage of extensive financial research tools, appealing to a broader range of investors and traders. Providing reliable security methods, inserting precautions in place to prevent influence on markets, and using regulatory guidelines can encourage responsible and suitable usage of algorithms for the field of cryptocurrency price prediction, improving a further visible and accurate financial system.

References

[1] "Artificial intelligence techniques: An introduction to their use for modelling environmental systems," *Mathematics and Computers in Simulation*, vol. 78, no. 2, pp. 379–400, 2008, special Issue: Selected Papers of the MSSANZ/IMACS 16th Biennial Conference on Modelling and Simulation, Melbourne, Australia, 12-15 December 2005.

[2] "Sentiment analysis: A comparative study on different approaches," *Procedia Computer Science*, vol. 87, pp. 44–49, 2016, fourth International Conference on Recent Trends in Computer Science Engineering (ICRTCSE 2016).

[3] "Does investor sentiment on social media provide robust information for bitcoin returns predictability?" *Finance Research Letters*, vol. 38, p. 101494, 2021.

[4] "On-chain analytics for sentiment-driven statistical causality in cryptocurrencies," *Blockchain: Research and Applications*, vol. 3, no. 2, p. 100063, 2022.

[5] "Sentiment and trading decisions in an ambiguous environment: A study on cryptocurrency traders," *Journal of International Financial Markets, Institutions and Money*, vol. 80, p. 101622, 2022.

[6] "Sentiment, google queries and explosivity in the cryptocurrency market," *Physica A: Statistical Mechanics and its Applications*, vol. 605, p. 128016, 2022.

[7] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," *SMU Data Science Review*, vol. 1, no. 3, p. 1, 2018.

[8] J. B. Awotunde, R. O. Ogundokun, R. G. Jimoh, S. Misra, and T. O. Aro, "Machine learning algorithm for cryptocurrencies price prediction," in *Artificial Intelligence for Cyber Security: Methods, Issues and Possible Horizons or Opportunities*. Springer, 2021, pp. 421–447.

[9] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[10] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research [review article]," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[11] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," *Journal of Computational and Applied Mathematics*, vol. 365, p. 112395, 2020.

[12] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021.

[13] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*. Springer, 2012, pp. 278–291.

[14] M. Crosby, P. Pattanayak, S. Verma, V. Kalyanaraman *et al.*, "Blockchain technology: Beyond bitcoin," *Applied Innovation*, vol. 2, no. 6-10, p. 71, 2016.

[15] T.-M. Dulău and M. Dulău, "Cryptocurrency – sentiment analysis in social media," *Acta Marisiensis. Seria Technologica*, vol. 16, pp. 1 – 6, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:210076114

[16] V. D'Amato, S. Levantesi, and G. Piscopo, "Deep learning in predicting cryptocurrency volatility," *Physica A: Statistical Mechanics and its Applications*, vol. 596, p. 127158, 2022.

[17] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015, vol. 72.

[18] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the first ACM conference on Online social networks*, 2013, pp. 27–38.

[19] J. Groß, *Linear regression*. Springer Science & Business Media, 2003, vol. 175.

[20] X. Guo, W. Yu, and X. Wang, "An overview on fine-grained text sentiment analysis: Survey and challenges," in *Journal of Physics: Conference Series*, vol. 1757, no. 1. IOP Publishing, 2021, p. 012038.

[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] investing, "investing ," investing.com, 2021. [Online]. Available: https://www.investing.com/

[23] 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.

[24] Kash, "Bitcoin Tweets," Kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets

[25] A. M. Khedr, I. Arif, M. El-Bannany, S. M. Alhashmi, and M. Sreedharan, "Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey," *Intelligent Systems in Accounting, Finance and Management*, vol. 28, no. 1, pp. 3–34, 2021.

[26] M. V. Koroteev, "Bert: A review of applications in natural language processing and understanding," 2021.

[27] J. Lakshmi, "Stochastic gradient descent using linear regression with python," *International Journal on Advanced Engineering Research and Applications*, vol. 2, no. 7, pp. 519–524, 2016.

[28] X. Li *et al.*, "Using" random forest" for classification and regression." *Chinese Journal of Applied Entomology*, vol. 50, no. 4, pp. 1190–1197, 2013.

[29] Y. Li, C. Zou, M. Berecibar, E. Nanini-Maury, J. C.-W. Chan, P. Van den Bossche, J. Van Mierlo, and N. Omar, "Random forest regression for online capacity estimation of lithium-ion batteries," *Applied energy*, vol. 232, pp. 197–210, 2018.

[30] Y. Liu and A. Tsyvinski, "Risks and returns of cryptocurrency," *The Review of Financial Studies*, vol. 34, no. 6, pp. 2689–2727, 2021.

[31] I. E. Livieris, E. Pintelas, S. Stavroyiannis, and P. Pintelas, "Ensemble deep learning models for forecasting cryptocurrency time-series," *Algorithms*, vol. 13, no. 5, p. 121, 2020.

[32] F. F. Lubis, Y. Rosmansyah, and S. H. Supangkat, "Gradient descent and normal equations on cost function minimization for online predictive using linear regression with multiple variables," in *2014 International Conference on ICT For Smart Society (ICISS)*. IEEE, 2014, pp. 202–205.

[33] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: an applied review," *International Journal of Remote Sensing*, vol. 39, no. 9, pp. 2784–2817, 2018.

[34] S. McNally, J. Roche, and S. Caton, "Predicting the price of bitcoin using machine learning," in *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*. IEEE, 2018, pp. 339–343.

[35] M. Milutinović *et al.*, "Cryptocurrency," - , no. 1, pp. 105–122, 2018.

[36] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[37] E. Pintelas, I. E. Livieris, S. Stavroyiannis, T. Kotsilieris, and P. Pintelas, "Investigating the problem of cryptocurrency price prediction: a deep learning approach," in *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 99–110.

[38] L. Rognone, S. Hyde, and S. S. Zhang, "News sentiment in the cryptocurrency market: An empirical comparison with forex," *International Review of Financial Analysis*, vol. 69, p. 101462, 2020.

[39] G. Serafini, P. Yi, Q. Zhang, M. Brambilla, J. Wang, Y. Hu, and B. Li, "Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[40] Z. Shahbazi and Y.-C. Byun, "Improving the cryptocurrency price prediction performance based on reinforcement learning," *IEEE Access*, vol. 9, pp. 162 651–162 659, 2021.

[41] K. Singpurwala, "Sentiment analysis trading indicators," July 2021.

[42] P. Sokkhey and T. Okazaki, "Hybrid machine learning algorithms for predicting academic performance," *Int. J. Adv. Comput. Sci. Appl*, vol. 11, no. 1, pp. 32–41, 2020.

[43] P. Sudhir and V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 205–211, 2021.

[44] Triple.A, "Wide Range of Users ," triple-a.io, 2021. [Online]. Available: https://triple-a.io/crypto-ownership-data/#:~:text=As%20of%202023%2C%20we%20estimated,420%20million%20crypto%20users%20worldwide

[45] L. Vaddi, V. Neelisetty, B. C. Vallabhaneni, and K. B. Prakash, "Predicting crypto currency prices using machine learning and deep learning techniques," *Int. J*, vol. 9, no. 4, 2020.

[46] Viktor Manhov, Hanxiong Zhang, "Forecasting Cryptocurrency Markets using Artificial Intelligence and Machine Learning Tools," www.frontiersin.org, 2022. [Online]. Available: https://www.frontiersin.org/research-topics/37594/forecasting-cryptocurrency-markets-using-artificial-intelligence-and-machine-learning-tools

[47] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, no. 2014, pp. 1–32, 2014.

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden