Sentiment Analysis and Bitcoin: Exploring the Historical Correlation Between Bitcoin Price Movement and Reception on Twitter

Yutong Hu
2550030
yutong.hu@emory.edu
Department of Computer Science

Wenzhuo Ma
2562156
wma44@emory.edu
Department of Computer Science

Sixing Wu
2608986
swu338@emory.edu
Department of Computer Science

***Abstract*—Bitcoin's high volatility and its connection to public sentiment make social media data crucial for analyzing price movements. This project investigates the impact of sentiment analysis and time series data processing on the accuracy of predicting Bitcoin price fluctuations. We employ Natural Language Processing (NLP) tools, such as NLTK and VADER, to process and classify sentiments from Twitter comments. The resulting sentiment polarity scores are combined with historical Bitcoin price data to form the input for predictive models. Both Random Forest (RF) and Long Short-Term Memory (LSTM) networks are utilized, with each model trained on differently processed time series data. This integrated approach leverages sentiment analysis and advanced machine learning techniques to assess whether incorporating sentiment can enhance Bitcoin price prediction.**

1. Introduction

Bitcoin's high volatility and potential for significant financial returns have garnered worldwide attention, creating a strong demand among traders, investors, and market analysts for tools to predict its price movements [1]. In recent years, social media—especially Twitter—has emerged as a critical platform where public sentiment about Bitcoin is frequently expressed. This project hypothesizes that public sentiment on platforms like Twitter may not only serve as an early indicator but also act as a driver of Bitcoin price fluctuations [2]. By exploring the correlation between Twitter sentiment and Bitcoin price action, and building models to predict the Bitcoin price movement, this study aims to provide insights into the performance of incorporating sentiment analysis into prediction.

Despite extensive research on Bitcoin's volatility, the role of social media sentiment, particularly on platforms like Twitter, remains underexplored. Previous studies have predominantly focused on economic news, regulatory changes, and market dynamics, but the direct influence of short-term public sentiment on Bitcoin price movements has received less attention. Furthermore, the impact of user demographics on sentiment trends, and their potential contribution to price fluctuations, is still a significant gap in existing literature.

This project addresses these gaps by analyzing the historical correspondence between Bitcoin price movements and sentiment trends derived from Twitter data. Additionally, it explores whether sentiment varies across demographic groups, providing a more nuanced understanding of public perception and its influence on Bitcoin. Given the unpredictable and emotion-driven nature of cryptocurrency markets, this study seeks to unravel the relationship between social media sentiment and price dynamics.

By leveraging advanced tools such as Natural Language Processing (NLP) for sentiment analysis and machine learning models for predictive analytics, the project aims to uncover actionable patterns and causal links. These insights can help investors and analysts better understand market trends. Furthermore, examining demographic influences on sentiment may offer broader societal insights, ultimately enriching the understanding of Bitcoin's market behavior and future price potential.

2. Related Work

2.1. Bitcoin Price Prediction

For the time-series forecasting on bitcoin prices, traditional methods such as Autoregressive Integrated Moving Average (ARIMA), often combined with machine learning techniques like Artificial Neural Networks (ANN), Random Forests, and Decision Trees, have been gradually supplanted by more robust deep learning approaches. These newer methods typically handle time-series problems more effectively [3]

Common deep learning models include the Long Short-Term Memory (LSTM) Networks, Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and Recurrent Neural Networks (RNN). [4]Researches by Aggarwal et all [5] and Phaladisailoed [6] indicate that LSTM outperforms other deep learning modals in the prediction accuracy. Consequently, LSTM-based models have become the preferred choice among researchers. For instance, Livieris et all [7]used a hybrid CNN-LSTM model and Hamayel et all [8] leveraged the Bidirectional LSTM for predicting bitcoin prices.

Despite the advancements in accuracy achieved by these approaches, reliance solely on historical Bitcoin data can

limit performance and interpretability. This highlights the need for incorporating additional input resources to enhance predictive capabilities.

2.2. Prediction On Bitcoin Prices With Tweet Sentiment Analysis

Recent developments in Natural Language Processing (NLP) have enabled researchers to integrate diverse resources into Bitcoin price prediction, including sentiment analysis of relevant tweets. Arslan et all [9] used two distinctive networks– EMD-LSTM and Stacked Sentiment LSTM– to respectively analyze the historical bitcoin data exclusively and combine the sentiment analysis as well as the output of EMD-LSTM. Critien et all [10] explored the LSTM, BiLSTM, and CNN and compared their effectiveness in this context.

These studies effectively merged several networks to incorporate tweet sentiment analysis into cryptocurrency analytics. However, their methodologies often involve assembling multiple LSTM models, which can be complex and time-consuming. Additionally, these studies have not thoroughly examined deeper insights from tweet information, such as the demographic features of the users posting tweets.

Our research aims to utilize a smaller model size while maintaining a reasonable accuracy rate, and analyze how the demographic features impact tweeters' sentiment about bitcoins.

3. Intended Proposed Approaches

First of all, we will employ a two-step approach to process the Twitter comments data. The first step is to use NLP methods to convert the unstructured text into structured datasets. Then the numerical data can be imputed into the prediction model. We use NLTK (Natural Language Toolkit) to do the tokenization. By removing the noises such as punctuation, symbols, and stop words, we can focus on relevant content.The technologies like bag of words (BoW) and frequency-inverse document frequency (TF-IDF) are used. Bag of words (BoW) is used to make each tweet represented by the frequency of words appearing in the text and frequency-inverse document frequency (TF-IDF) adjusts the word frequency by considering its prevalence across the entire dataset. These transformations provide us with structured numerical features that quantify the textual data and prepare it for sentiment analysis.The second step is to analyze the sentiment expressed in each of the comments qualitatively. We use VADER (Valence Aware Dictionary and sEntiment Reasoner), which is a sentiment analysis tool designed for social media text. This technique works well especially for short, informal texts. By inserting in the tokenized and processed data, VADER will assign four sentiment scores to each tweet: positive, negative, neutral and compound score. The compound score is a normalized measure of overall sentiment, ranging from -1 (most negative) to +1 (most positive). When we successfully label the tweets by sentiment analysis, we are ready to relate it with the price movements of Bitcoin.

Besides processing the Twitter comments data, we apply a set of machine learning techniques into both historical price data and sentiment analysis obtained at the second step to predict the price movement of bitcoin. LSTM (Long Short-Term Memory Networks) will be used in modeling bitcoin price by capturing long-term dependencies. Because LSTM is capable of retaining a long period of data, by incorporating the sentiment scores from tweet analysis and time series data of bitcoin (including the open time, open price, high price, low price, close price), we can capture its influence of public sentiment on the fluctuation of bitcoin price movement. In the next layer, the bidirectional long short-term memory networks might be utilized. The BiLSTM can improve the prediction accuracy due to its bidirectional reading capability, providing more comprehensive patterns. In addition to LSTM and BiLSTM, random forest, which is able to leverage various technical indicators of bitcoin(such as moving averages and trade volume, and the sentiment scores, capturing the nonlinear relations in the model), will also be implemented as a baseline model.

4. Intended System Design

4.1. Architecture

Our architecture consists of five parts: Data Ingestion Layer, Data Preprocessing and Analysis Layer, Modeling and Prediction Layer, Performance Evaluation Layer, and Reporting and Visualization Layer. In the Data Ingestion Layer, Twitter sentiment data and Bitcoin price data will be imported from kaggle dataset and coincodex dataset respectively, ensuring a seamless start for data processing. In the Data Preprocessing and Analysis Layer, we will first apply cleaning techniques to improve the quality of the data, then handling potential missing values and anomalies, and normalizing data if necessary. Some NLP techniques such as tokenization and stop-word removal might also be used in Twitter data form text preprocessing. In the Modeling and Prediction Layer, we will first do sentiment analysis and then employ models, such as Time Series Prediction and Deep Learning Models, for prediction. In this layer, we will also optimize model parameters to improve prediction accuracy. The Performance Evaluation Layer will first divide the data set into three parts: training set, testing set, and validation set, ensuring the model generalization works well on new data. We will also use appropriate metrics like accuracy and F1-score for sentiment classification, and MSE or RMSE for price prediction. Based on evaluation results, we intend to optimize our models by hyperparameter tuning. In the final Reporting and Visualization Layer, Matplotlib and Seaborn will be used to create interactive visualization to depict the relationship between sentiments and Bitcoin price movements.

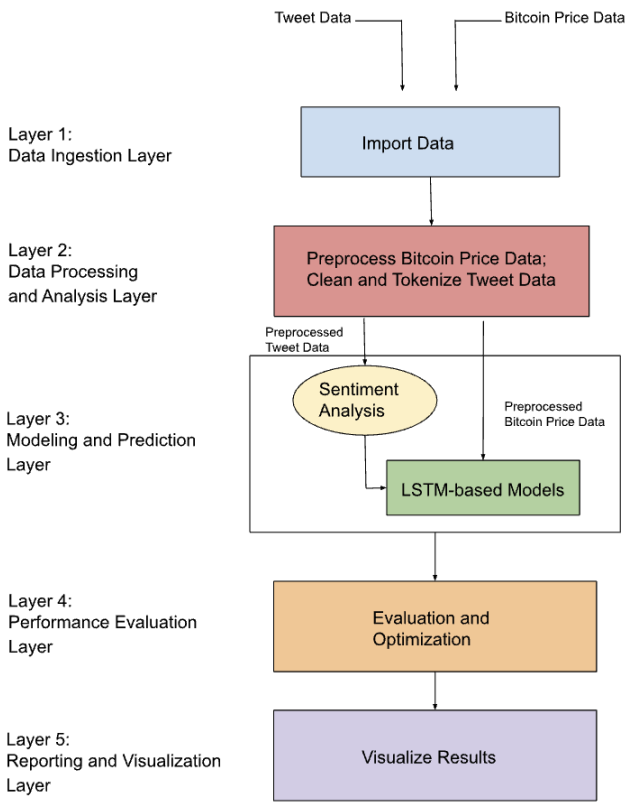

Figure 1. The Architecture for This Project

4.2. Languages, Software Tools, and Data sets

The language we decided to use primarily is Python. It has comprehensive support for data science and machine learning libraries. It is highly suitable to handle natural language processing and predictive modeling tasks. Python has a great ecosystem. It supports libraries such as Scikit-learn for machine learning algorithms, NLTK for processing Twitter data, and tools for deep learning applications. For Twitter data, a lot of python supportive softwares will be used, including NLP(Natural Language Processing) tools such as NLTK (Natural Language Toolkit), bag of words (BoW), frequency-inverse document frequency (TF-IDF), to tokenize data efficiently and accurately. The sentiment analysis tool we used is VADER (Valence Aware Dictionary and sEntiment Reasoner) which can give us the most accurate result in sentiment analysis in tweets.

For predicting bitcoin price, we use LSTM (Long Short-Term Memory Networks), which will be implemented with Keras packages as well as techniques like Dense and Dropout.

For the software tools, the development environment we are going to use is Pycharm. It is an integrated development environment that offers robust features for Python code development, debugging, and system management, facilitating a streamlined workflow. Version control and collaboration are processed by using GitHub. It can make sure the project can be updated, integrated, and maintained efficiently. Additionally, Kaggle is the neighborhood for us to gain pivotal datasets which helps us get the tweet dataset [11] that can be analyzed in our project. The Bitcoin price data is gained from CoinCodex [12]. All these can help us get the trend of the Bitcoin price and correlation with public sentiment which is shown in Twitter data.

5. Experiment

5.1. Dataset Preprocessing

There are 215 unique dates(ranging from 2021-02-05 to 2022-12-27) in our bitcoin tweets dataset. For each date, we randomly selected up to 100 instances. Aftering dropping null value and cleaning text by removing urls, emojis, and some special characters, we finally get a preprocessed bitcoin tweet dataset with the cardinality of 21345.

5.2. Sentiment Scores

The sentiment analysis data set consisted of 13,428 entries. There are 3436 negative scores (25.6% of the total) and 9992 positive scores (74.4% of the total). This part gauges public sentiment surrounding Bitcoin by analyzing sentiment scores derived from textual content. The data has been processed by sklearn and the "SentimentIntensityAnalyzer" package in nltk. The different number of entries from the data after the previous preprocess part is due to the clearness of unmeaningful data, which has a sentiment score of 0. The sentiment scores ranged from -1.0 to 1.0.

Statistical Measure	
Mean	0.281
Median	0.402
Mode	0.440(appearing 554 times)
Standard Deviation(TD)	0.47

TABLE 1. SUMMARY OF STATISTICAL MEASURE OF SENTIMENT SCORES

The statistical analysis to the data are done by using Boxplot and Histogram.

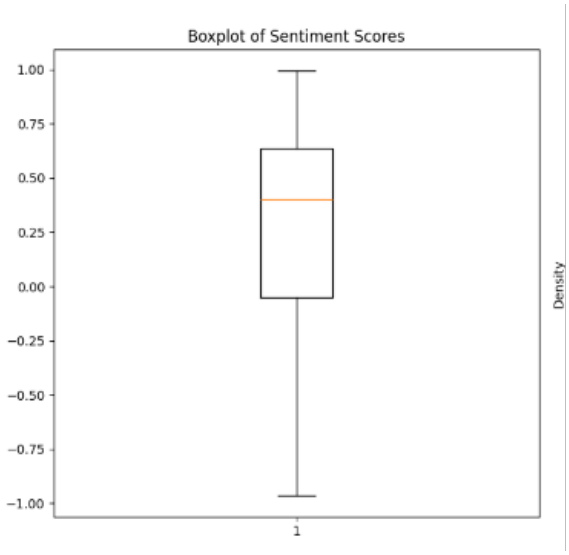

Figure 2. Boxplot of Sentiment Scores

The boxplot shows a median higher than the mean, indicating a negatively skewed distribution, with a longer tail of negative sentiments. This further suggests that while the majority of sentiments are positive, the impact of negative sentiments is significant. The interquartile range of the sentiment data, mostly compressed between 0.0 and 0.75, indicates that most sentiments lean towards the positive, but significant outliers at both ends highlight the presence of extreme sentiments, both positive and negative.

This distribution pattern has direct implications for the Bitcoin market. The accumulation of positive sentiment is more likely to attract new investors to the market, potentially driving up Bitcoin prices. This distribution confirms a predominantly positive sentiment within the data set. Shows people have a general optimism about Bitcoin.

The histogram depicting sentiment scores primarily clustered around 0.5 indicates that the majority of texts display a mildly positive sentiment toward Bitcoin. The range of sentiment scores, spanning from -1.0 to 1.0 with an average of 0.28, and using 0.1 as the interval width, reveals a broad diversity in public sentiment about Bitcoin. There are two peaks in the graph. One is at -0.2 and the other is at 0.5. While the peak at 0.5 is much higher and denser than the one at -0.2 the graph is a single peak histogram and most people are optimistic about Bitcoin. And this histogram is a typical right-leaning graph, which intuitively shows that the general sentiment is positive. The predominantly positive sentiment reflects optimism about Bitcoin's future prospects, possibly linked to its potential for growth as an investment tool or means of payment.

However, the variability also highlights deep divisions and uncertainties concerning Bitcoin. These may stem from its extreme price volatility, potential regulatory risks, concerns over market manipulation, and technical and security issues. Such variability could lead to sharp price fluctuations in Bitcoin's market value, as the market is sensitive to both positive and negative news. This distribution confirms a

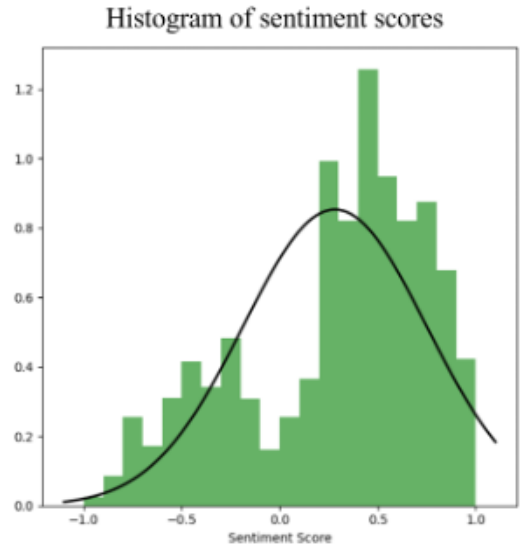

Figure 3. Histogram of Sentiment Scores

predominantly positive sentiment within the data set. Shows people have a general optimism about Bitcoin.

6. Intended System Design

6.1. Architecture

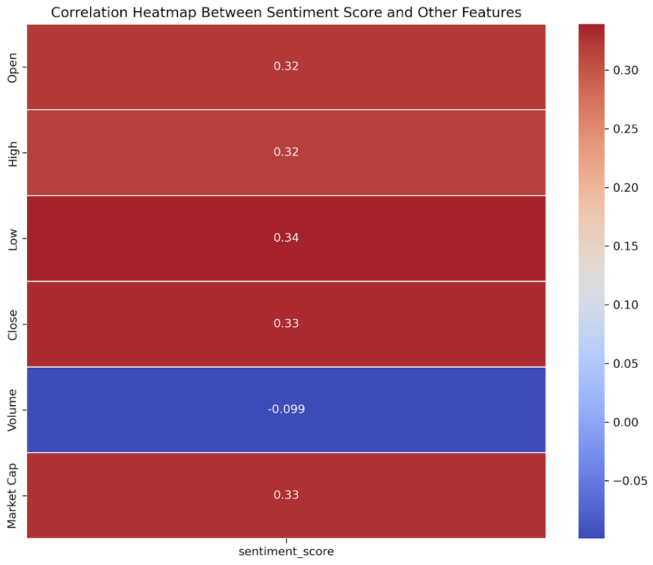

Figure 4. Correlation Heatmap Between Sentiment Score and Other Bitcoin Market Features

After preprocessing all tweet messages and generating corresponding sentiment scores, we analyzed the correlation

between sentiment scores and various price features, including open, close, high, low, volume, and market capitalization. Initially, we excluded data with sentiment scores of 0, as neutral sentiment does not contribute to understanding overall public sentiment. To represent public sentiment on a daily basis, we used the average sentiment score for each day, which falls within the range of -1 to 1.

We then visualized the correlation between sentiment scores and all price attributes (from 2021/02/05 to 2022/12/27) using a heatmap. The results indicated that sentiment scores were moderately positively correlated with open, close, high, low, and market prices. Conversely, sentiment scores were negatively correlated with trading volume on the corresponding day. This finding aligns with expectations, as positive public sentiment is typically associated with better Bitcoin price performance, whereas negative public sentiment tends to coincide with lower trading volumes.

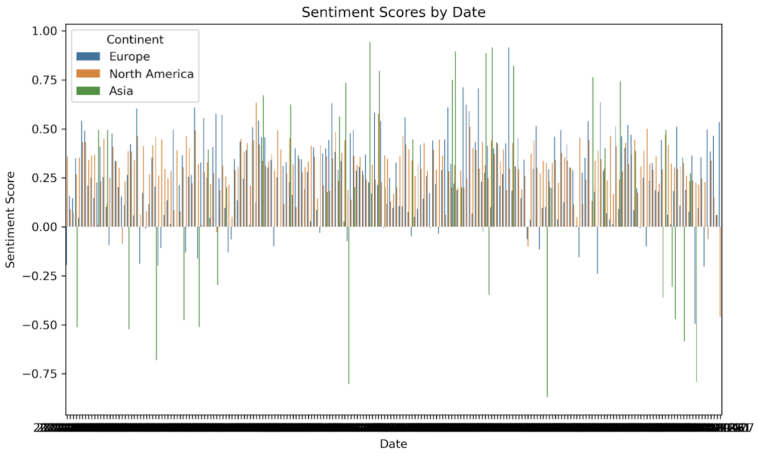

Figure 5. The Correlation Between Sentiment Score and Tweeter's Location

In addition to analyzing the correlation between sentiment scores and all price attributes, we also conducted a correlation analysis between sentiment scores and the geographical location of tweet users. We grouped all tweet messages into three regions: Europe, North America, and Asia—representing the three primary Bitcoin trading continents. We then calculated the average sentiment score for each continent to represent public sentiment on a given date, resulting in sentiment values within the range of -1 to 1.

In our results, blue lines represent Europe, orange lines represent North America, and green lines represent Asia. The box plot indicates that public sentiment in Asia tends to be more volatile. Specifically, sentiment in Asia shows higher positive peaks and lower negative troughs, suggesting greater fluctuations compared to other regions.

6.2. Bitcoin Prediction

6.2.1. Bitcoin Price Presrocess. By analyzing the trend and seasonal data of the Bitcoin Price, we can successfully eliminate the residual in the raw data. Using different periods, we can get data without residual in daily, weekly, monthly, and quarterly. We did this analysis in all data, open price, close price, high price, low price, market cap, and volume. The graph shown is the close price's each period. we can see that the daily data is a little different from the original data while when the period becomes longer, the difference between raw data and data without residual becomes larger and larger. The other graphs can be found in the attachments.

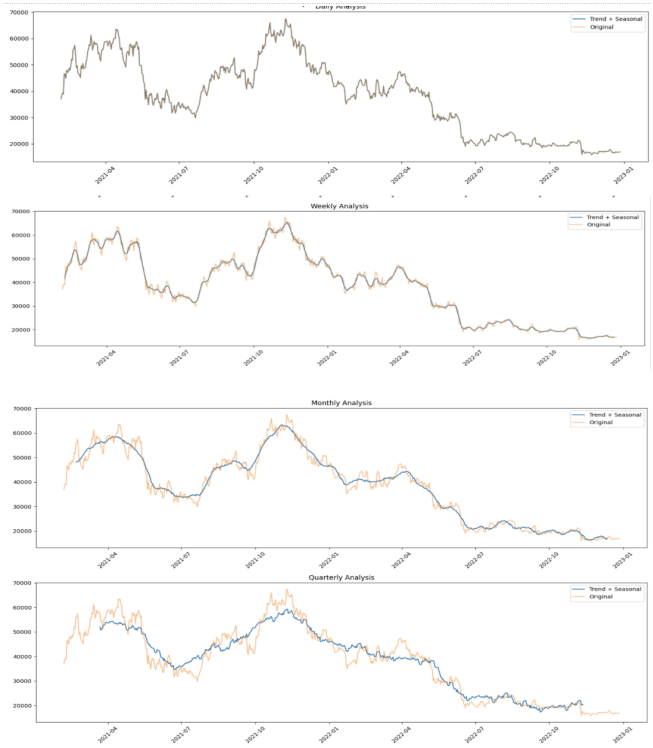

Figure 6. Bitcoin Pirce without residual (open)

6.2.2. Random Forest Model Approach. In our analysis, the Random Forest model is employed to predict Bitcoin's closing price, considering the complex and nonlinear nature of the cryptocurrency market. The decision to use Random Forest was motivated by its ability to handle a mix of feature types, minimize overfitting by averaging results across several trees, and its inherent capability to manage feature interactions, making it well-suited for financial datasets with interdependent variables like price and sentiment. Six distinct scenarios representing different levels of data processing and temporal granularity were performed in this section.

The baseline prediction model utilized unprepossessed Bitcoin price features exacted from the history data. The second model incorporated all the Bitcoin price features (including Open, Close, Volume, High, Low, and Market Cap) and sentiment analysis, capturing public perception towards Bitcoin. Preprocessed Bitcoin price data was used instead for the remaining four prediction models. Weekly, monthly, and quarterly Bitcoin price data without residual is utilized, which indicates the result under different scales.

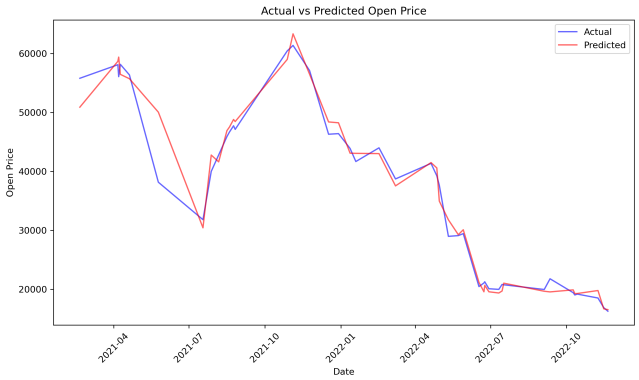

Figure 7. Bitcoin (Close) Price Prediction On Unprocessed Bitcoin Time Series Data Using Random Forest

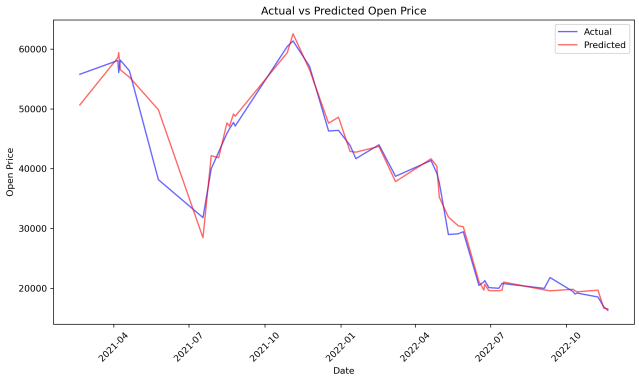

Figure 8. Bitcoin (Close) Price Prediction On Unprocessed Bitcoin Time Series Data With Sentiment Scores Using Random Forest

For each of the prediction scenarios, we grouped the data and calculated an average sentiment score (from 5.2) to the corresponding trading day, capturing public perception towards Bitcoin. To incorporate the effect of prior day events on the current day's price and sentiment, we shifted each feature by one day, thus creating lagged variables to account for the temporal dependence of price movements. We then split the dataset into training and testing subsets, with 80% allocated to training and 20% for testing, ensuring a robust evaluation of model performance. After training the Random Forest Regressor on the training data, we used it to predict the closing price for the test set and compared these predictions against actual prices. We maintain the temporal order and split the dataset into training and testing subsets, with 80% allocated to training and 20% for testing, ensuring a robust evaluation of model performance.

In the model evaluation, we use Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared values, which give us a comprehensive insight into the performance of the prediction, and also indicate in which ways we should use this model. We also visualize the actual values versus predicted values for all six scenarios in a time-series plot, which allows us to assess how well the model captured the price trends and provides an intuitive way to compare predictions against the actual price behavior to better understand the performance.

6.2.3. LSTM Model Approach. Our approach to predicting Bitcoin prices employs Long Short-Term Memory (LSTM) networks across six distinct scenarios, each representing different levels of data processing and temporal granularity. LSTM model is known for its ability of processing and pertaining the information over several steps

The first scenario utilizes unprocessed Bitcoin data incorporating sentiment analysis, providing the highest granularity while capturing market sentiment influence. The second scenario maintains the raw data structure but excludes sentiment features, serving as a baseline for assessing sentiment impact. The remaining four scenarios progressively process the data and aggregate the processed one into daily, weekly, monthly, and quarterly periods, allowing us to examine the trade-off between noise reduction and information preservation at different time scales. For the last four scenarios, no sentiment features are provided.

For each scenario, we implement the MinMaxScaler for both features and target variables to normalize the data within a consistent range. The preparation process includes generating sequence data using a sliding window approach, with various lookback periods tested to determine optimal temporal dependencies. We maintain the temporal order of the data while performing an 80-20 train-test split to ensure realistic evaluation of the model's predictive capabilities.

The LSTM model architecture consists of two main LSTM layers followed by a dense layer for final prediction. The first LSTM layer returns sequences to the second layer, allowing for deep temporal feature extraction. We incorporate dropout layers between the LSTM layers to prevent overfitting, with rates determined through grid search optimization. The model's hyperparameters, including the number of LSTM units (32, 50, or 64) and learning rates (0.001 or 0.01), are systematically optimized for each scenario to ensure optimal performance.

Our training strategy employs the Adam optimizer with mean squared error as the loss function. We implement early stopping with a patience of 20 epochs, monitoring validation loss to prevent overfitting while ensuring adequate model training. The training process utilizes a batch size of 32 and reserves 10% of the training data for validation. This approach allows us to balance computational efficiency with model stability. For model evaluation, we employ a comprehensive set of metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared values. These metrics provide complementary insights into the model's performance across different aspects of prediction accuracy. We supplement these quantitative metrics with visualizations of actual versus predicted values and training history analysis to better understand the model's learning patterns and prediction characteristics. Our implementation provides consistent methodology application across all scenarios while maintaining flexibility for scenario-specific adjustments.

By comparing results across all six scenarios, we aim to understand how different levels of data processing and tem-

poral aggregation affect prediction accuracy. This comparative analysis helps identify the optimal balance between data granularity and noise reduction while assessing the value of sentiment analysis in Bitcoin price prediction. The approach also enables us to determine whether certain temporal scales are more predictable than others, providing insights into the most effective timeframes for Bitcoin price prediction using LSTM networks.

7. Analysis

7.1. Random Forest Model

The Random Forest model effectively captured the general trend of actual close prices, demonstrating its ability to identify key price movements in the volatile cryptocurrency market. This transparency and reliability make Random Forest particularly valuable for evaluating feature importance and understanding market dynamics. However, the RF model incorporating sentiment analysis had little improvement compared to the RF model without sentiment analysis, where applying sentiment analysis outputs an MSE of 5669970.89, MAE of 1466.39, and R^2 of 0.97, and model without sentiment analysis outputs an MSE of 5712939.98, MAE of 1485.02, and R^2 of 0.97. This indicates that sentiment analysis might not improve the prediction significantly.

Since the overall MSE value is large, the model should not be used for predicting the price of an exact day. Instead, it might work as a tool for trend analysis and sentiment prediction, offering insights into broader price movement patterns rather than precise valuations.

7.2. LSTM Model

For LSTM, first of all, based on the prediction results over unprocessed time-series bit coin data, LSTM model involving sentiment analysis had only slightly better but generally really similar performance than that LSTM without sentiment scores does: the former one has MSE of 765682.12, MAE of 639.74, and R^2 of 0.73 while the latter one has MSE of 780126.41, MAE of 629.01, R^2 of 0.73. This implies that, in fact, sentiment analysis does not significantly improve the BTC price prediction.

Furthermore, looking over the results on processed time-series Bitcoin price data, we can find that processed data with daily period significantly outperforms those with other periods: it has the lowest MSE(90933.37), MAE(191.68) and largest R^2(0.98). And we also compare it with the previous two LSTM models based on unprocessed Bitcoin price data: the LSTM model with the processed data (daily period) performs considerably better than them. The former two models have almost nine times more MSE and three times more MAE than it does, while they have much lower R^2 of 0.73. That indicates processing time-series Bitcoin can largely improve the prediction performance.

Figure 9. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Daily Period) Without Sentiment Scores Using RF

Figure 10. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Weekly Period) Without Sentiment Scores Using RF

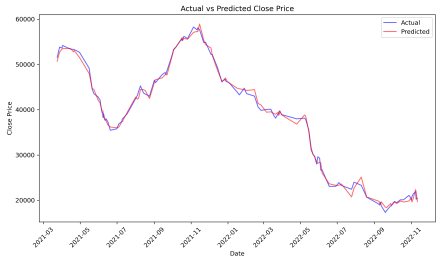

Figure 11. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Monthly Period) Without Sentiment Scores Using RF

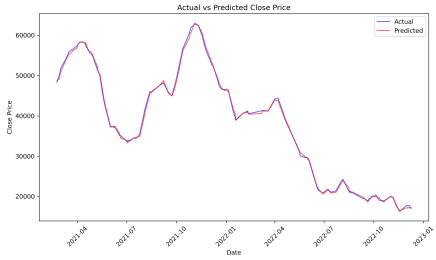

Figure 12. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Quarterly Period) Without Sentiment Scores Using RF

Evaluation Metric	
MSE	5712939.98
MAE Value	1485.02
R^2	0.97

TABLE 2. THE EVALUATION OF BITCOIN (CLOSE) PRICE PREDICTION ON UNPROCESSED BITCOIN TIME SERIES DATA WITHOUT SENTIMENT SCORES USING RF

Evaluation Metric	
MSE	5669970.89
MAE Value	1466.39
R^2	0.97

TABLE 3. THE EVALUATION OF BITCOIN (CLOSE) PRICE PREDICTION ON UNPROCESSED BITCOIN TIME SERIES DATA WITH SENTIMENT SCORES USING RF

Evaluation Metric	
MSE	3124103.73
MAE Value	1259.24
R^2	0.98

TABLE 4. THE EVALUATION MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(DAILY PERIOD) USING RF

Evaluation Metric	
MSE	513272.97
MAE Value	529.63
R^2	0.99

TABLE 5. THE EVALUATION MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(WEEKLY PERIOD) USING RF

Evaluation Metric	
MSE	228318.55
MAE Value	390.95
R^2	0.99

TABLE 6. THE EVALUATION MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(MONTHLY PERIOD) USING RF

Evaluation Metric	
MSE	557631.35
MAE Value	614.91
R^2	0.99

TABLE 7. THE EVALUATION MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(QUARTERLY PERIOD) USING RF

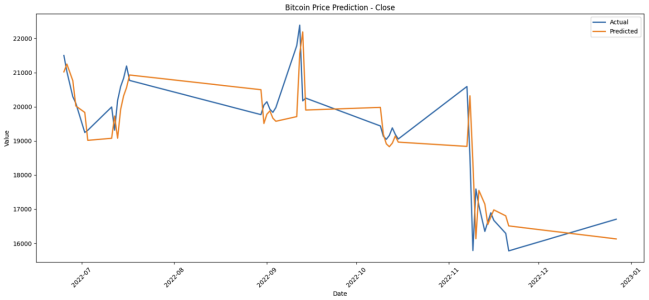

Figure 13. Bitcoin (Close) Price Prediction On Unprocessed Bitcoin Time Series Data With Sentiment Scores Using LSTM

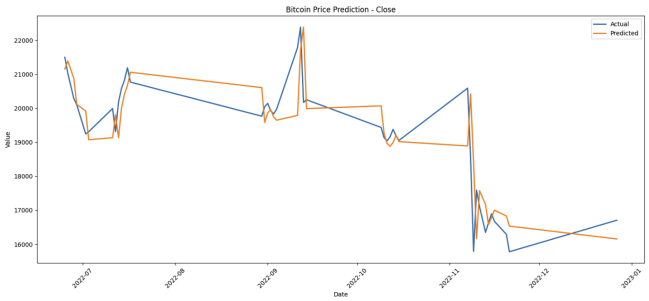

Figure 14. Bitcoin (Close) Price Prediction On Unprocessed Bitcoin Time Series Data Without Sentiment Scores Using LSTM

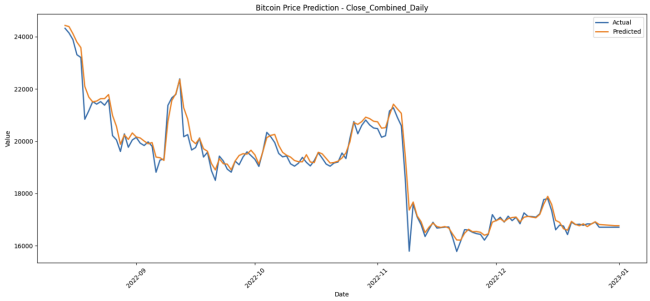

Figure 15. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Daily Period) Without Sentiment Scores Using LSTM

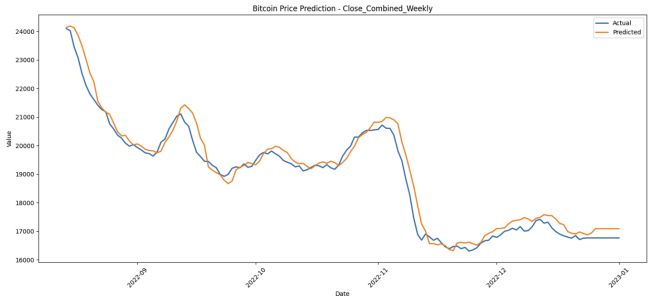

Figure 16. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Weekly Period) Without Sentiment Scores Using LSTM

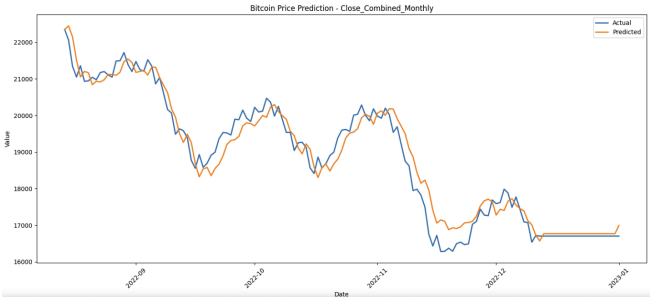

Figure 17. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Monthly Period) Without Sentiment Scores Using LSTM

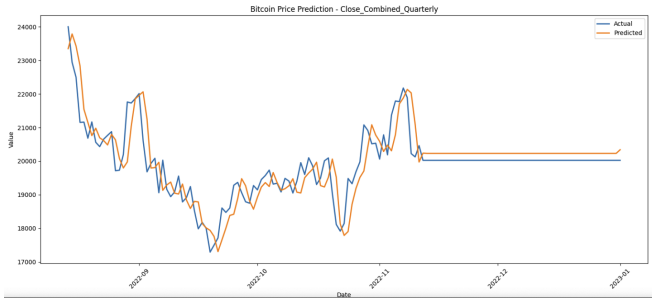

Figure 18. Bitcoin (Close) Price Prediction On Processed Bitcoin Time Series Data(Quarterly Period) Without Sentiment Scores Using LSTM

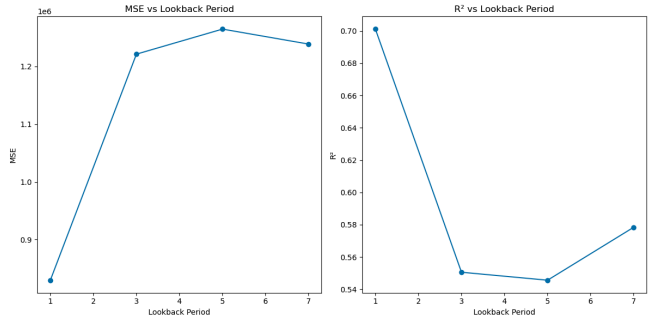

Figure 19. MSE and R^2 Against Lookback Periods

Evaluation Metric	
MSE	765682.12
MAE Value	639.74
R^2	0.73
Statistical Measure	
Mean Actual Value	19237.77
Mean Predicted Value	19186.12
Min Actual Value	15779.97
Max Actual Value	22387.94

TABLE 8. THE EVALUATION AND STATISTICAL MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON UNPROCESSED BITCOIN TIME SERIES DATA WITH SENTIMENT SCORES USING LSTM

Evaluation Metric	
MSE	780126.41
MAE Value	629.01
R^2	0.73
Statistical Measure	
Mean Actual Value	19237.77
Mean Predicted Value	19261.21
Min Actual Value	15779.97
Max Actual Value	22387.94

TABLE 9. THE EVALUATION AND STATISTICAL MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON UNPROCESSED BITCOIN TIME SERIES DATA WITHOUT SENTIMENT SCORES USING LSTM

Evaluation Metric	
MSE	90933.37
MAE Value	191.68
R^2	0.98
Statistical Measure	
Mean Actual Value	18922.71
Mean Predicted Value	19063.22
Min Actual Value	15779.97
Max Actual Value	24323.00

TABLE 10. THE EVALUATION AND STATISTICAL MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(DAILY PERIOD) USING LSTM

Evaluation Metric	
MSE	114201.59
MAE Value	264.70
R^2	0.97
Statistical Measure	
Mean Actual Value	18923.49
Mean Predicted Value	19042.25
Min Actual Value	16304.16
Max Actual Value	24097.07

TABLE 11. THE EVALUATION AND STATISTICAL MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(WEEKLY PERIOD) USING LSTM

Evaluation Metric	
MSE	142768.65
MAE Value	293.38
R^2	0.95
Statistical Measure	
Mean Actual Value	18921.66
Mean Predicted Value	18970.78
Min Actual Value	16279.72
Max Actual Value	22355.12

TABLE 12. THE EVALUATION AND STATISTICAL MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(MONTHLY PERIOD) USING LSTM

Evaluation Metric	
MSE	320851.54
MAE Value	420.95
R^2	0.69
Statistical Measure	
Mean Actual Value	19931.74
Mean Predicted Value	19990.96
Min Actual Value	17291.98
Max Actual Value	24001.53

TABLE 13. THE EVALUATION AND STATISTICAL MEASURE OF BITCOIN (CLOSE) PRICE PREDICTION ON PROCESSED BITCOIN TIME SERIES DATA(QUARTERLY PERIOD) USING LSTM

8. Conclusion

Based on the results from both Random Forest models and LSTM models, involvement of sentiment analysis into Bitcoin price prediction does not show important improvement. However, decomposing the time-series Bitcoin price data based on daily period and dropping the residuals really enhance both RF models' and LSTM models' performance.

8.1. Limitation

Firstly, LSTM models with unprocessed Bitcoin price data perform poorly regardless of the involvement of sentiment analysis, which somehow contradicts our anticipation of LSTM's performance. So we still need to fine-tune our LSTM model to achieve better results. In that scenario, we did not compare the Random Forest models and LSTM models, for we believe that comparison will be somewhat meaningless.

Moreover, because we have limited tweet data, only a subset of Bitcoin price data has corresponding tweet data within our selected time interval. This causes a small data size as well as the discontinuity in our data, which could influence our prediction quality.

8.2. Future Direction

We hope to overcome the alignment issue when we tried to incorporate sentiment scores into Bitcoin price prediction with processed Bitcoin data in the future. This step can help develope a more systematic comparison between models' performance which can provide some insights for the impact of sentiment analysis and processing of data on the performance.

In addition, we also attempted to predict the trend of Bitcoin price, which has its results leveraging LSTM models included in the Appendix section. But the LSTM models performed really poorly in this task. So we want to refine our LSTM models and also add other models for trend prediction.

And for future enhancements of our project on predicting Bitcoin prices using sentiment analysis, we aim to refine our approach by integrating advanced modeling techniques and expanding our feature set. We plan to develop more sophisticated sentiment analysis methods that consider not just sentiment polarity but also its context and intensity. Additionally, we intend to explore multilevel time series modeling to capture intricate temporal patterns and implement machine learning algorithms like Graph Neural Networks and reinforcement learning to compare their effectiveness. By enhancing our feature engineering and incorporating event-driven models that respond to real-time market changes, we expect to significantly improve the accuracy and responsiveness of our predictive models.

References

[1] M. S. Ahmed, A. A. El-Masry, A. I. Al-Maghyereh, and S. Kumar, "Cryptocurrency volatility: A review, synthesis, and research agenda," *Research in International Business and Finance*, vol. 71, p. 102472, Aug. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0275531924002654

[2] T. Yelkenci, B. Dobrucalı Yelkenci, G. Vardar, and B. Aydoğan, "Exploring the relationship between digital trails of social signals and bitcoin returns," *Studies in Economics and Finance*, vol. 41, no. 1, pp. 125–147, Jan. 2024, publisher: Emerald Publishing Limited. [Online]. Available: https://doi.org/10.1108/SEF-12-2022-0572

[3] W. Chen, H. Xu, L. Jia, and Y. Gao, "Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants," *International Journal of Forecasting*, vol. 37, no. 1, pp. 28–43, Jan. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207020300431

[4] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2194, p. 20200209, Feb. 2021, publisher: Royal Society. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0209

[5] A. Aggarwal, I. Gupta, N. Garg, and A. Goel, "Deep Learning Approach to Determine the Impact of Socio Economic Factors on Bitcoin Price Prediction," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Aug. 2019, pp. 1–5, iSSN: 2572-6129. [Online]. Available: https://ieeexplore.ieee.org/document/8844928

[6] T. Phaladisailoed and T. Numnonda, "Machine Learning Models Comparison for Bitcoin Price Prediction," Jul. 2018, pp. 506–511.

[7] I. E. Livieris, N. Kiriakidou, S. Stavroyiannis, and P. Pintelas, "An Advanced CNN-LSTM Model for Cryptocurrency Forecasting," *Electronics*, vol. 10, no. 3, p. 287, Jan. 2021, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2079-9292/10/3/287

[8] 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, Dec. 2021, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2673-2688/2/4/30

[9] S. Arslan, "Bitcoin Price Prediction Using Sentiment Analysis and Empirical Mode Decomposition," *Computational Economics*, May 2024. [Online]. Available: https://doi.org/10.1007/s10614-024-10588-3

[10] J. V. Critien, A. Gatt, and J. Ellul, "Bitcoin price change and trend prediction through twitter sentiment and data volume," *Financial Innovation*, vol. 8, no. 1, p. 45, May 2022. [Online]. Available: https://doi.org/10.1186/s40854-022-00352-7

[11] "Bitcoin Tweets." [Online]. Available: https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets

[12] "Bitcoin (BTC) Historical Data | CoinCodex." [Online]. Available: https://coincodex.com/crypto/bitcoin/historical-data/

Appendix

Bitcoin Price without Residual

We have done the residual elimination, Bitcoin preprocess for all data of Bitcoin. Including close price, high price, low price, market cap, and volume.

Bitcoin Price Trend Prediction with LSTM

We also predict the trend of the BTC (close) price with the LSTM model. The trend labels are up (BTC price grows), sideways(BTC price does not change), and down(BTC price flops). And the results of prediction in different scenarios are shown in Fig. 14-19. Based on those tables, our LSTM model performs somewhat poorly in predicting trends. Among those scenarios, LSTM model always tend to only reflect two labels of three. Therefore, we might need to better scale and adjust our trend threshold to get a better performance.

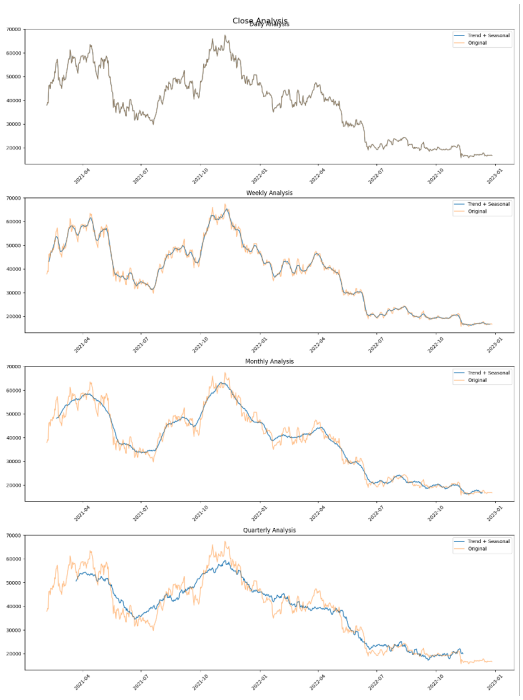

Figure 20. Bitcoin Preprocess (Close Price)

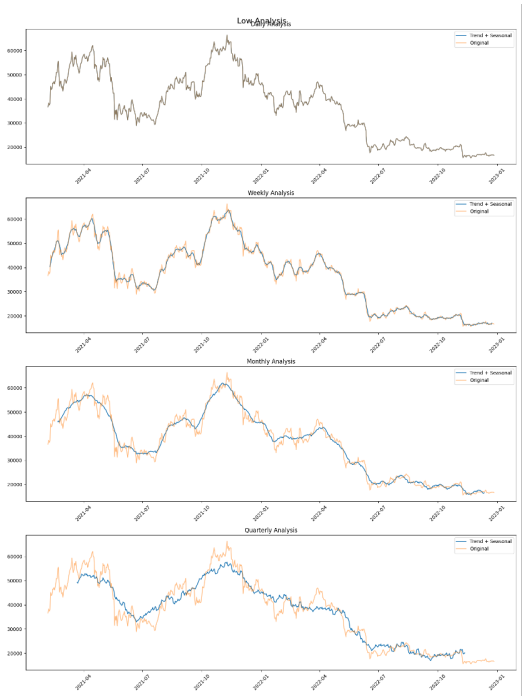

Figure 21. Bitcoin Preprocess (Low Price)

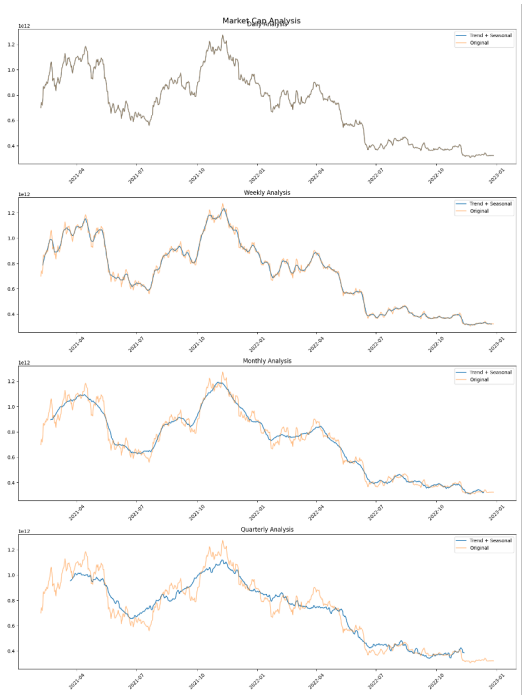

Figure 22. Bitcoin Preprocess (Market Cap)

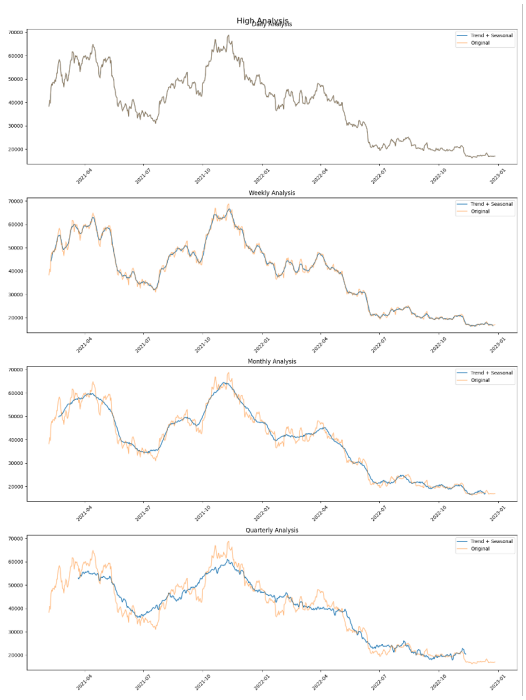

Figure 23. Bitcoin Preprocess (High Price)

Figure 24. Bitcoin Preprocess (Volume)

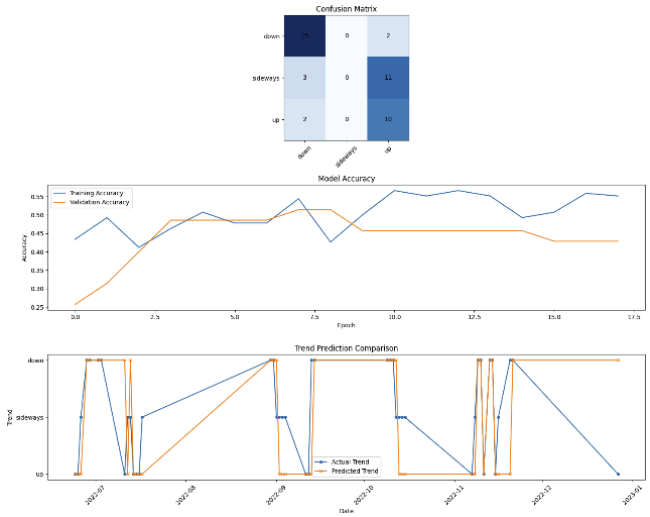

Figure 25. Bitcoin (Close) Price Trend Prediction On Unprocessed Bitcoin Time Series Data With Sentiment Scores Using LSTM

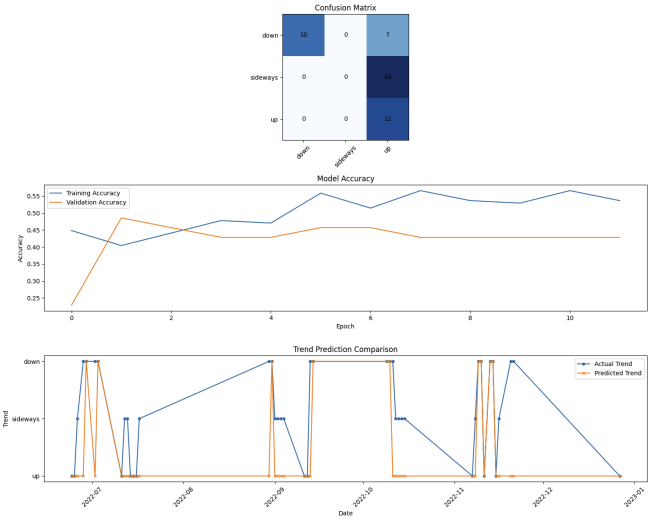

Figure 26. Bitcoin (Close) Price Trend Prediction On Unprocessed Bitcoin Time Series Data Without Sentiment Scores Using LSTM

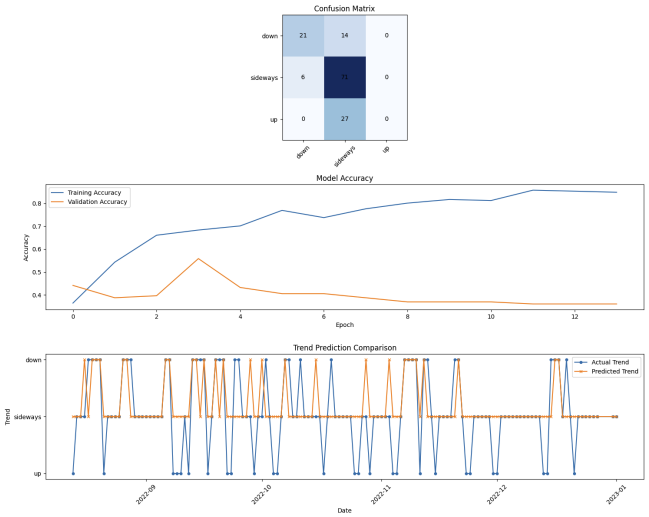

Figure 27. Bitcoin (Close) Price Trend Prediction On Processed Bitcoin Time Series Data(Daily Period) Using LSTM

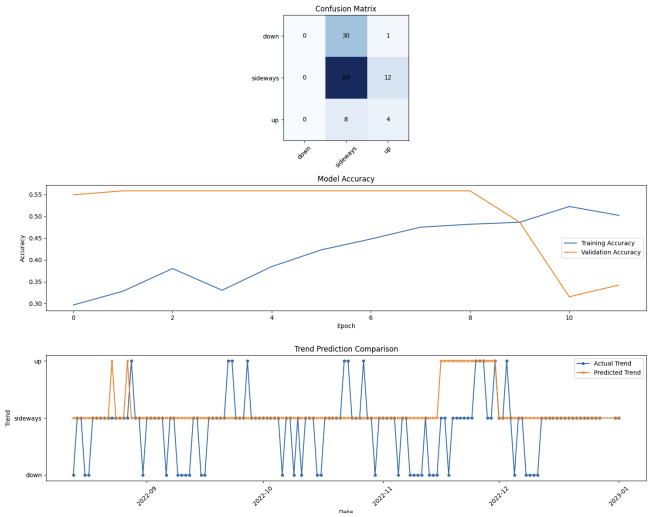

Figure 29. Bitcoin (Close) Price Trend Prediction On Processed Bitcoin Time Series Data(Monthly Period) Using LSTM

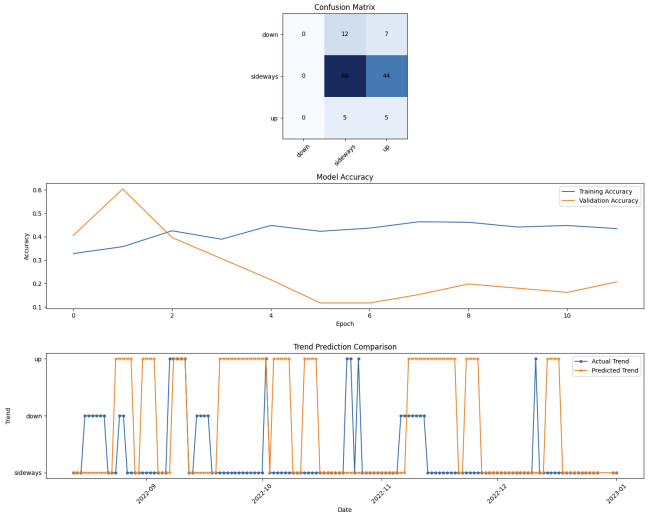

Figure 28. Bitcoin (Close) Price Trend Prediction On Processed Bitcoin Time Series Data(Weekly Period) Using LSTM

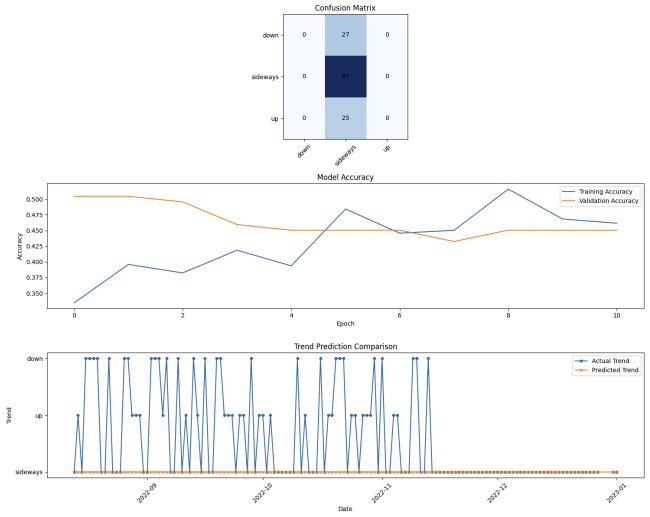

Figure 30. Bitcoin (Close) Price Trend Prediction On Processed Bitcoin Time Series Data(Quarterly Period) Using LSTM