Bitcoin in North America: Sentiment Analysis of Regulations and Price Prediction

Final Report

Juan Carlos Ramos

Student ID: 500393737

Ryerson University

Table of Contents

Abstract

The dynamic and relatively nascent field of cryptocurrency is increasingly impacted by regulatory measures in North America, particularly by the Commodity Futures Trading Commission (CFTC), the Securities and Exchange Commission (SEC), the Financial Industry Regulatory Authority (FINRA) and the Canadian Securities Administrators (CSA). This project will attempt to address the need to understand the sentiment and widespread adoption of cryptocurrency in light of these regulatory frameworks. The primary research question focuses on how current and proposed regulations by these bodies influence public sentiment towards cryptocurrencies. The study will also seek to investigate whether these regulatory announcements of proposed rules and implementation dates could predict cryptocurrency price fluctuations, especially in the context of "sell the news" events, where prices might drop following major announcements or regulatory updates. One such recent example being the SEC's approval of listing and trading of the spot bitcoin ETF.

The data for this study will be sourced from publicly available documents, including current and proposed rules, regulations, and official statements from the websites of the CFTC, SEC, FINRA and CSA. These will provide a comprehensive view of the regulatory landscape in North America. Additionally, cryptocurrency price data corresponding to the timeline of these regulatory announcements will be incorporated to examine potential correlations between regulatory actions and market reactions (bullish/bearish). The price data will be sourced from one of the many crypto exchanges that track historical crypto prices.

To analyze this broad and unstructured dataset, I've chosen to employ the theme of text mining and sentiment analysis techniques. Text mining will be used to extract relevant information and

key themes from the regulatory documents. Sentiment analysis will then be applied to gauge the tone and sentiment of these documents, categorizing them as positive, negative, or neutral in the context of cryptocurrency. This sentiment analysis will be correlated with cryptocurrency adoption trends and market data to identify any significant relationships.

For the prediction aspect, the study will use time-series analysis to explore the impact of regulatory announcements on cryptocurrency prices. The project will use machine learning algorithms to analyze the relationship between the timing of these announcements, the language used in the regulation and subsequent price movements, thereby assessing the feasibility of predicting price changes based on regulatory news. A revision to this abstract is the focus on Bitcoin prices instead of multiple crypto currencies to keep the research relatively cleaner. Regulations mostly focus on Bitcoin and Ether and it has been shown that most other crypto currencies follow the same price movement.

The primary tools for these analyses will include Python programming language, utilizing libraries such as NLTK for natural language processing, Pandas for data manipulation, and scikit-learn for machine learning models. Additionally, visualization tools like Matplotlib and Seaborn can be used to present the findings in an easy to understand manner.

In summary, this project aims to provide a nuanced understanding of the impact of North American regulatory measures on cryptocurrency sentiment and market behavior. The findings are expected to offer valuable insights for investors, regulatory bodies, and academics and person's like myself interested in the intersection of finance, technology, and regulation.

Literature Review

The intersection of sentiment analysis and financial markets has become an increasingly important area of study, especially within the volatile realm of cryptocurrencies. This literature review aims to synthesize the current body of research on sentiment analysis of social media posts as it relates to the prediction of cryptocurrency prices, as well as the sentiment analysis of regulatory filings and their subsequent impact on market movements. Exploring these two areas should help in answering the primary research question guiding this review which is centered on understanding how current and proposed regulations (and the language used wherein) by financial and governmental bodies influence public sentiment towards cryptocurrencies and, by extension, their market prices.

Cryptocurrencies, by their very nature, are highly susceptible to public sentiment, given their speculative nature and the relative infancy of the market. Unlike traditional financial assets, cryptocurrencies do not have a long history of market behavior or a wealth of financial data to draw upon, making sentiment analysis a particularly valuable tool for predicting market trends. Many investors have begun using sentiment analysis as a viable trading strategy in not only traditional financial (Trad-Fi) markets but in crypto as well. Companies sell dashboards and APIs that connect to social media sites such as X or Reddit to gauge public sentiment to determine any price fluctuations. Because of this opportunity, there is a wealth of research in the use of social media sentiment analysis and market price predictions, as will be seen in (6). There also exists research that conducts sentiment analysis on regulatory filings to gauge the impact it poses on any relevant stakeholders such as broker dealers (4). It is generally known of course that regulatory announcements cause bearish reactions in the targeted field, but it is not immediately quantifiable as to the effect that the price would go down. Recently however, the approval of the spot Bitcoin ETF caused the price of bitcoin to spike almost 100%. Similarly, economists track which words are used during Federal Reserve/Bank of Canada interest rate announcements to see whether the next announcement will trend “hawkish” or “dovish”. This research project sits in between both areas, and can utilize the methodologies used previously in an area that hasn’t fully been explored, which is analysis and how the use of language within regulatory filings pertaining to crypto can be used to predict price movements and its magnitude.

Research Papers

Sentiment and Uncertainty About Regulation

The comprehensive study conducted by Tara M. Sinclair and Zhoudan Xie delves into the nuanced impacts of regulatory sentiment and uncertainty on the U.S. economy, leveraging advanced natural language processing techniques to analyze a vast corpus of news articles spanning from 1985 to 2021. Their work culminates in the creation of monthly indexes that quantify regulatory sentiment and uncertainty, alongside categorical indexes tailored to 14 distinct regulatory policy areas. Their analysis sheds light on the intricate dynamics between regulatory perceptions and economic performance, offering many insights specific to the overall project.

Methodological Approach:
The methodological rigor and innovative use of natural language processing set this study apart and is the reason why this review is beginning with it. By constructing indexes from a broad array of news sources over an extensive period, the research offers a granular view of the regulatory landscape's impact on the economy, surpassing previous methods that relied on more limited or qualitative assessments.

A major takeaway is the use of the “Loughran and McDonald (LM) dictionary (originally developed in Loughran and McDonald (2011)) to assess the sentiment and uncertainty in the regulatory sections of the relevant news articles in the baseline analysis. It was constructed specifically for the domain of finance, using a corpus of corporate 10-K reports (Loughran and McDonald, 2011). Because of its domain relevance, the LM dictionary has been frequently used in economic research (for example, Fraiberger (2016); Calomiris et al. (2020); Ostapenko (2020)). The 2018 version of the dictionary comprises sentiment word lists in several categories, including 2,355 words in the negative category, 354 words in the positive category, and 297 words in the uncertainty category.” (1)

The existence of this sentiment dictionary and its usage will be pivotal to determine the sentiment of crypto specific regulatory text. In the paper the authors used a standard formula to calculate sentiment scores. The “regulatory sentiment score” of an article is the difference between the proportion of positive words and the proportion of negative words in the regulatory section of the article. So a positive sentiment score indicates an overall positive tone in the news about regulation, and a negative score means an overall negative tone. One thing to consider, just as the authors do, is that the dictionary skews far more negative than positive as seen in the proportion of words (2,355 words in the negative category, 354 words in the positive category), and could potentially bias the results of this projects sentiment analysis disproportionately negative. To counteract the negativity bias, the authors also used the Harvard General Inquirer (GI) dictionary and the Lexicoder Sentiment Dictionary (LSD). The GI dictionary being a general-purpose lexicon originally developed in the 1960s and has been widely used in various disciplines. It covers several broad valence categories, including lists of 2,005 negative words and 1,637 positive words. The LSD is a comprehensive sentiment lexicon combining three pre-existing dictionaries and tailored primarily to political news (Young and Soroka, 2012).5 The LSD comprises 2,857 negative words and 1,709 positive words. (1)

Although specific to news articles pertaining to regulations, the same formula can be applied to the actual regulations in theory, and any combination of these dictionaries can be used as part of the project's end goal.

Negative Regulatory Sentiment and Economic Performance:
As to the paper's results, a pivotal discovery of Sinclair and Xie's research is the significant, lasting downturn in economic output and employment triggered by adverse shocks to regulatory sentiment. Specifically, their data illustrate that a negative shift in regulatory sentiment can lead to a reduction in industrial production by up to 0.61% and employment by up to 0.28% within a year of the shock. This relationship underscores the profound influence that negative perceptions of the regulatory environment can have on the broader economy and is something to keep in mind when assessing the regulatory impact on crypto.

Sector-Specific Sensitivities:
The research goes further to dissect the impact of regulatory sentiment and uncertainty within specific policy domains. It identifies transportation, consumer safety and health, general business and trade, and energy regulations as areas where sentiment and uncertainty exhibit a more pronounced effect on economic performance. For instance, negative sentiment shocks related to transportation and consumer safety and health regulations are linked to substantial, persistent declines in future output. The authors achieved this through the use of “noun chunks” to identify regulation related news articles by industry. Although for the purposes of this project the industry is specified, (cryptocurrency) the idea of the noun chunks can be used to potentially distinguish the impact of a regulation on a more granular level such as differentiating between stablecoins, altcoins, CBDC etc.

Regulatory Uncertainty's Limited Impact:
Contrary to the pronounced effects of regulatory sentiment, the study finds that increases in regulatory uncertainty generally have a negligible or only transient impact on economic outcomes. This distinction between the effects of sentiment and uncertainty is a crucial aspect of their findings, highlighting the more significant role of sentiment in economic fluctuations. It will be interesting to see if the words in the “uncertain" category have the same effect on crypto and thus have a smaller impact.

An Analysis of Speculative Language in SEC 10-K Filings

Jonathan L. Pulliza's research offers a nuanced exploration into speculative language within SEC 10-K filings, employing sentiment analysis techniques to discern patterns of speculation in these critical financial documents. By training a model on the MPQA corpus, Pulliza explores the complex terrain of speculative sentences, applying this model to a corpus of SEC 10-K documents over a five-year span. This approach illuminates the varying concentrations of speculative terms within documents, particularly those related to risk factors such as projects, taxes, and pensions. Due to the nature of regulatory documents and its speculative and regulatory language this paper was chosen for review and the same models can be used in framing future outlooks of crypto regulation.

High Concentration of Speculative Language:
The study reveals that documents laden with speculative language exhibit a distinct lexical pattern, differing significantly from the broader corpus. This suggests that heightened speculation is often linked to discussions around potential risks and uncertainties facing firms, providing a deeper understanding of the strategic language used in financial disclosures. This differs from most regulations, as the creators seek to define roles and responsibilities of firms by

minimizing speculation and not leaving much room for interpretation of the rules so that roles and responsibilities are clearly defined. However regulators generally state which risks the regulations seek to fix and the language used can potentially measure the impact it will have on crypto.

Methodological Approach:
The study stands out for its methodological rigor, leveraging sentiment analysis and machine learning techniques to systematically extract and analyze speculative language. This approach not only enhances the reliability of the findings but also contributes to the methodological advancements in the field of financial text analysis. The models utilized for this study were Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Decision Trees.
For the purposes of this project, inclusion of speculative language as a variable may take away from words that convey “uncertain” sentiment as part of the lexicon. SEC 10-K filings in general suggest to investors a company's forecast along with relevant financial information. Regulatory documents generally use “plain english” in what will be enforced and the technical requirements needed for stakeholders to abide by them. It appears to be a matter of semantics, but the models can still be applied with some adjustments.

Bitcoin price change and trend prediction through twitter sentiment and data volume

Jacques Vella Critien and colleagues' research investigates the potential of Twitter sentiment and tweet volume in predicting Bitcoin price trends. Their study builds on previous work by not only aiming to predict the direction of price changes but also the magnitude of these changes. Utilizing sentiment extracted from Twitter data, alongside the volume of tweets, the researchers present findings from various experiments exploring the relationship between sentiment and future Bitcoin prices at different time intervals. They aim to identify the optimal time frame in which expressed sentiment becomes a reliable indicator of price change. Two neural network models, one based on recurrent nets and another on convolutional networks, are explored and evaluated. Additionally, a model to predict the magnitude of change is introduced, framed as a multi-class classification problem, which when used in conjunction with a price trend prediction model, yields more reliable predictions.

Magnitude Prediction:
The novel contribution of this research is in predicting the magnitude of price changes, which goes beyond the current state-of-the-art that primarily focuses on the direction of price movements. Seeing how this project seeks to predict magnitude, the models used within this research paper can be used with modifications. The paper used a Bitcoin pricing dataset and

cleaned the data so that the high and low prices were removed from the feature list so as to only keep the average price per minute. This is more conducive when working with social media sites like twitter where the volume of tweets is high. But it may just be beneficial and simpler to follow the daily closing price set out by the data because the rate of which regulations are put out is nowhere near the volume of tweets.

Temporal Granularity: The research delves into the temporal aspect of sentiment analysis, exploring different time intervals to determine when sentiment becomes a reliable predictor of price changes. This could provide insights into how quickly crypto regulation announcements and proposals are reflected in its prices, and the models for time lags will be investigated as an option, but it makes sense that a lag exists, as more analysts get more time to review crypto regulation to assess their broader impact.

Model Comparison: By comparing recurrent and convolutional neural network models, the study provides insights into the effectiveness of different neural network architectures in analyzing sentiment data and predicting asset prices. Three different models, (i) using an LSTM, (ii) CNN and (iii) Bidirectional Long Short Term Memory Cells (BiLSTM), were implemented for predicting whether the following day's closing price will increase or decrease. For the three price direction prediction models (Direction-LSTM, Direction-CNN and Direction-BiLSTM) at different time-lags of 1, 3 and 7 days. Another prediction model tries to predict the magnitude of the change of closing day prices as a multi-class classification problem. This is done by predicting which interval the closing day price changes would fall into. (3) All of these price prediction models could be used in conjunction with sentiment analysis and regulatory filings to predict bitcoin price.

Can SEC Comment Letters of Regulation A Offerings Serve as Quality Signals for Investors and Firms? Preliminary Results are Encouraging

David S. Krause's study delves into the role of Securities and Exchange Commission (SEC) comment letters as potential quality signals in the context of Regulation A (Reg A) offerings. This exploration, set against the backdrop of the JOBS Act's implementation, leverages sentiment analysis to scrutinize the nuances within these letters and their correlation with the SEC's qualification decisions for proposed offerings. The findings suggest a positive link between the sentiment conveyed in comment letters and the likelihood of an offering's qualification, thus enriching the signaling theory within the realm of Reg A offerings.

This research holds significant implications for the project focused on sentiment analysis of proposed cryptocurrency regulations. Krause's methodology and insights can be adapted to analyze public comments and regulatory communications within the cryptocurrency sector. By applying sentiment analysis, we can gauge the regulatory climate and stakeholders' sentiments towards proposed regulations, offering a nuanced understanding of the potential impacts and reception of these regulations.

Sentiment as a Predictor:
The positive relationship between the sentiment in SEC comment letters and offering qualification highlights the predictive power of sentiment analysis. This approach can be applied to assess sentiments in public comments on proposed cryptocurrency regulations, providing indicators of the regulations' acceptance and potential enactment.

Signaling Theory Application:
Krause's application of signaling theory to Reg A offerings underscores the importance of signals in regulatory and investment environments. For cryptocurrency regulations, analyzing public comments through the lens of signaling theory could reveal how different stakeholders signal their approval, concerns, or opposition to regulatory proposals.

Methodological Approach:
The study's methodological approach, combining sentiment analysis with statistical validation, offers a robust framework for evaluating sentiments in regulatory communications. This framework can be adapted to analyze sentiments in the context of cryptocurrency regulations, facilitating a data-driven understanding of stakeholder perspectives.

To conduct sentiment analysis, the SentimentIntensityAnalyzer from the Natural Language Toolkit (NLTK) was utilized. This rule-based sentiment analysis tool calculates the sentiment of a piece of text by assigning a positive (+1), negative (- 1), or neutral (0) score to each word in the text, and then combines those metrics to calculate an overall sentiment score for the text. This approach required the extraction of SEC comment letters from EDGAR for each firm in the sample and the conversion of the pdf files into text files. By using the NLTK library and a pre-trained sentiment analysis model, it is possible to create a unique sentiment score for each text (4).

Krause however was aware of the limitation of the sentiment analyzer and suggested that it would be better to use a domain-specific sentiment analysis lexicon that is trained on a corpus of financial and legal texts so that the results accurately reflect the unique language and terminology used in this domain.

Accessing, Extracting, and Analyzing the Textual Content within Reg A Form 1-A Part II Filings Using Python

Another paper by David S. Krause, titled "Accessing, Extracting, and Analyzing the Textual Content within Reg A Form 1-A Part II Filings Using Python," focuses on the methodology for analyzing textual information in Regulation A (Reg A) filings, particularly Part II, which contains detailed business descriptions, risk factors, and financial information. Krause outlines the process of using Python, along with libraries such as BeautifulSoup and TextBlob, to efficiently preprocess, extract, and analyze textual data from these filings.

Methodological Approach:
Krause emphasizes the importance of analyzing textual content in Reg A filings to complement financial statement analysis, aiding in the evaluation of investment risks and rewards.
The paper provides a step-by-step guide on using Python to access, extract, and perform sentiment analysis on the textual content of Reg A filings, highlighting the role of libraries like BeautifulSoup and TextBlob. Challenges such as the unstructured nature of the text, language complexity, and data accessibility are discussed, along with the potential for advanced AI and machine learning techniques to address these challenges.

The methodologies outlined in Krause's paper can be adapted for analyzing proposed cryptocurrency regulations. By leveraging Python to extract and parse textual content from regulatory proposals, stakeholders can gauge the sentiment, understand the regulatory focus areas, and identify potential impacts on the cryptocurrency market. A challenge that was highlighted is that while many regulatory filings are publicly available, accessing and extracting the data in a structured and standardized format can be challenging. So a standardized format may be required when normalizing crypto regulations from different regimes as part of this project.

Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting

The research paper "Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting" by Markus Frohmann and colleagues explores an innovative approach to predict Bitcoin's future price by combining time series forecasting with sentiment analysis derived from blogs.

Hybrid Approach:

The study introduces a hybrid model that integrates time series forecasting of Bitcoin prices with sentiment analysis obtained from blogs. This approach aims to enhance prediction accuracy by considering not only historical price data but also public sentiment towards Bitcoin. They investigated four types of ML-based algorithms: LSTM networks, TCNs, the D-Linear method, and linear regression, along with some simple baselines which include Exponential Smoothing, Fast Fourier Transform, naive mean and naive drift.(6) This hybrid approach could serve as a potential model of forecasting future bitcoin prices against regulatory sentiment analysis.

Sentiment Analysis with BERT:
A significant contribution of this paper is the use of a fine-tuned BERT model for sentiment analysis, a different application in the context of Bitcoin price prediction. BERT is a pre-trained transformer model that has been used previously to analyze tweets. The authors employ a BERT-based sentiment analysis to evaluate the sentiments expressed in microblogs, providing a deeper understanding of public sentiment towards Bitcoin. This model however won't be of much use for the project as it can't be used on regulations due to a specific lexicon being required.

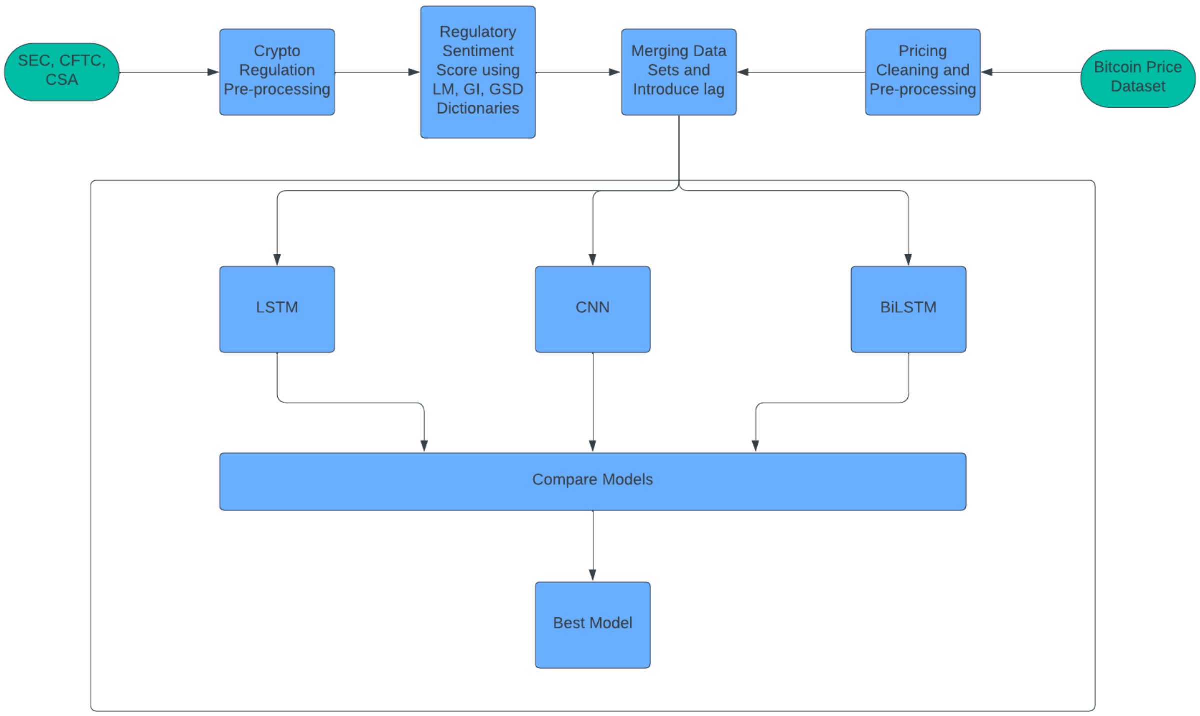
Performance Metrics:
The models are evaluated based on Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), with the hybrid models that utilize linear regression showing the best performance in terms of these metrics.

Methodological Approach:
The methodologies outlined in this paper can be applied to the broader project of using sentiment analysis in crypto regulation across North America by:

- Incorporating Regulatory Sentiments: Analyzing sentiments from regulatory announcements, discussions, and news articles using a modified BERT model to understand the impact of regulatory changes on cryptocurrency markets.
- Predictive Analysis: Integrating sentiment analysis with time series forecasting of cryptocurrency prices to assess how regulatory sentiments influence market dynamics and price movements.
- Weighted Sentiment Scores: Applying the weighting scheme based on the influence or credibility of the source (e.g., regulatory bodies, influential market analysts) to gauge the impact of regulatory sentiments on the market.

Methodological Framework

Adopting methodologies from prior works, this study proposes a hybrid approach integrating sentiment analysis of regulatory filings with time-series market data analysis. It outlines the use of advanced NLP techniques for sentiment extraction and machine learning algorithms for price prediction, facilitated by Python and its libraries.



1. Using Krause's analysis on extracting and analyzing textual content within Reg A filings using Python. This methodology can be leveraged to process and analyze textual data from cryptocurrency regulatory documents normalizing them across all jurisdictions. A standard format would make the text cleaner and in better shape for sentiment analysis.
2. Sinclair and Xie's Work provided insights into the impact of regulatory sentiment on the economy, employing NLP to create sentiment indexes. This methodology along with the Loughran and McDonald (LM) dictionary, Harvard General Inquirer (GI) dictionary and the Lexicoder Sentiment Dictionary (LSD) can be adapted to analyze sentiments in cryptocurrency regulatory filings, draft proposals and final rules, offering a quantitative measure of regulatory sentiment. The “regulatory sentiment score” standard formula is also something that will be utilized to measure how positive or negative or uncertain the crypto regulations are.
3. Pulliza's research highlights the use of sentiment analysis to identify speculative language in SEC filings. This approach could be applied to scrutinize speculative

sentiments in cryptocurrency regulatory texts, aiding in the prediction of market reactions to regulatory news.

4. To forecast the price of bitcoin three different models, (i) using an LSTM, (ii) CNN and (iii) Bidirectional Long Short Term Memory Cells (BiLSTM), can be implemented for predicting whether the following day's closing price will increase or decrease. Different time lags can also be introduced for the three price direction prediction models (Direction-LSTM, Direction-CNN and Direction-BiLSTM) at different time-lags of 1, 3 and 7 days. The lags should be beneficial to track how sentiment of regulations change over time as further subject matter experts analyze and assess the impact of the regulation. LSTM was used in both Critien et al.'s and Frohmann et al.'s research.
5. A weighting scheme used by Frohmann et al.'s can be adapted for analyzing sentiments in regulatory filings and to differentiate which federal body has more or less impact on cryptocurrency prices.

Research Questions (revised)

The primary research question focuses on how current SEC enforcement actions by these affect public sentiment towards cryptocurrencies. The study will also seek to investigate whether these enforcement actions and their sentiment could be used to predict cryptocurrency price fluctuations, especially in the context of "sell the news" events, where prices might drop following major announcements.

Applied Methodology

Scope of the Research

The project aims to use machine learning algorithms to predict short-term closing prices of Bitcoin. The dataset for this project is obtained from Yahoo which contains historical data for the chosen cryptocurrency. The objective is to compare the predicted prices with the actual prices and identify which model is the most accurate in utilizing sentiment analysis to predict price. To answer the research question, we will explore different machine learning algorithms and time series analysis techniques.. We will compare the efficiency and stability of these techniques to identify the most effective ones for our purpose.

The following steps outline a broad and revised structured methodology for analyzing cryptocurrency market data compared:

Data Collection: The first step relevant historical price data for bitcoin was exported as a csv from yahoo finance. The enforcement actions from the SEC were then scraped from the SEC website using EDGAR.

Data Preparation: Any missing or erroneous data, transforming the data into a suitable format, and normalizing it to account for differences in scale are adjusted before obtaining the sentiment analysis utilizing the Loughran and McDonald (LM) dictionary.

Model Selection: LSTM, BiLSTM and CNN models were selected as the models due to historical research showing these being the most promising.

Model Implementation: The models are implemented and applied to the prepared merged dataframe.

Data Visualization: Exploratory data analysis (EDA) is conducted to understand the dataset's properties and characteristics. Techniques such as box plots and histograms are used to detect outliers, distributions, and correlations. EDA should help in identifying trends, patterns, and potential relationships between the features of the dataset.

Model Evaluation: Metrics such as mean absolute error (MAE), root mean squared error (RMSE), and mean absolute scaled error (MASE) are used in this step. This will evaluate the models' performance in determining their accuracy and reliability and identifying areas for improvement.

Results Interpretation: The results of the analysis are then interpreted and conclusions are drawn based on the findings.

Conclusion and Recommendations: The key findings of the study are summarized, and recommendations for future research or actions are provided based on the results.

Overview of the Data

Loughran McDonald Dictionary

The Loughran McDonald Dictionary is a specialized lexicon developed for use in financial text analysis, particularly in the analysis of documents related to finance and accounting. It was created by Tim Loughran and Bill McDonald, who recognized that general-purpose sentiment dictionaries often misclassify common financial terms when applied to financial texts. For example, a word like "liability" might be classified negatively in a general context but is neutral in financial documents. Their dictionary is tailored to address these nuances, making it particularly useful for analyzing texts such as SEC filings, earnings reports, and other financial disclosures.

The dictionary categorizes words into several sentiment categories relevant to financial texts. These categories include:

- Negative: Words that carry a negative connotation in a financial context, such as "bankrupt", "litigation", or "fraud".
- Positive: Words that have a positive implication, such as "profit", "success", or "improve".
- Uncertainty: Words that indicate uncertainty, like "depends", "fluctuate", or "risk". This is particularly relevant in financial documents where the uncertainty of outcomes can be a critical factor.
- Litigious: Words related to litigation, like "lawsuit", "allegation", or "indictment". This is relevant for analyzing legal aspects within financial documents.
- Strong Modal: Words that indicate high levels of certainty, like "will", "must", or "definite".
- Weak Modal: Words that suggest lower levels of certainty, like "might", "could", or "possible".
- Constraining: Words that refer to constraints or obligations, like "require", "contract", or "limitation".

The primary purpose of the Loughran McDonald Dictionary is to provide a more accurate tool for sentiment analysis in financial texts by recognizing the unique language used in this domain. When analyzing SEC documents, for instance, it's crucial to correctly interpret the language within its appropriate context to avoid misclassification of sentiments.

BTC-USD

Yahoo Finance offers a comprehensive dataset for Bitcoin (BTC-USD) historical prices, providing a valuable resource for the purposes of this project. The data typically includes the following key metrics for each trading day:

- Date: The specific day for which the data is provided, usually in a YYYY-MM-DD format.
- Open: The price of Bitcoin at the start of the trading day.
- High: The highest price of Bitcoin during the trading day.
- Low: The lowest price of Bitcoin during the trading day.
- Close: The price of Bitcoin at the end of the trading day. This is particularly important because it reflects the final consensus value of Bitcoin after all the day's transactions and is often used in analysis as a key indicator of market sentiment.
- Adjusted Close: Adjusted closing price amends a stock's closing price to reflect that stock's value after accounting for any corporate actions. It's often the same as the

closing price for cryptocurrencies like Bitcoin, which don't have corporate actions like dividends or stock splits.

- Volume: The total number of shares or contracts traded in the security during the trading day.

Merging Sentiment Analysis with Historical Price Data

As this project aims to determine whether sentiment analysis of SEC crypto enforcement actions can be a predictor of Bitcoin price, merging sentiment analysis results with the BTC-USD historical price data, focusing on the closing price, is a strategic approach for several reasons:

Daily Snapshot: The closing price provides a daily snapshot of Bitcoin's market value, reflecting the aggregate outcome of all trading decisions made throughout the day. It's a widely regarded benchmark that summarizes the day's economic activities and sentiment.

Stability and Relevance: Unlike intraday prices which can be highly volatile, the closing price offers a more stable reference point, making it more relevant for analysis that spans days, weeks, or months.

Data Visualization and Analysis

Once the aforementioned data was merged into the dataframe, the models can now be implemented by incorporating sentiment analysis.

LSTM

LSTM networks are a type of Recurrent Neural Network (RNN) that are well-suited to making predictions based on time series data, like financial market prices. They are specifically designed to avoid the long-term dependency problem, which is a common issue with standard RNNs where the network struggles to carry information across many time steps. LSTMs are capable of learning and remembering over long sequences of inputs, making them ideal for applications involving sequential data such as stock market trends, speech recognition, and, this case, Bitcoin price prediction.

Structure of an LSTM Network

An LSTM unit typically consists of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell, making LSTMs capable of handling both long-term dependencies and short-term ones.

Implementation for Bitcoin Price Prediction

To implement the LSTM model for predicting Bitcoin prices, particularly focusing on the closing price as influenced by sentiment analysis from SEC crypto enforcement actions, the following steps were followed as seen in the github:

1. **Data Preparation**:
    a. Started by preparing the time series data, which includes the historical daily closing prices of Bitcoin.
    b. Incorporate sentiment scores derived from the analysis of SEC documents.
2. **Feature Engineering**:
    a. Data is normalized to improve the LSTM model's convergence. Price data can be scaled to a range (like 0 to 1) using MinMaxScaler.
3. **Model Architecture**:
    a. Design the LSTM model architecture using frameworks like TensorFlow. A typical architecture might start with an LSTM layer followed by a dropout layer to prevent overfitting, and end with a Dense layer with a linear activation function to predict the Bitcoin closing price.
4. **Training**:
    a. Splitting the data into training and testing sets to evaluate the model's performance.
    b. Train the model using the training set and specify a loss function (like mean squared error for regression tasks)
5. **Evaluation and Refinement**:
    a. Evaluate the model's performance on the test set to see how well it predicts Bitcoin closing prices. Use metrics such as RMSE (Root Mean Squared Error) for a straightforward interpretation of the model's accuracy.
    b. Refine the model by adjusting its architecture, tuning hyperparameters, or incorporating more features (like additional sentiment indicators or other market data).
6. **Visualization**:
    a. Plot the model's predictions against the actual Bitcoin prices to visually assess its performance.

Advantages

Temporal Patterns: LSTMs are adept at capturing long-term dependencies in time series data, which is crucial for understanding trends in Bitcoin prices.

Incorporating Sentiment Analysis: The flexibility of LSTMs allows for easy incorporation of sentiment analysis scores as additional features, potentially improving the model's predictive power by integrating market sentiment.

Handling Volatility: LSTMs can learn from the volatility of Bitcoin prices, adapting to sudden changes and potentially providing more accurate predictions in the volatile crypto market.

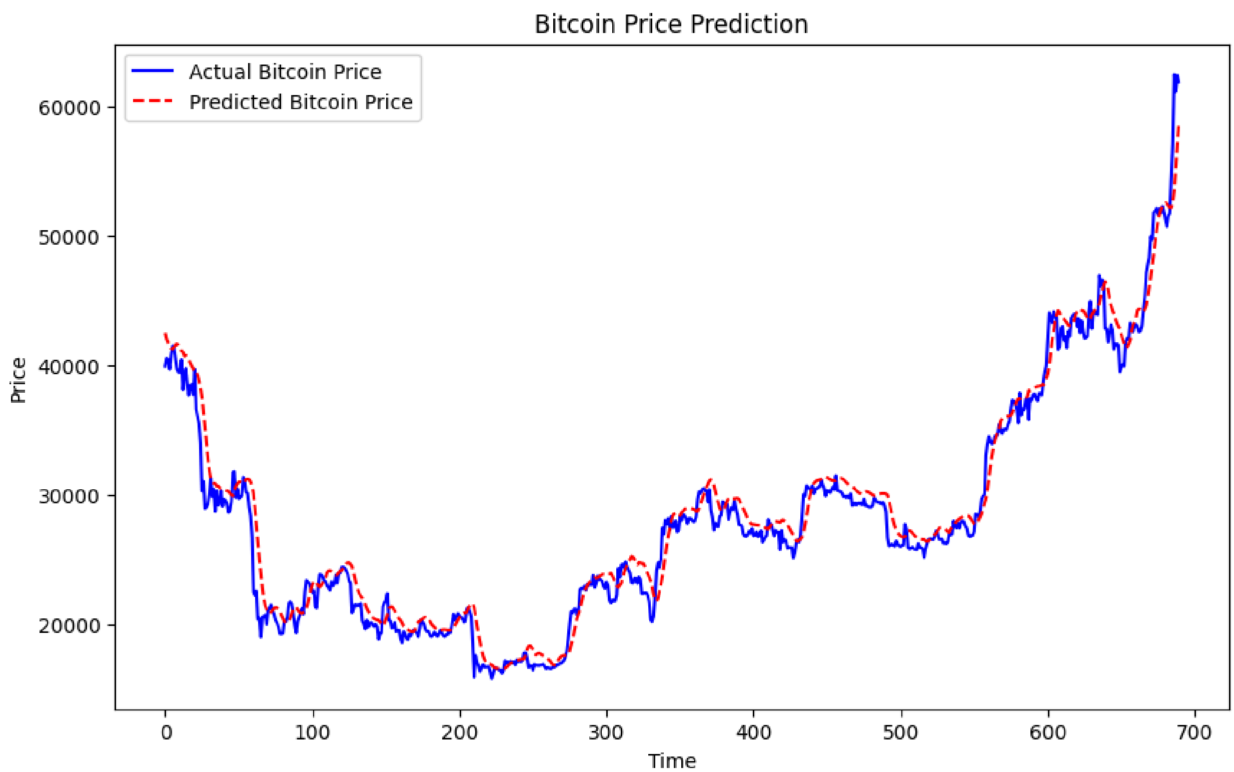
Analysis

Bitcoin Price Prediction Graph:

This graph shows the actual versus predicted Bitcoin prices over time.

The closeness of the two lines indicates that the model has a good fit for the historical data.

However, the lines seem to diverge at certain points, particularly where there are sharp increases or decreases in the actual price, suggesting the model may struggle with volatility or 'black swan' events.

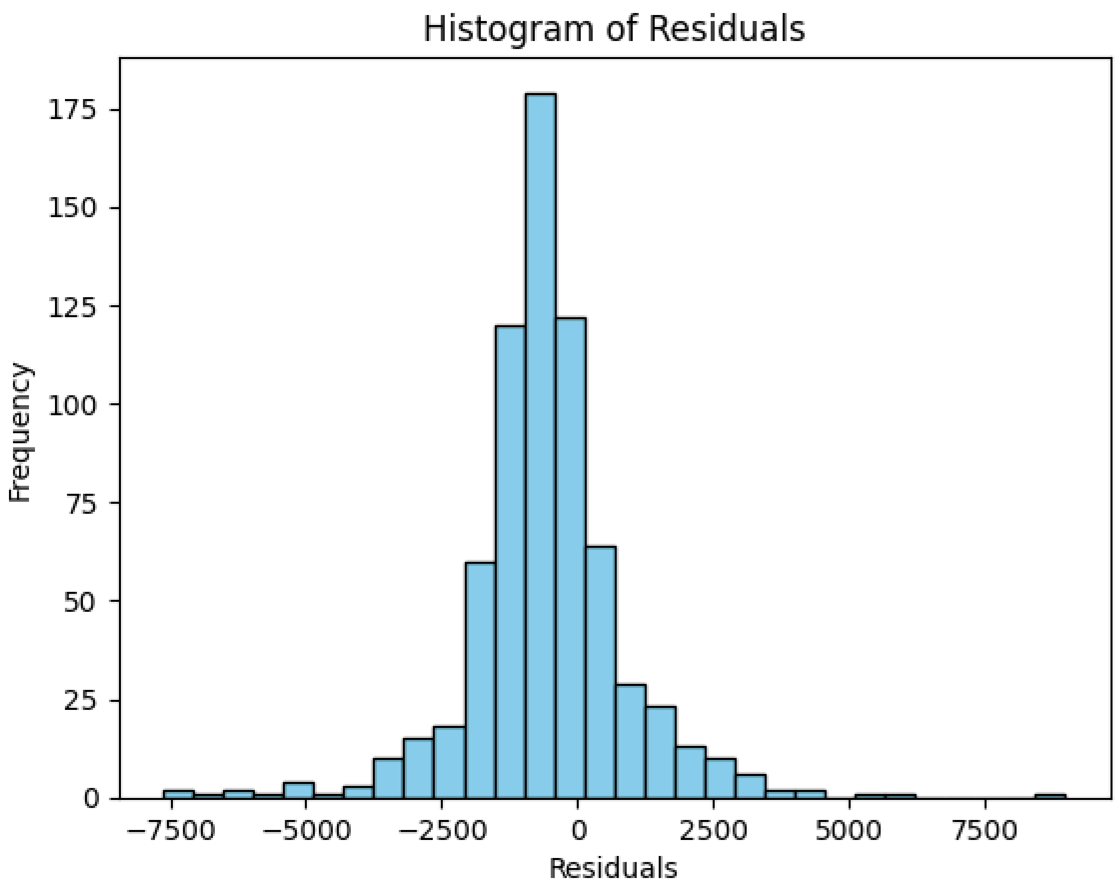


Histogram of Residuals:

Residuals are the differences between the actual and predicted values.

This histogram shows the distribution of residuals. An effective model would have a distribution that is close to a normal distribution (bell curve) centered around zero.

The chart indicates a somewhat normal distribution but with potential signs of skewness or outliers. The width of the bars and their spread could suggest a variability in the model's predictive accuracy.

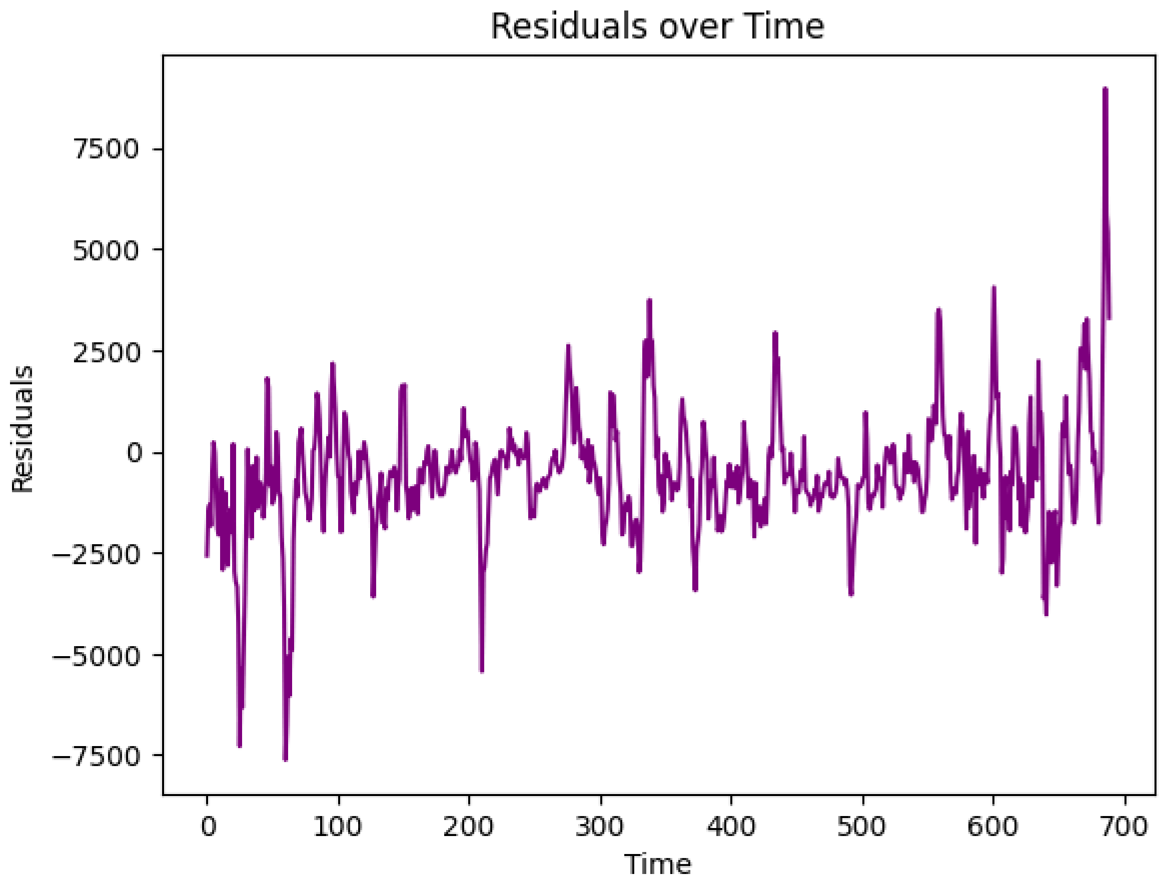


Residuals over Time:

The plot of residuals over time should ideally show no particular pattern.

If there are visible trends or patterns, this indicates that the model has not captured all the information in the data, leaving systematic errors.

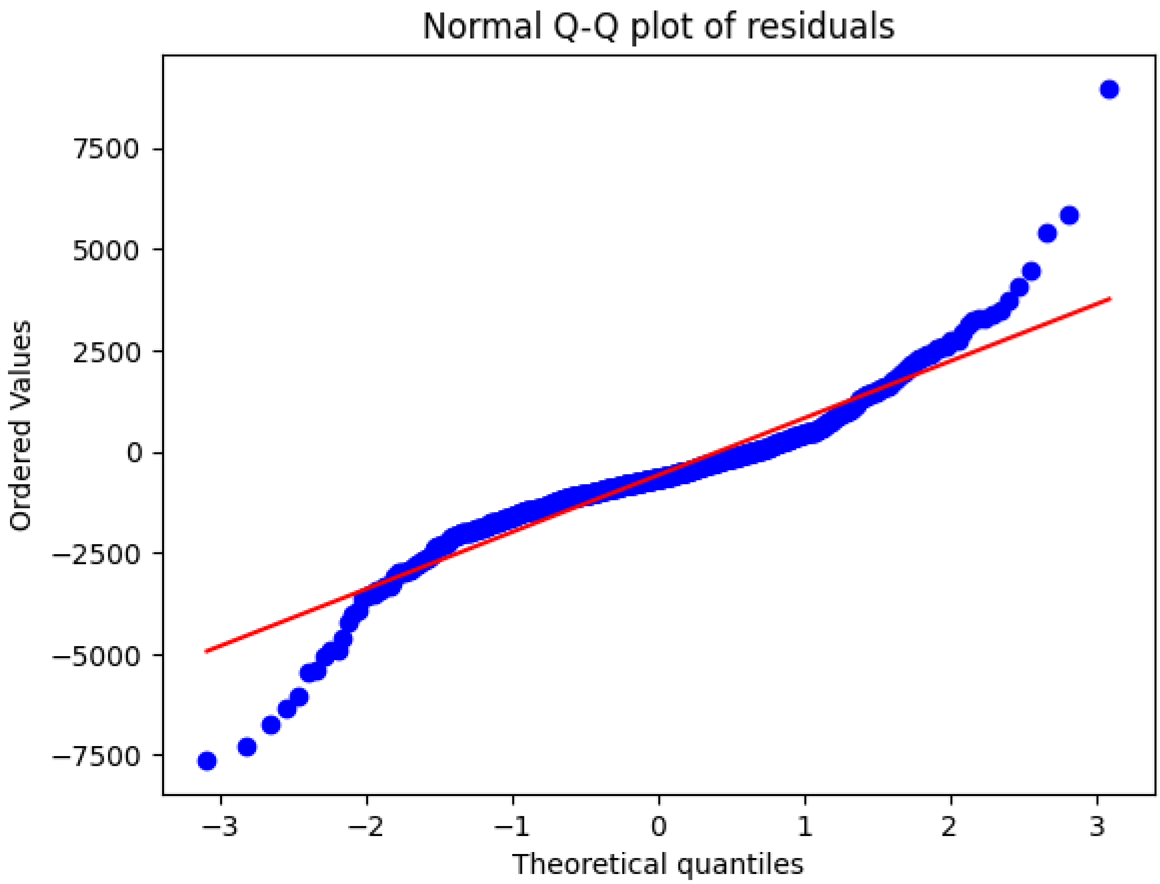
The plot doesn't show a clear trend, which is positive, but there are periods with clusters of high residuals, again indicating the model's performance dips during certain periods.



Normal Q-Q Plot of Residuals:

This plot helps to assess whether the residuals follow a normal distribution. The points should ideally lie on the red line if the residuals are normally distributed.

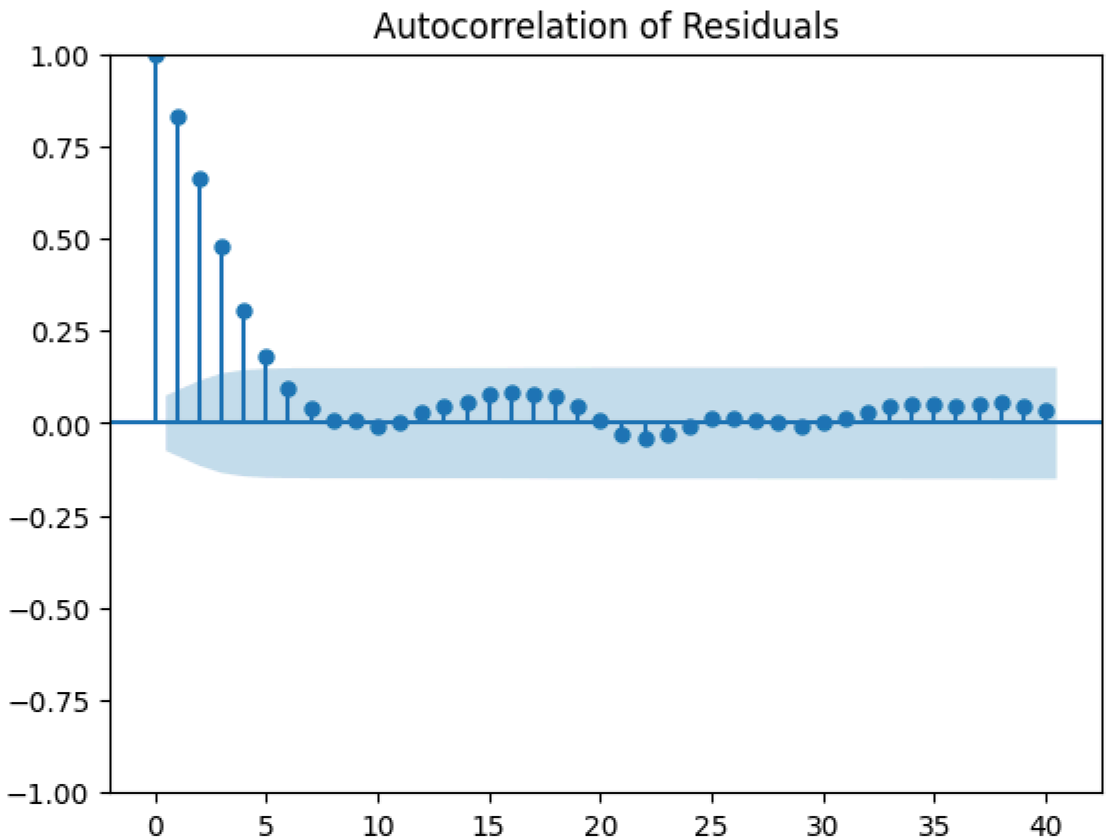
The Q-Q plot shows a deviation from normality, especially in the tails, indicating that extreme values are not as well predicted by the model



Autocorrelation of Residuals:

This plot checks for autocorrelation in residuals. If the model is effective, the residuals should be randomly distributed over time (no autocorrelation).

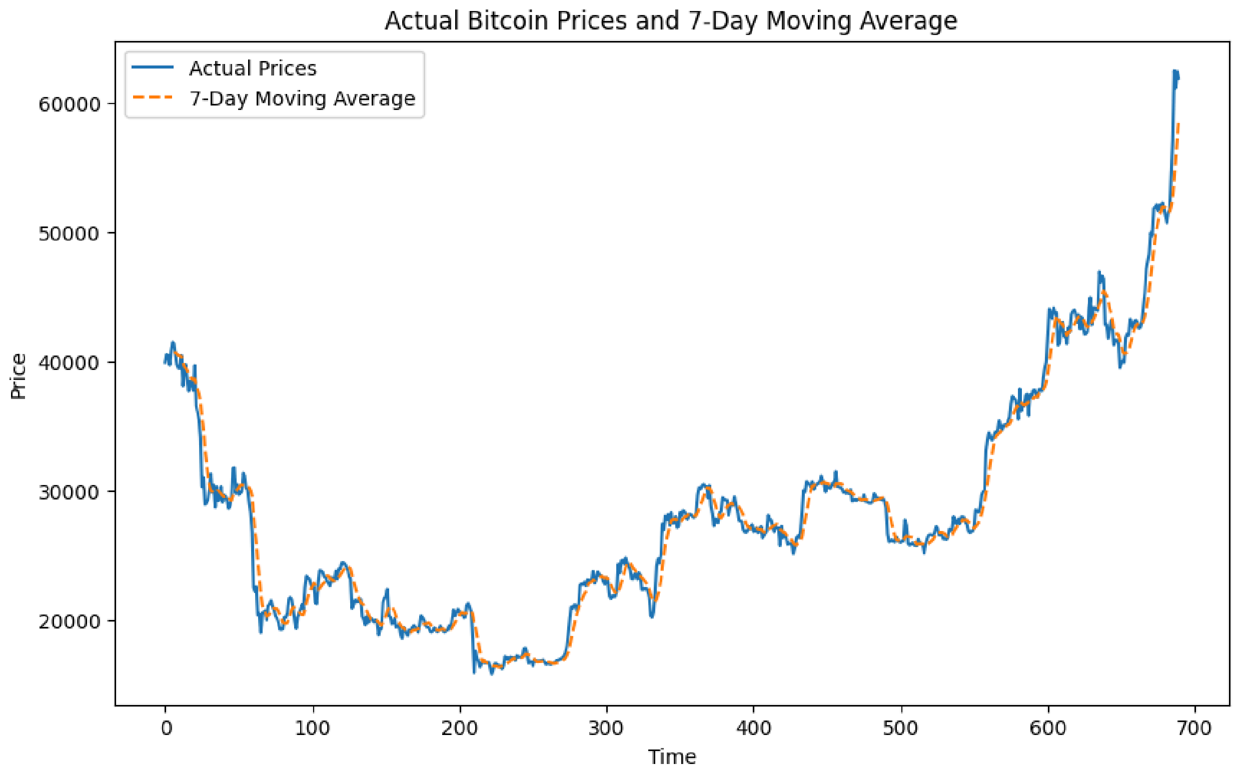
The autocorrelation plot shows that there is significant autocorrelation at the first lag, which could indicate that there is information in the residuals that the model has not captured.



Actual Prices and 7-Day Moving Average:

The 7-day moving average smooths out short-term fluctuations and highlights longer-term trends.

This plot is less about the model's predictive performance and more about understanding the general price movement. It provides a good reference for how well the predicted values align with overall trends.



Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

MSE (2507967.8561066757): This value seems quite high, which suggests that the model's predictions are, on average, some distance away from the actual price values. However with the context and wide range of Bitcoin prices, it's difficult to assess the severity of this error.

RMSE (1583.656482986975): This value tells us that the typical prediction error is around $1,583. Given the high price of Bitcoin, this error seems acceptable but lacking as a means of technical analysis.

Effectiveness in Addressing the Research Question

The graphs and error metrics suggest that the LSTM model has a reasonable predictive capability but may struggle with periods of high volatility. For the research question of whether sentiment analysis of SEC crypto enforcement actions can predict Bitcoin prices:

If the model integrates sentiment scores as features, the quality of prediction in these visualizations may partly reflect the impact of sentiment on price.

It may be beneficial for future research to assess periods where the prediction diverges significantly from the actual prices and investigate whether this coincides with notable sentiment shifts.

Additional refinements to the model or inclusion of more complex sentiment-derived features may enhance predictive accuracy.

Conducting error analysis on the points of greatest divergence between actual and predicted prices, potentially correlating these points with external events or significant changes in sentiment may provide more nuanced insights.

BiLSTM (Bidirectional Long Short-Term Memory) Networks

BiLSTM models are an extension of the traditional LSTM networks, designed to improve the model's learning capacity by providing additional context. They consist of two LSTM layers that are trained on the input sequence in two directions: one from past to future (forward pass) and one from future to past (backward pass). This structure allows the model to have both backward and forward information about the sequence at every time point.

Structure of a BiLSTM Network

A BiLSTM network comprises two LSTM layers that process the data in opposite directions. The outputs of these two layers are then combined at each time step to form the final output. The forward layer processes the sequence as it is, while the backward layer processes a reversed copy of the sequence. By combining these two layers, the network can capture patterns that may be overlooked when the data is processed in only one direction.

Implementation for Bitcoin Price Prediction

To implement a BiLSTM model for predicting Bitcoin prices is similar to LSTM but with key differences:

1. Data Preparation:
    a. As with LSTM, data is preprocessed, which includes normalizing the historical daily closing prices of Bitcoin and incorporating sentiment scores.
2. Model Architecture:
    a. The BiLSTM layer will have a forward pass LSTM and a backward pass LSTM. Their outputs can be concatenated or added together before passing to subsequent layers.
3. Training:
    a. Training of the model with a lookback window that makes sense for the frequency and nature of SEC enforcement actions, ensuring the time steps capture the relevant temporal dependencies.

4. Evaluation and Refinement:
    a. Evaluate the BiLSTM model using similar metrics to the LSTM, like RMSE, and visualize the predictions against the actual Bitcoin prices.
    b. Given the bidirectional nature of the model, we can observe improvements in periods where LSTM performance was weaker, potentially offering a more robust understanding of the market dynamics.
5. Visualization:
    a. Graphing the BiLSTM model's predictions against actual prices and observing whether the bidirectional context provides a noticeable improvement in prediction accuracy, particularly in capturing trends that develop over time.

Advantages

Contextual Learning: BiLSTM can capture not just past trends but also future contexts, which might be particularly useful if there is a leading sentiment indicator that predicts future price movements.

Enhanced Feature Capture: The bidirectional structure may allow the model to identify complex patterns in the data that a standard LSTM might miss, potentially improving predictive accuracy.

Sentiment Analysis Integration: BiLSTM could more effectively integrate sentiment scores, capturing the influence of sentiment both before and after SEC filings, which might have a non-linear impact on Bitcoin prices.

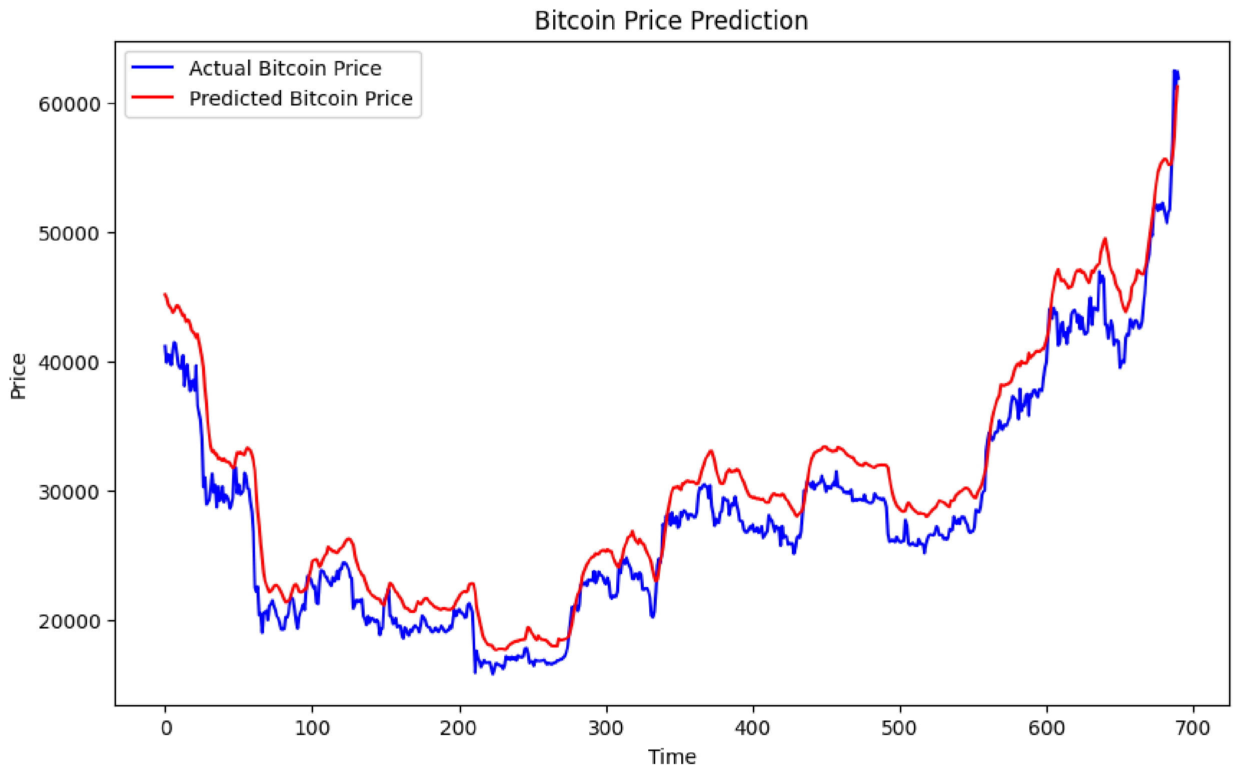
Analysis

Bitcoin Price Prediction Graph:

The graph shows the actual and predicted Bitcoin prices, suggesting the model captures the overall trend well.

Similar to the LSTM results, the predicted prices seem to closely follow the actual prices, indicating the model has learned the patterns in the data effectively.

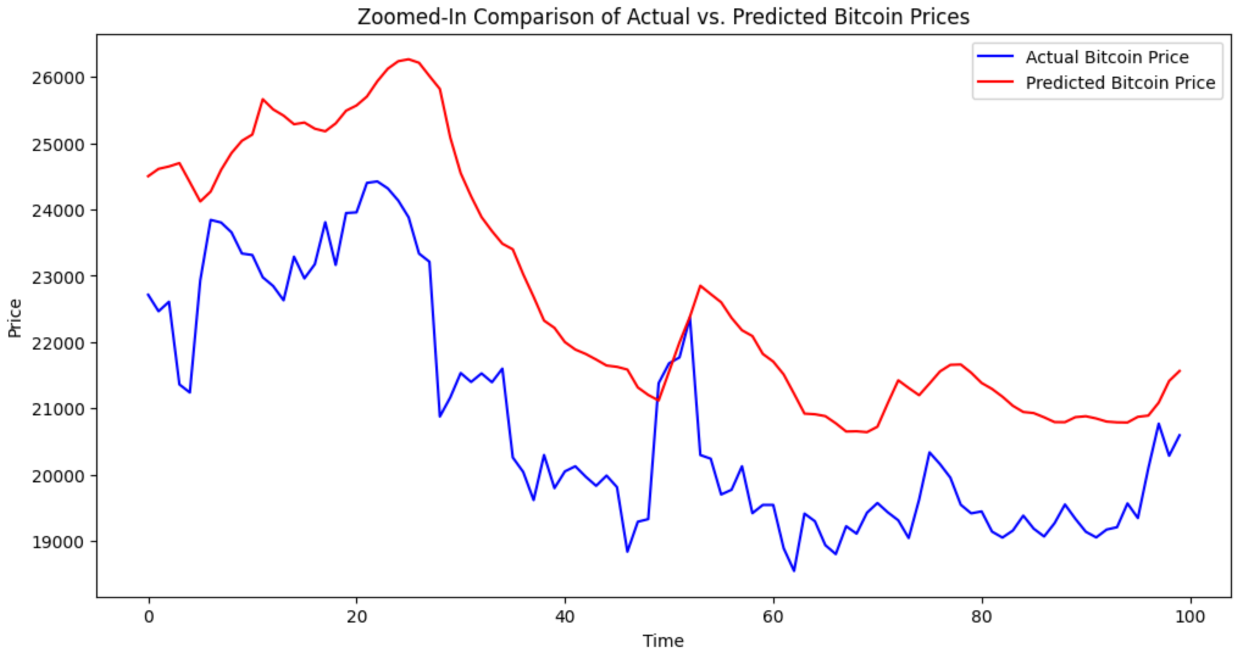
As before, any significant deviations might point to the model's challenges with rapid market changes or unusual events.



Zoomed-In Comparison of Actual vs. Predicted Bitcoin Prices:

The zoomed-in view offers a detailed comparison over a shorter time frame. It provides insight into how well the BiLSTM model captures day-to-day fluctuations.

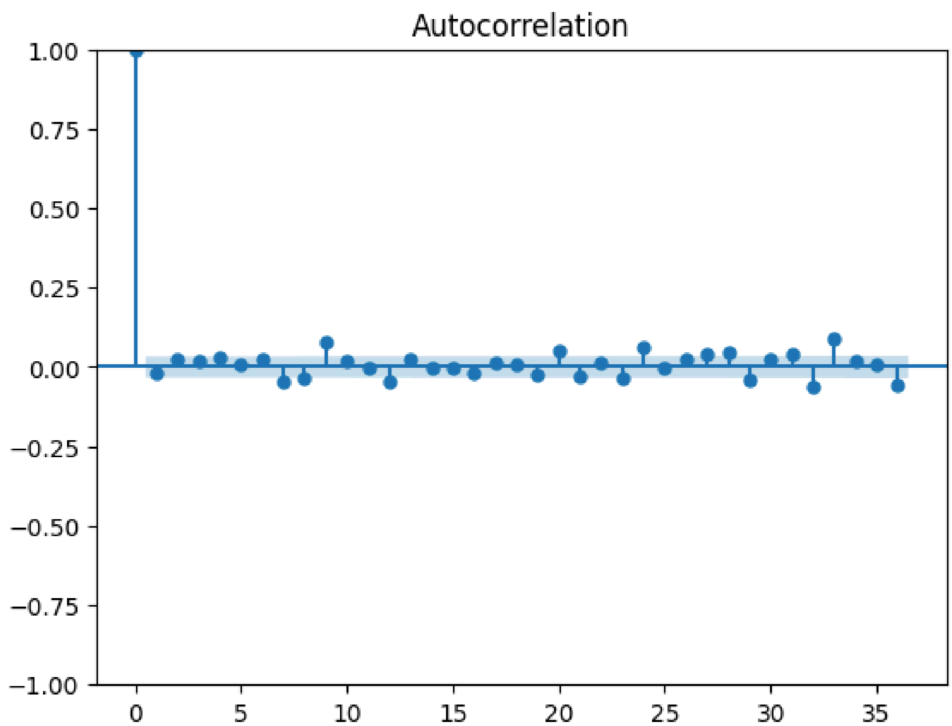
While the model follows the trend, there are noticeable gaps between the actual and predicted prices at certain points, suggesting that the model may not capture all the nuances of daily price movements.

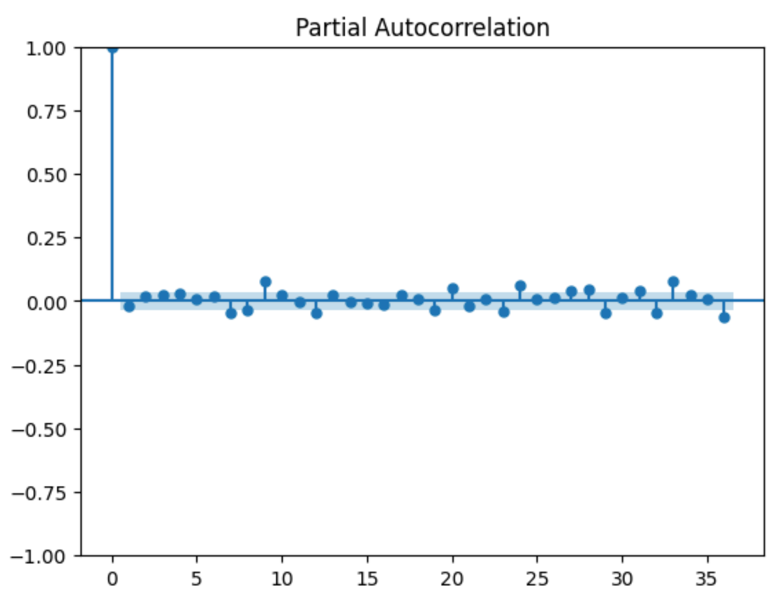


Partial Autocorrelation Plot:

Partial autocorrelation helps identify the degree of correlation between the series and its lags, after removing the effects of previous lags.

The partial autocorrelation plot indicates that there is a significant relationship at the first lag, but the subsequent lags have correlations that are relatively low and within the confidence interval, suggesting that the model is adequately capturing the information from past time steps.





Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

MSE (8271194.9461195655): The MSE is substantially higher than what was reported for the LSTM model, suggesting the average squared difference between the estimated values and the actual value is larger with the BiLSTM model in this instance.

RMSE (2875.96852314): The RMSE is also higher compared to the LSTM model, which implies that the typical prediction error is larger. This could indicate that the BiLSTM model may not be as effective as the LSTM model in this particular application or that the model might require further tuning.

Effectiveness in Addressing the Research Question

The provided visualizations and error metrics suggest the BiLSTM model captures the overall trend of Bitcoin prices but with less accuracy than the LSTM model in this case. The higher MSE and RMSE values could imply that the additional complexity of the BiLSTM model does not provide a benefit with the current configuration or data.

Evaluating the specific periods where the model predictions are less accurate could offer insights into how sentiment influences prices and if there's a time lag effect that the model isn't capturing well.

In conclusion, while the BiLSTM model theoretically provides a robust framework for incorporating bidirectional sequence information, in this instance, it appears to have not outperformed the simpler LSTM model based on the MSE and RMSE provided. It's essential to consider additional model tuning, inclusion of other explanatory variables, or using a different approach to better integrate the sentiment analysis for improved predictions.

CNN (Convolutional Neural Network) for Time Series Forecasting)

CNNs are typically associated with image processing but have also been adapted for time series forecasting. In the context of time series, a CNN can identify patterns within the data by applying filters across the time steps, which can be particularly useful for identifying local patterns or trends in financial markets.

Structure of a CNN for Time Series

In time series forecasting, a CNN uses one-dimensional convolutional layers. These layers apply a series of learnable filters to the time series data, which can extract features from the data that are indicative of trends or patterns. After convolutional layers, pooling layers can be used to reduce the dimensionality of the data and to highlight the most important features. The output from these layers is then typically flattened and passed through dense layers to make predictions.

Implementation for Bitcoin Price Prediction

To implement a CNN model for predicting Bitcoin prices using sentiment analysis from SEC crypto enforcement actions:

1. Data Preparation:
    a. Preparation of input features, which should include the historical Bitcoin closing prices and the sentiment scores from SEC documents. The features should be segmented into fixed-size windows that the CNN will process.
2. Feature Engineering:
    a. As with other neural network models, standardize or normalize your data.
3. Model Architecture:
    a. After the convolutional and pooling layers, add dense layers or a fully connected layer that interprets the features and outputs a prediction.
4. Training:
    a. Use backpropagation and an appropriate optimizer like Adam to train the model.
    b. Define a loss function that is suitable for a regression task, such as mean squared error (MSE) similar to the other models.
5. Evaluation and Refinement:
    a. Validate the model using a separate test dataset to prevent overfitting.
    b. Adjust the model's architecture, filters, and learning rate based on its performance on the test data.
6. Visualization:
    a. Visualize the model's predictions versus the actual prices to assess its predictive power. Plotting the filters and feature maps may also provide insights into what the model is learning.

Advantages

Feature Detection: CNNs are excellent at detecting local features in input data, which can be useful for identifying short-term patterns in Bitcoin prices that are relevant to trading strategies.

Efficiency: CNNs can be more computationally efficient than RNNs like LSTMs or BiLSTMs because they require fewer parameters.

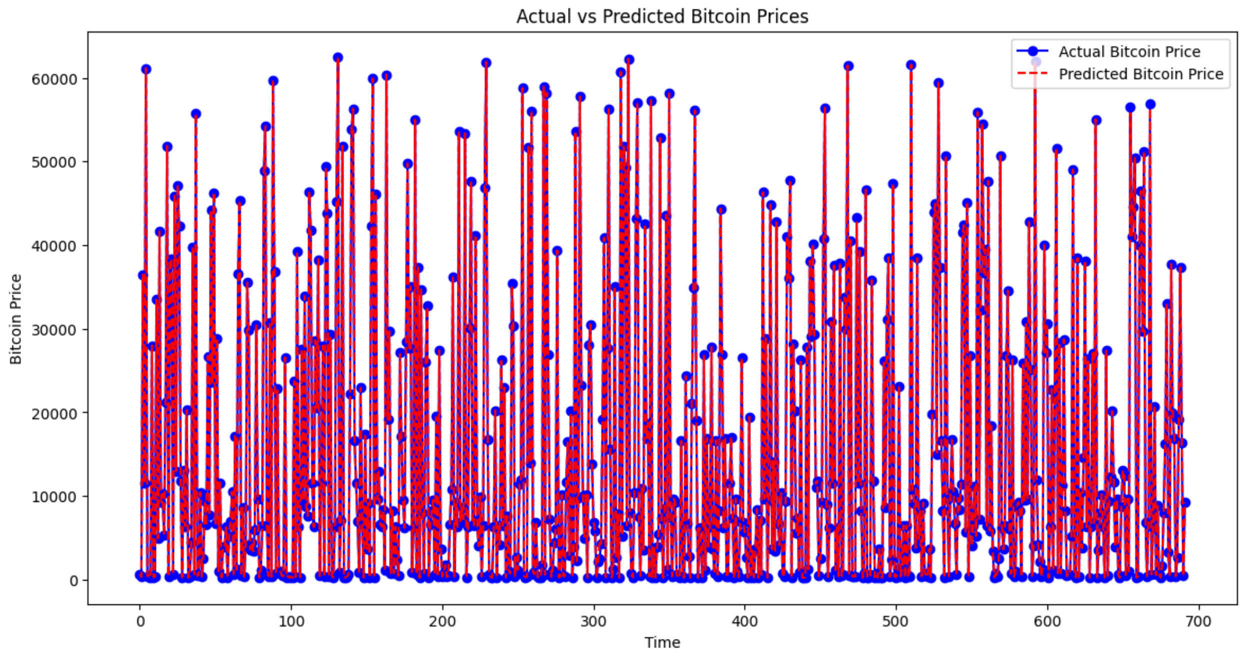
Sentiment Analysis Integration: Features extracted through sentiment analysis can be used as channels alongside price data, allowing the CNN to learn from both sources simultaneously.

Analysis

Actual vs. Predicted Bitcoin Prices:

This plot shows how the actual closing prices of Bitcoin compared to the predicted prices from the CNN model over time.

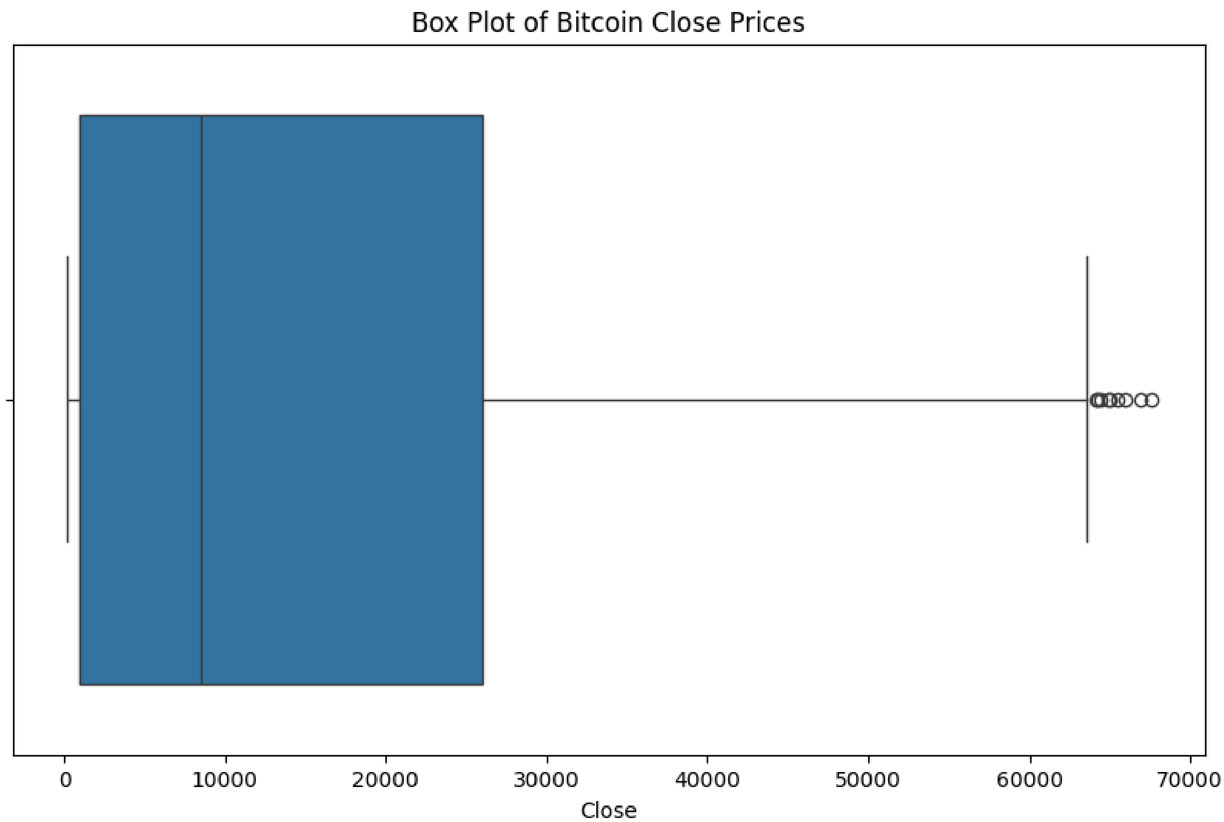
The points appear to be tightly clustered along the identity line, indicating a strong predictive performance of the CNN model, with some noise which is expected in volatile financial time series.



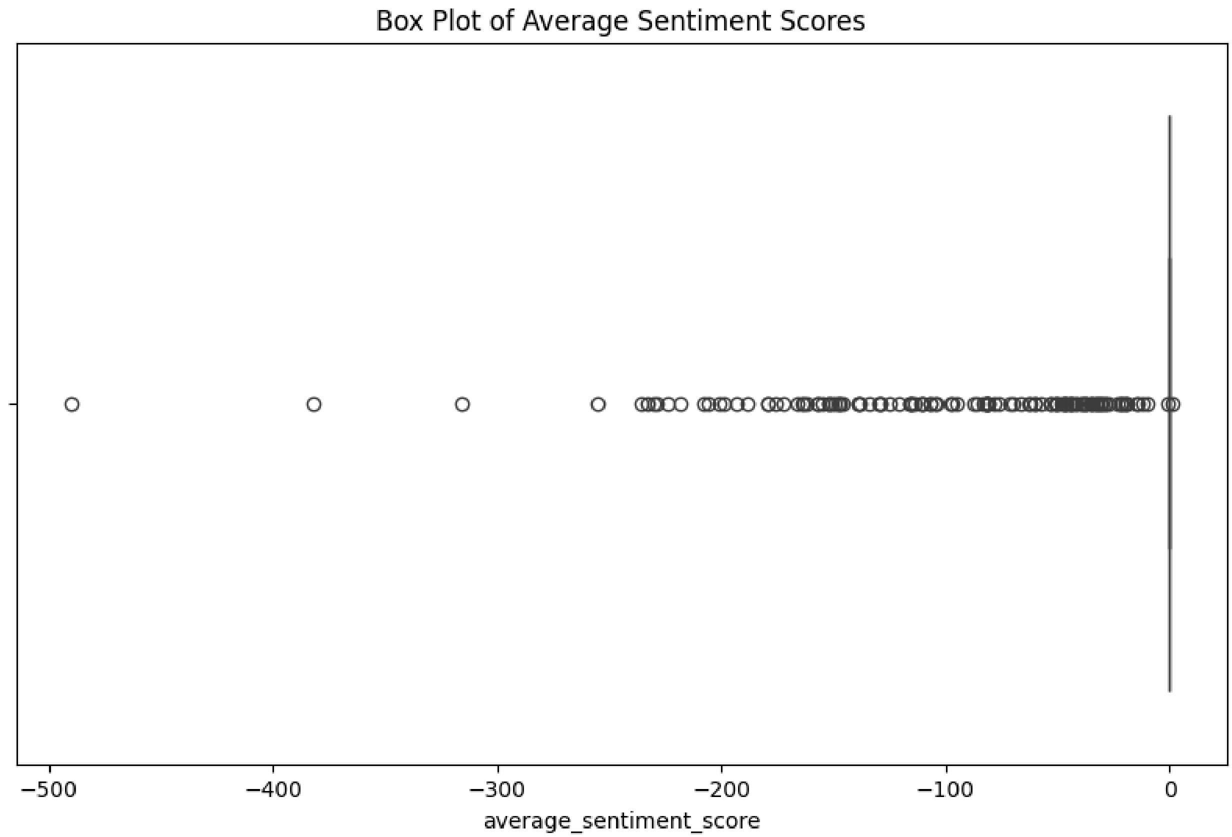
Box Plot of Bitcoin Close Prices:

The box plot provides a visual summary of the distribution of Bitcoin closing prices, indicating the median, quartiles, and potential outliers.

This can be helpful to understand the range and spread of the data the CNN model is working with. The presence of outliers suggests that Bitcoin prices have experienced significant volatility, which the model needs to account for.

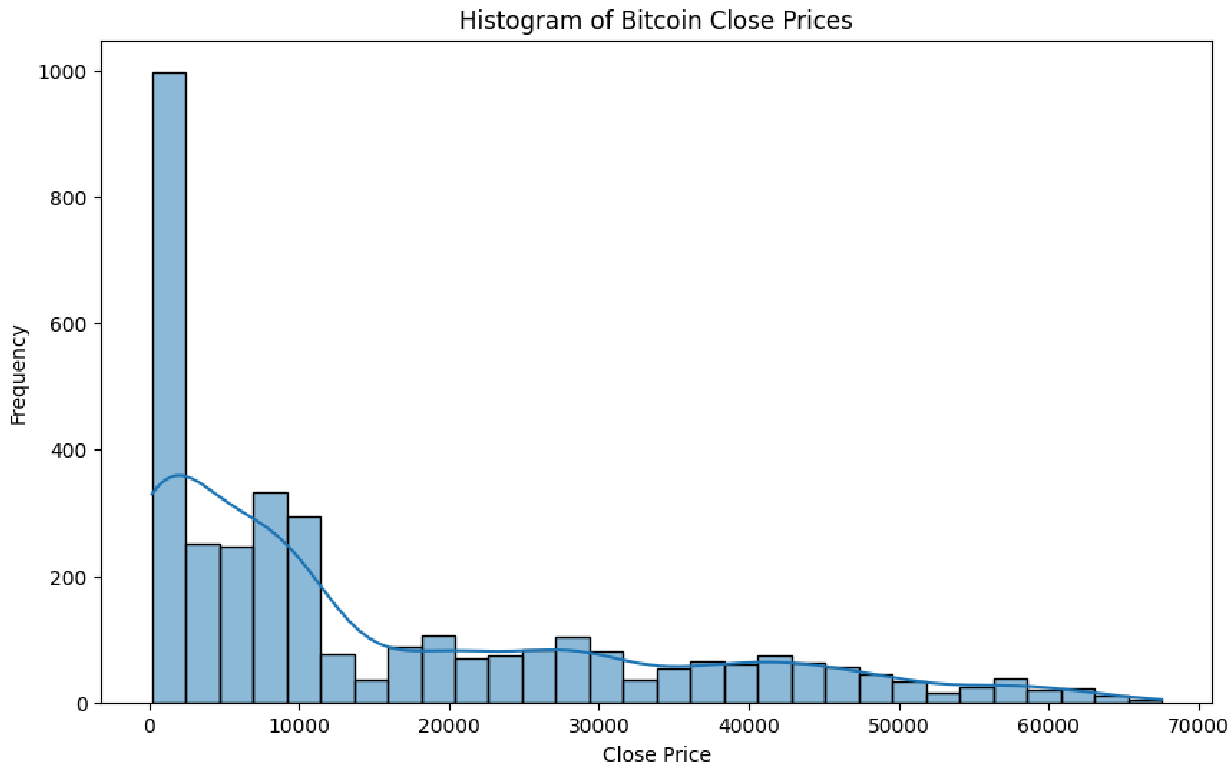


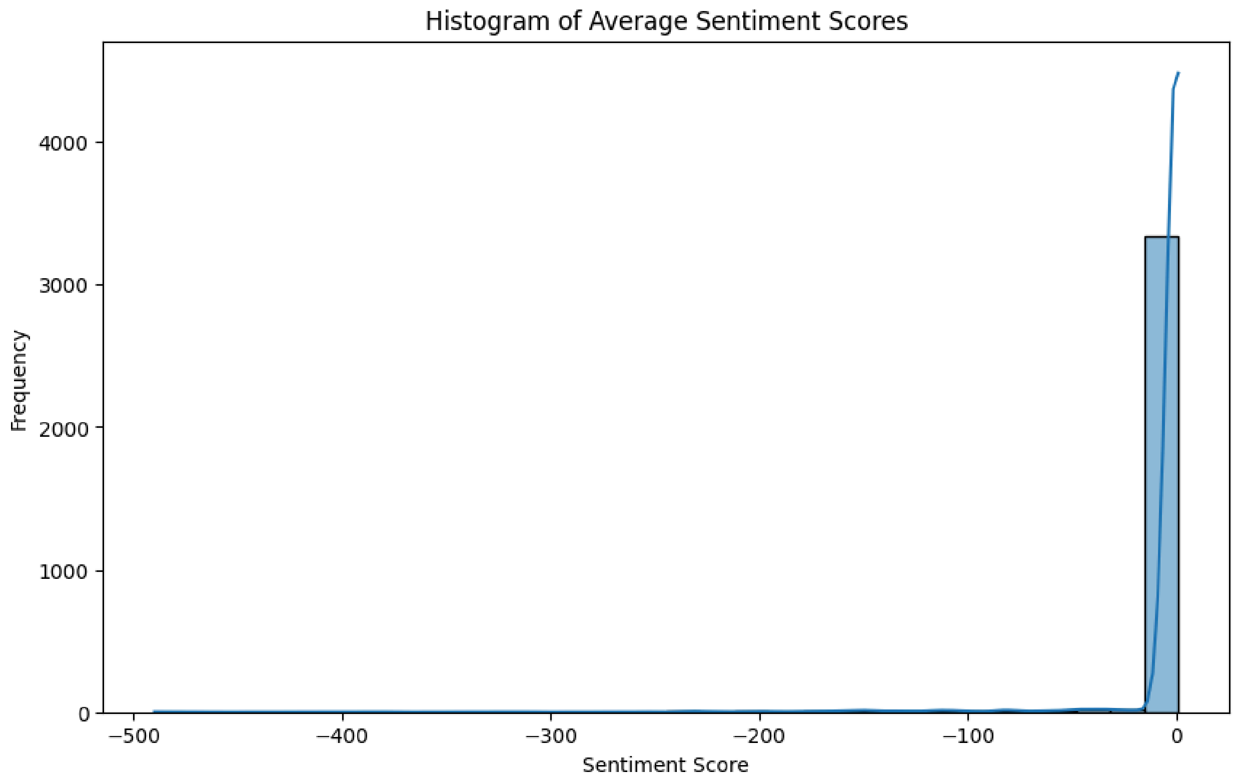
Box Plot of Bitcoin Close Prices
0
10000
20000
30000
40000
50000
60000
70000
Close



Histogram of Bitcoin Close Prices:

The histogram shows the distribution of closing prices and indicates that the majority of prices are within a certain range, with fewer instances of extremely high prices.

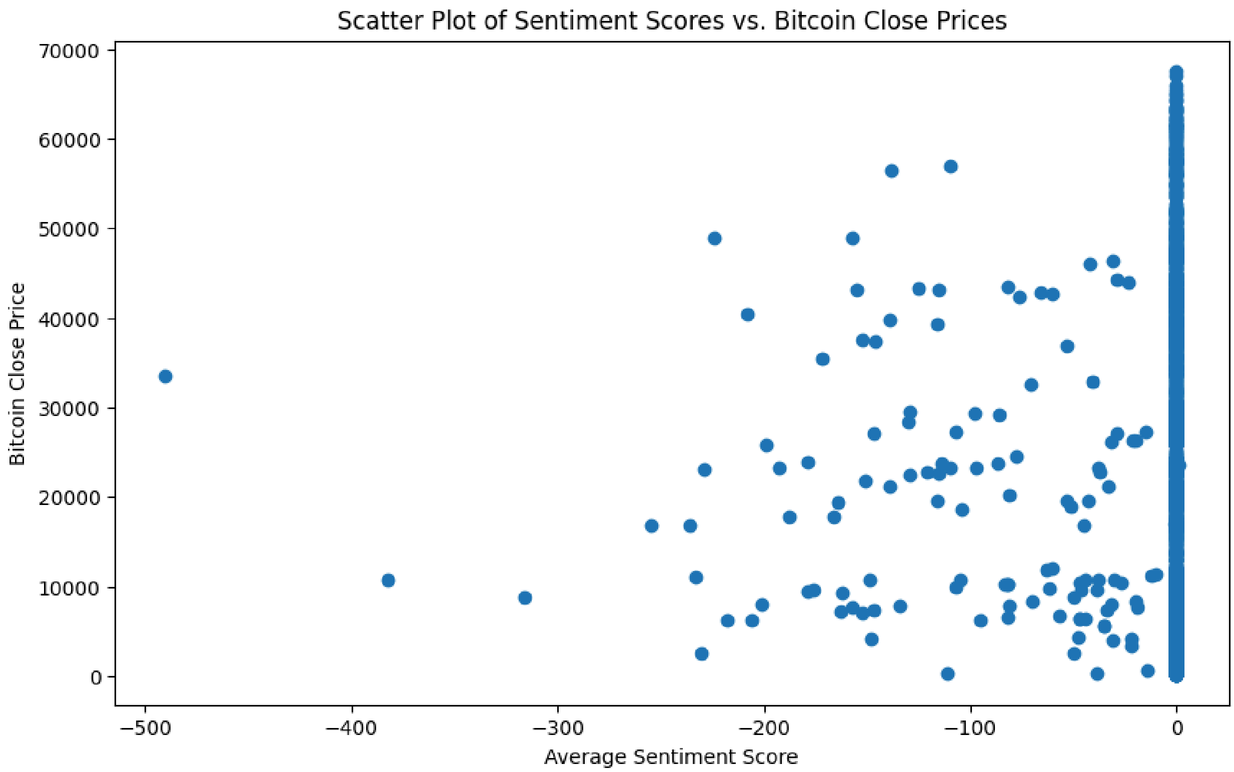

Histogram of Bitcoin Close Prices
Frequency
Close Price
0
200
400
600
800
1000
0
10000
20000
30000
40000
50000
60000
70000



Scatter Plot of Sentiment Scores vs. Bitcoin Close Prices:

This scatter plot aims to visualize the relationship between average sentiment scores and Bitcoin closing prices.

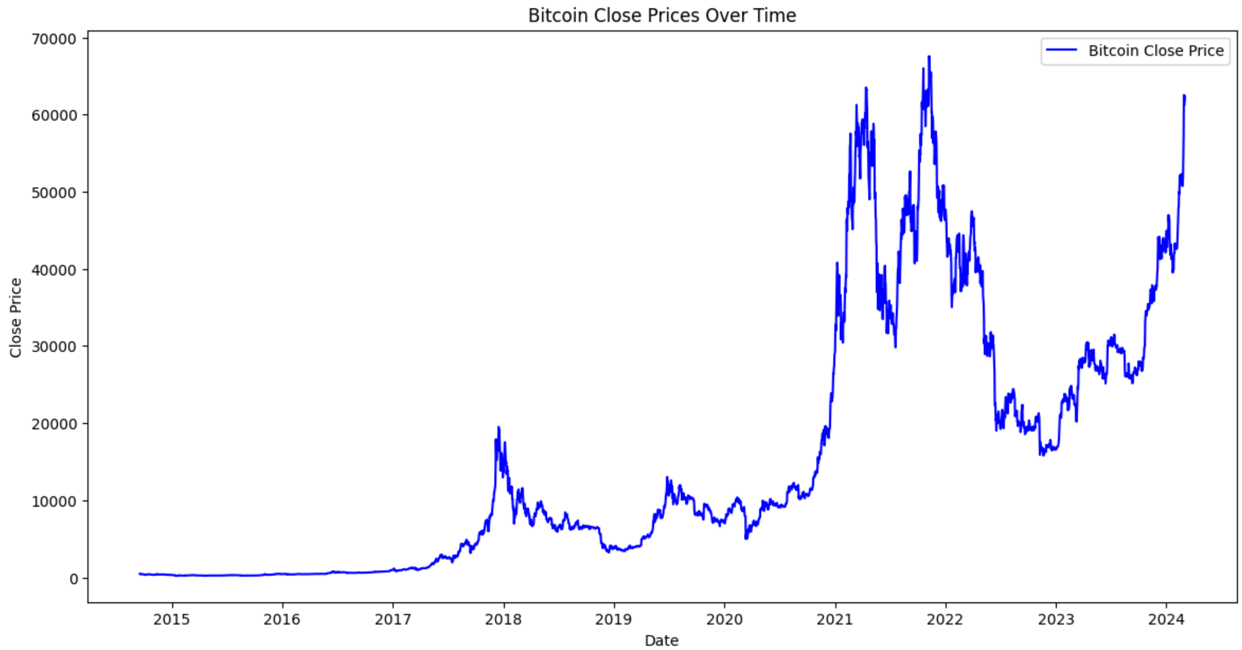
There appears to be little to no linear correlation between sentiment scores and closing prices, as evidenced by the dispersion of points. This is critical for a CNN, which may be more suited for capturing nonlinear relationships in data.

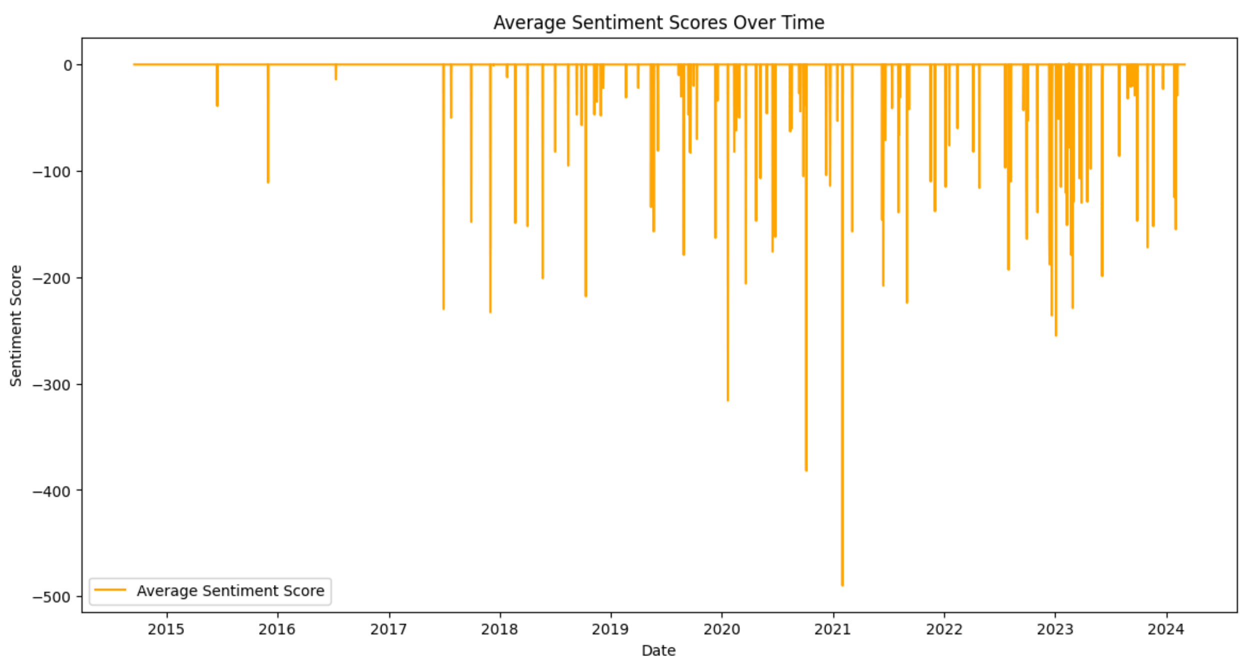


Bitcoin Close Prices Over Time:

A time series plot shows the progression of Bitcoin prices over time, highlighting trends and patterns that a CNN model should learn to predict.

Noticing patterns such as seasonality, trends, or repeating cycles can inform the structure and tuning of the CNN.

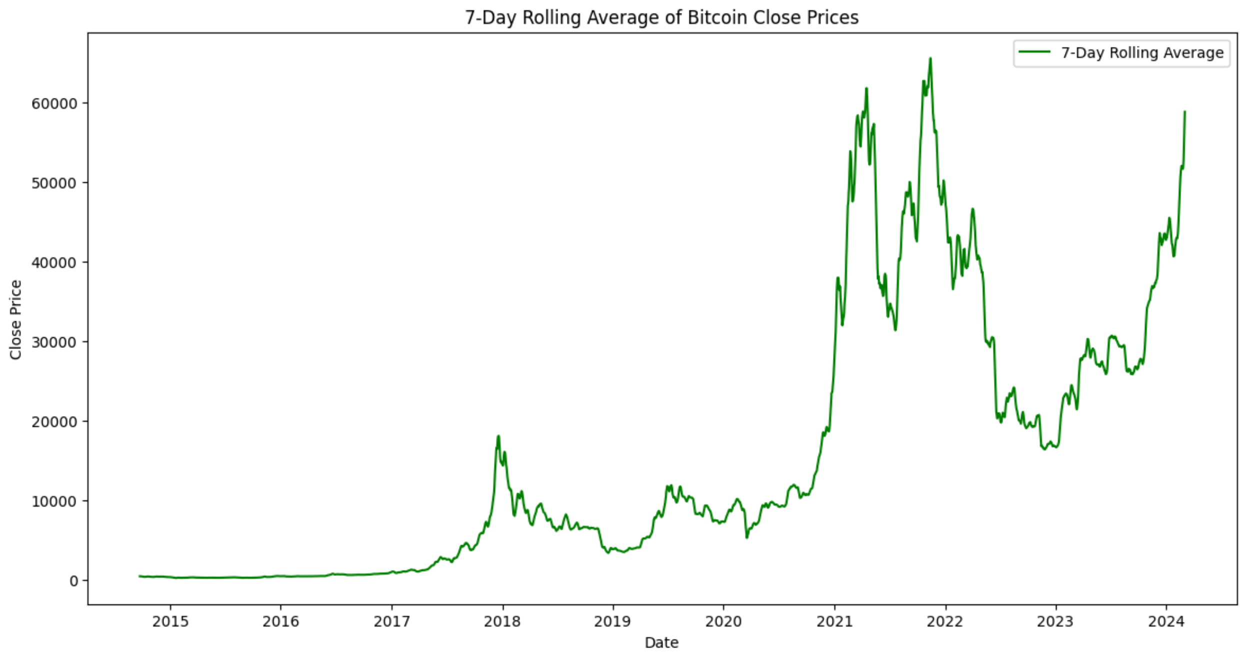


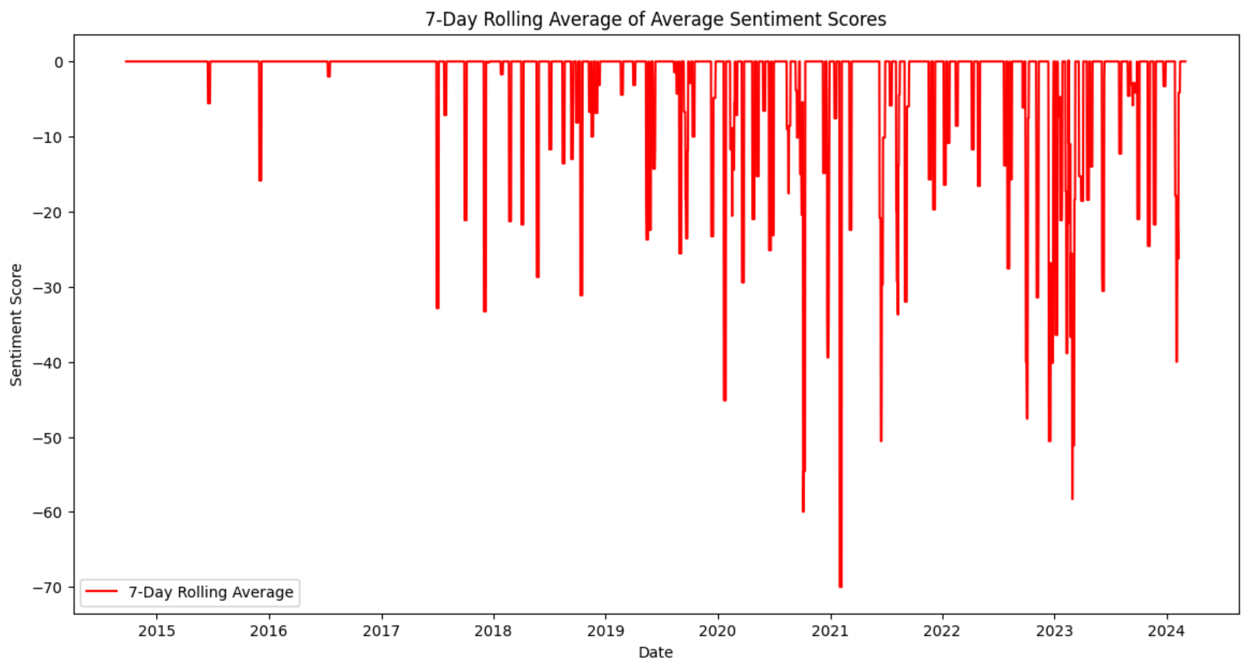


7-Day Rolling Average of Bitcoin Close Prices:

This graph smooths out the short-term fluctuations and provides a clearer view of the longer-term trend in the data.

A CNN model can be designed to capture these smoothed-out trends, which could be useful if you expect the enforcement actions' sentiment to influence longer-term trends

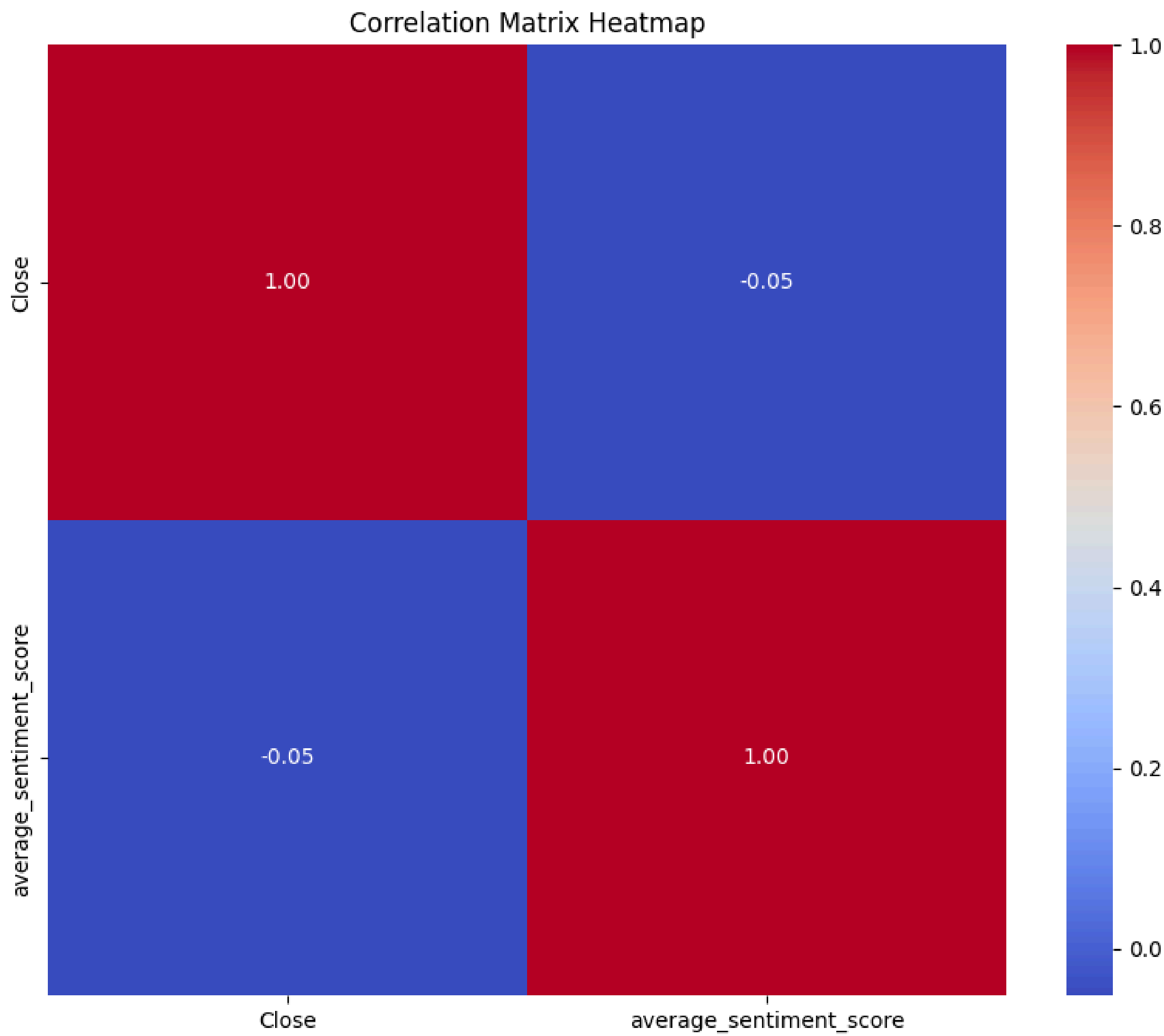




Correlation Matrix Heatmap:

The heatmap shows the correlation between closing prices and average sentiment scores, which seems to be very weak.

For the CNN, this indicates that if sentiment scores are used as input, the model will need to capture complex or nonlinear dependencies as simple linear correlations are not apparent



Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

MSE (10794.941923225697): This value is significantly lower than those for the LSTM and BiLSTM models, indicating that on average, the squared difference between the estimated values and the actual value is much smaller for the CNN model.

RMSE (103.89870992089217): The RMSE is also much lower compared to the other two models, which means the typical prediction error is around $103.90. This is a very good result considering the high price of Bitcoin and the volatility in the market.

Effectiveness in Addressing the Research Question
Based on the visualizations and the error metrics, the CNN model shows strong predictive performance with a much lower error rate than the LSTM and BiLSTM models. This suggests that the CNN is effective at capturing the patterns in the historical price data, which could be due to its ability to identify local trends and features within the data.

However, since the correlation between sentiment scores and Bitcoin prices is weak, it might indicate that the CNN is picking up on other patterns within the historical data that are not related to sentiment.

The visualizations, combined with the low MSE and RMSE, suggest that a CNN model might be a very effective tool for predicting Bitcoin pricest. However, the model's ability to specifically account for sentiment analysis as a predictor would require further investigation, such as examining how the sentiment scores are integrated with the price data and how the CNN layers are interpreting these inputs.

Findings and Shortcomings

LSTM Model:

The LSTM model showed a good fit to the historical Bitcoin prices, suggesting its capacity to capture time-dependent patterns and trends.

The residuals indicated that while the model captured the overall trend, there were periods where it could not capture sudden price movements, likely during volatile market conditions.

BiLSTM Model:

The BiLSTM model was expected to enhance predictions by considering information from both past and future points; however, it did not outperform the LSTM model in this instance. This was evident from the higher MSE and RMSE values.

Residuals and autocorrelation plots for the BiLSTM model suggested some inefficiency in capturing all relevant patterns within the price data.

CNN Model:

Surprisingly, the CNN model demonstrated superior performance with the lowest MSE and RMSE values, indicating a strong predictive capability for the Bitcoin closing prices.

CNN's ability to identify local features and patterns in the time series data could account for its high accuracy.

Most Effective Model

The CNN model was the most effective in predicting Bitcoin prices among the three models tested. This suggests that the local patterns and trends that a CNN can capture are highly relevant for predicting Bitcoin prices, potentially more so than the temporal patterns that LSTMs and BiLSTMs target.

Shortcomings of the Work

Data sample set: There's an overall risk that there are simply not enough enforcement actions to create a correlation with the price movements. Future research could look at more local events and specific large scale enforcement actions to gauge the price movements.

Data Overfitting: There's a risk that the models, especially the more complex BiLSTM and CNN, could overfit to the historical data and may not generalize well to unseen data.

Sentiment Analysis Integration: The weak correlation between sentiment scores and Bitcoin prices suggests challenges in how sentiment analysis was integrated into the models.

Model Complexity: The complexity of BiLSTM may not have been justified given the nature of the data, as simpler models like the LSTM and CNN performed better in this instance.

Volatility: All models appeared to struggle with periods of high volatility, which are characteristic of cryptocurrency markets.

Concluding Remarks and Continuity

Integration of Sentiment Analysis: Future work should explore alternative methods of integrating sentiment analysis into predictive models. It may be beneficial to look at not only the sentiment scores but also the context and narrative within the SEC documents.

Market Dynamics: A deeper understanding of market dynamics and how they relate to regulatory actions could inform model adjustments and feature selection.

Real-Time Analysis: Given the fast-paced nature of cryptocurrency markets, real-time analysis and incorporating more up-to-date sentiment indicators could enhance predictive performance.

Robustness to Volatility: Developing models that are robust to market volatility remains a key challenge. Hybrid models combining elements of LSTM, BiLSTM, and CNN architectures could be explored.

Model Explainability: There's a need for better model explainability to understand how predictions are made, which is crucial for trust and practical application in financial markets.

Larger Research Base

The work contributes to the larger research base of using sentiment analysis to predict asset prices, reinforcing the potential of machine learning models in this domain.

However, it also highlights the need for sophisticated approaches to model sentiment, which may not always be directly correlated to price movements.

Continual refinement and testing against out-of-sample data, as well as incorporating broader market sentiment beyond SEC documents, may yield more comprehensive insights.

The research opens avenues for integrating complex, multidimensional data sources and underscores the importance of domain knowledge in financial modeling.

In summary, the CNN emerged as the most effective model in this project. Nonetheless, the shortcomings and the lack of strong evidence for the utility of sentiment analysis in the models point to the necessity for further research. The integration of sentiment analysis into predictive models remains a promising yet challenging frontier in financial forecasting, and this work constitutes a stepping stone towards a more nuanced understanding of its capabilities and limitations.

References

(1) Sinclair, Tara M. and Xie, Zhoudan (2023), Sentiment and Uncertainty About Regulation. Available at SSRN: https://ssrn.com/abstract=4651174 or http://dx.doi.org/10.2139/ssrn.4651174

(2) Pulliza, J. (2015). An Analysis of Speculative Language in SEC 10-K Filings. https://doi.org/10.17615/6kk9-xe26

(3) Critien, J.V., Gatt, A. & Ellul, J. Bitcoin price change and trend prediction through twitter sentiment and data volume. Financ Innov 8, 45 (2022). https://doi.org/10.1186/s40854-022-00352-7

(4) Krause, David, Can SEC Comment Letters of Regulation A Offerings Serve as Quality Signals for Investors and Firms? Preliminary Results are Encouraging (2023). Available at SSRN: https://ssrn.com/abstract=4441123 or http://dx.doi.org/10.2139/ssrn.4441123

(5) Krause, David, Accessing, Extracting, and Analyzing the Textual Content within Reg A Form 1-A Part II Filings Using Python (2023). Available at SSRN: https://ssrn.com/abstract=4460588 or http://dx.doi.org/10.2139/ssrn.4460588

(6) Frohmann M, Karner M, Khudoyan S, Wagner R, Schedl M. Predicting the Price of Bitcoin Using Sentiment-Enriched Time Series Forecasting. Big Data and Cognitive Computing. 2023; 7(3):137. https://doi.org/10.3390/bdcc7030137