

Article

Enhancing Bitcoin Price Prediction with Deep Learning: Integrating Social Media Sentiment and Historical Data

Hla Soe Htay , Mani Ghahremani  *  and Stavros Shiaeles 

School of Computing, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, UK; up2232262@myport.ac.uk (H.S.H.); stavros.shiaeles@port.ac.uk (S.S.)

* Correspondence: mani.ghahremani@port.ac.uk

Abstract: Bitcoin, the pioneering cryptocurrency, is renowned for its extreme volatility and speculative nature, making accurate price prediction a persistent challenge for investors. While recent studies have employed multivariate models to integrate historical price data with social media sentiment analysis, this study focuses on improving an existing univariate approach. By incorporating sentiment and tweet volume data into a multivariate framework, we systematically evaluated the benefits of this integration. Among the five LSTM-based models developed for this study, the Multi-LSTM-Sentiment model achieved the best performance, with the lowest mean absolute error (MAE) of 0.00196 and root-mean-square error (RMSE) of 0.00304. These results underscore the significance of including social media sentiment in predictive modelling and demonstrate its potential to enhance decision-making in the highly dynamic cryptocurrency market.

Keywords: bitcoin; LSTM; cryptocurrency forecasting; twitter sentiment analysis; multivariate time series

1. Introduction



Academic Editor: Eui-Nam Huh

Received: 20 December 2024

Revised: 28 January 2025

Accepted: 31 January 2025

Published: 4 February 2025

Citation: Htay, H.S.; Ghahremani, M.; Shiaeles, S. Enhancing Bitcoin Price Prediction with Deep Learning: Integrating Social Media Sentiment and Historical Data. *Appl. Sci.* **2025**, *15*, 1554. <https://doi.org/10.3390/app15031554>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Bitcoin, introduced in 2009, is the first decentralized cryptocurrency that leverages blockchain technology to enable secure and anonymous online transactions. Known for its high volatility—it is about eight times more volatile than the stock market and 20 times more volatile than the US dollar [1]—Bitcoin has experienced significant price fluctuations. From a trading price of USD 0.06 in July 2010, it reached a peak of USD 69,000 in November 2021 before declining to USD 16,530 by the end of 2022. On 14 March 2024, it hit a record high of USD 75,830, indicating its incredible appreciation and increasing importance in the financial market [2]. Due to its capped supply of 21 million coins, Bitcoin is sometimes referred to as “digital gold” [3]. Its volatility and market capitalisation have made it a focal point in financial markets, driving the need for reliable forecasting techniques to navigate the associated risks and rewards.

Recurrent Neural Networks (RNNs) are designed to process sequential data, making them suitable for tasks involving temporal patterns. Long Short-Term Memory (LSTM) networks, a specialized form of RNNs, excel in capturing long-term dependencies and are widely used in applications like language modelling, time-series forecasting, and sentiment analysis [4,5]. Multivariate LSTM networks extend this by incorporating multiple input variables, enabling the modelling of complex relationships critical for tasks such as financial market prediction [6].

There exists a range of studies, particularly early studies, that have utilised machine learning in market prediction using either historical pricing data or social media sentiment analysis but not a combination of the two. For instance, recent research such as [7] focuses

solely on X (formerly known as Twitter) sentiment for Bitcoin price prediction. While sentiment analysis has shown potential, its limitations were noted as early as Philippas et al. [8], who relied on trends and sentiments alone, and were further corroborated by Abraham et al. [9], highlighting challenges.

On the other hand, multivariate machine learning models, such as Transformers and LSTM networks, have demonstrated significant promise in capturing the temporal relationships between sentiment and price fluctuations while addressing these limitations [10–13]. Although we acknowledge the advancements made in recent multivariate approaches, our research builds directly upon the foundational work of Zuvela et al. [14], which utilized a univariate LSTM model. By extending their methodology to a multivariate framework and incorporating social media sentiment alongside historical Bitcoin pricing data, we aim to systematically examine the impact of these enhancements. Additionally, our study employs a similar dataset to enable a direct comparison of results, providing clarity on the improvements achievable through multivariate modelling.

This research hence seeks to enhance prediction accuracy and contribute to more informed trading decisions in the volatile cryptocurrency market. Our study not only highlights the significance of integrating sentiment analysis with historical data but also positions itself as a critical evaluation of recent multivariate approaches within this domain.

The remainder of this paper is organized as follows: Section 3 describes the data collection, preprocessing, and model development. Section 4 presents the experimental results. Section 5 discusses the findings and limitations of the study. Finally, Section 6 concludes the paper with directions for future research.

2. Related Work

2.1. Definition

Understanding the foundational concepts behind Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and their multivariate extensions is crucial for contextualizing the advancements discussed in this study. This subsection provides a concise overview of these architectures and their relevance to financial time-series prediction.

A Recurrent Neural Network (RNN) is a specialized deep neural network designed to process and analyse sequential or time-series data. It is trained to generate machine learning models capable of making predictions or deriving insights based on sequential input data [15]. RNNs are particularly suited for tasks where context and temporal relationships are critical.

While traditional RNNs rely on sequential relationships from past states, Bidirectional RNNs (BRNNs) extend this capability by considering both past and future states to improve prediction accuracy. For instance, when predicting the next word in a sequence, BRNNs utilize both prior and subsequent context [15].

Long Short-Term Memory (LSTM) networks are a type of RNN architecture introduced by Hochreiter and Schmidhuber to address the challenges of long-term dependencies. LSTM networks are designed to retain important information while forgetting irrelevant details, making them particularly effective at overcoming the vanishing gradient problem that affects traditional RNNs. The vanishing gradient occurs during training when gradients diminish as they propagate backwards through time, leading to ineffective learning [16].

Multivariate LSTM models expand upon traditional LSTM networks by incorporating multiple input variables to predict a target variable. This approach allows the model to capture complex relationships between input variables, making it well-suited for tasks like stock market prediction and weather forecasting [6]. Such models are particularly valuable in financial applications, where multiple factors influence market dynamics.

2.2. Literature Review

To provide a comprehensive understanding of prior research, this subsection is divided into two parts: studies that did not incorporate sentiment analysis and those that did. This structure highlights the evolution of methodologies and the increasing emphasis on sentiment as a predictive feature in financial modelling. Table 1 summarises these key contributions.

Table 1. Summary of related work.

Study	Model	Features	Sentiment Analysis	Key Findings
Abraham et al. (2018) [9]	Linear models	Tweet volume, Google Trends, sentiment	Yes	Sentiment analysis unreliable; tweet volume a robust predictor.
Philippas et al. (2019) [8]	Multiple regression	Google Trends, Twitter data	Yes	Media attention impacts Bitcoin prices.
Mudassir et al. (2020) [17]	SVM, LSTM	Statistical indicators (e.g., MA, RSI)	No	Machine learning outperforms time-series models.
Zuvela et al. (2022) [14]	Univariate RNN LSTM	Historical price data	No	Recommended incorporating sentiment analysis.
Critien et al. (2022) [18]	Bi-LSTM, CNN	Sentiment scores, tweet volume	Yes	Voting classifiers enhance predictive accuracy.
Seabe et al. (2023) [19]	Bi-LSTM, GRU	Historical data, RNN architectures	No	Bi-LSTM outperforms other RNN variants.
Gu et al. (2024) [10]	FinBERT-LSTM	Financial news sentiment, historical prices	Yes	Integration of sentiment and price trends improves accuracy.
Saleem et al. (2024) [7]	Logistic regression	Sentiment variance	Yes	Negative sentiment impacts price declines more than positive sentiment.
Mardjo et al. (2024) [13]	HyBiLSTM	Social, economic variables	Yes	Hybrid models capture both linear and non-linear dynamics.

2.2.1. Studies Without Sentiment Analysis

Early studies such as Mudassir et al. [17] leveraged traditional machine learning models such as Support Vector Machines (SVMs) and LSTM with statistical indicators like moving averages (MA) and relative strength indices (RSI) to predict cryptocurrency prices. These models demonstrated improvements over conventional time-series methods.

Zuvela et al. [14] built on these approaches by employing univariate RNN LSTM models for Bitcoin price forecasting based solely on historical price data. While effective, their study highlighted significant overfitting issues and proposed the inclusion of additional features, such as sentiment scores, to improve generalisation.

Seabe et al. [19] expanded on these models by evaluating Bi-LSTM and GRU architectures for cryptocurrency price prediction. Their results emphasized the superior performance of Bi-LSTM due to its ability to leverage bidirectional learning for enhanced context understanding.

A comprehensive comparison of these and other models was conducted by Kehinde et al. [11], who evaluated RNN, LSTM, GRU, and Transformer models. Their findings highlighted the superior accuracy and efficient convergence of Transformer models across multiple metrics, such as mean absolute error (MAE) and root-mean-square error (RMSE). This study serves as a detailed reference for understanding the capabilities of these architectures in financial time-series prediction.

2.2.2. Studies with Sentiment Analysis

Moving beyond traditional models, Abraham et al. [9] examined cryptocurrency price prediction by integrating sentiment analysis and tweet volume into their models. While their findings indicated that sentiment analysis using VADER was unreliable due to neutral or positively biased tweets, tweet volume emerged as a robust predictor.

Philippas et al. [8] further demonstrated the importance of sentiment by analysing Google Trends and Twitter data. Their work established that Bitcoin prices were partially influenced by public sentiment, providing a foundational understanding of the interplay between sentiment and market dynamics.

Gu et al. [10] first utilised a pre-trained NLP model, FinBERT, for sentiment analysis in financial news. They incorporated this to make a hybrid FinBERT-LSTM model that combined sentiment analysis from financial news headlines with sequential data analysis of historical prices. Their model significantly improved predictive accuracy, achieving notable reductions in MAE and testing loss.

Critien et al. [18] investigated the utility of Twitter sentiment and tweet volume in predicting Bitcoin price direction and magnitude. They employed models like Bi-LSTM and CNN and introduced a voting classifier that refined predictions by aligning direction and magnitude outputs.

Saleem et al. [7] took a nuanced approach by integrating prospect theory and sentiment variance into a logistic regression framework. Their study highlighted the differential impacts of positive and negative sentiments on Bitcoin price fluctuations, emphasizing the need for refined sentiment metrics.

More recently, Mardjo et al. [13] proposed a hybrid HyBiLSTM model, combining ARIMAX, GARCHX, and Bi-LSTM for Bitcoin price prediction. Their inclusion of social and economic variables, along with a SHAP analysis for feature importance, demonstrated the effectiveness of hybrid models in capturing both linear and non-linear dynamics.

2.3. Comparison and Contribution

Our work builds on works such as Philippas et al. [8] and Gu et al. [10], which demonstrate the value of sentiment analysis in predicting Bitcoin prices, although their methodologies differ from ours in terms of dataset and architecture. To be more precise, Ref. [8] used a combination of Google Trends and Twitter data to measure media attention on Bitcoin, while Ref. [10] performed their own sentiment analysis using a pre-trained NLP model. Our study, however, utilises a dataset that includes Twitter sentiment scores to improve prediction accuracy.

Our work is also similar to [19], as we employ Bi-LSTM architectures, but we introduce multivariate inputs, such as sentiment scores, which were absent in their study. Additionally, our dataset differs from theirs when it comes to the start and end dates as well as in granularity, focusing on hourly data rather than daily records.

The primary contribution of this study is improving upon [14] by integrating sentiment analysis into their univariate LSTM approach. Our multivariate models demonstrate superior performance, as evidenced by significantly reduced MAE and RMSE values. Furthermore, our comparative analysis across multiple architectures underscores the importance of leveraging sentiment indices alongside historical pricing data.

By addressing these gaps, our research provides a robust framework for understanding the interplay between social sentiment and cryptocurrency price dynamics, setting the stage for future advancements in financial forecasting.

3. Proposed Methodology

The OSEMN framework was followed for our work, which is a five-step process of Obtaining Data, Scrubbing Data, Exploring Data, Modelling Data, and Interpreting Data [20]. This project built on the work of Zuvela et al. [14], who developed a univariate RNN LSTM model to predict Bitcoin (BTC) prices based solely on closing price data. To enhance model performance, this study incorporated social media sentiment data and tweet volume alongside multivariate LSTM models, inspired by previous research indicating that sentiment indices can improve predictive accuracy [9]. Additionally, error metrics such as mean absolute error (MAE), mean square error (MSE), mean absolute percentage error (MAPE), and root-mean-square error (RMSE) were employed for model evaluation, ensuring that results are comparable with prior findings and reliable for practical applications.

3.1. Data Collection

The datasets for this study were sourced from Data & Sons, published under an open license [21]. This dataset comprised historical Bitcoin (BTC) price data in USD alongside tweet sentiment scores, covering the period from 1 August 2017 to 21 January 2019.

The data, recorded hourly, provided 24 entries per day, each capturing the BTC price and the associated sentiment score for that hour. This granularity facilitated an in-depth analysis of the relationship between Bitcoin price trends and social sentiment, providing a robust foundation for model development.

The dataset contained 12,358 rows and 16 columns, representing an hourly snapshot of market and sentiment metrics. The primary features are listed below:

- Date: timestamp of recorded data.
- Compound_Score: aggregate sentiment score from social media posts.
- Total Volume of Tweets: number of tweets analysed.
- Count Negatives, Positives, and Neutrals: counts of tweets with respective sentiments.
- Sent Negatives and Positives: average sentiment score for negative and positive tweets.
- Count News and Bots: counts of news articles and bot-identified tweets.
- Open, High, Low, Close: hourly BTC prices.
- Volume BTC and Currency: trading volume in BTC and USD.

The correlation between these features is illustrated in Figure 1, which highlights the relationships between sentiment variables and Bitcoin price movements.

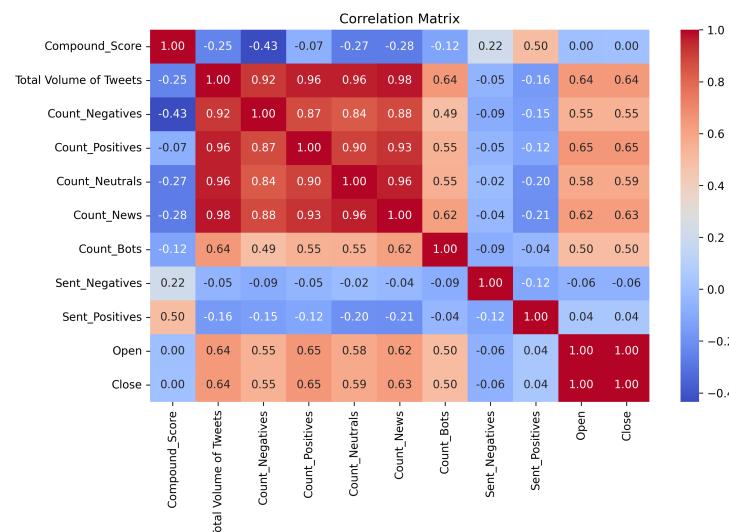


Figure 1. Correlation heatmap illustrating the relationships between features. Values range from -1 to 1 , with -1 indicating a strong negative correlation and 1 indicating a strong positive correlation.

3.2. Data Preprocessing

The multivariate LSTM model utilised features such as Close (the closing price), Sent_Positives, and Sent_Negatives for the primary analysis. Additional experiments incorporated the Total Volume of Tweets feature to assess its effect on prediction accuracy. The Close feature served as the target variable across all experiments, with sentiment data acting as auxiliary predictors.

MinMax scaling was applied to standardise values to a range between 0 and 1 to address scale disparities among features. The dataset was divided into training (60%), validation (20%), and test (20%) subsets, preserving temporal relationships to maintain the integrity of time-series data.

Later in Section 5, we explore an alternative data partitioning strategy of 70%–15%–15% to evaluate its impact on some of our models' performance. This adjustment aimed to assess the robustness of our best models, the Multi-LSTM-Sentiment and the Tuned Multi-LSTM-Sentiment models (see Sections 3.3.1 and 3.3.2), under different training conditions and data distributions.

Input sequences were structured with a length of 48 (equivalent to a two-day window of hourly data), enabling the model to learn temporal dependencies. The data were reshaped into the LSTM-required format of “samples, time steps, features”, where the training set encompassed 9838 samples, and the test set included 2424 samples.

For the last mode, the Hybrid LSTM-Volume model, the preprocessing steps were adjusted slightly to focus on two features: Close (closing price) and Total Volume of Tweets. The Close feature remained the target variable, while tweet volume data acted as an auxiliary predictor. This setup allowed the model to combine price information with social media activity for more robust predictions.

The reshaped dataset comprised 12,310 samples, each with 48 time steps and 2 features. The training set included 9838 samples, while the test set contained 2424 samples, maintaining the same data partitioning strategy and sequence structure as before.

3.3. Model Training

3.3.1. Multi-LSTM-Sentiment

A sequential LSTM architecture was developed first to predict Bitcoin price movements using multiple input features. We named the model Multi-LSTM-Sentiment and describe its training and validation in this section.

The features used included historical Bitcoin price data and Twitter sentiment scores, as these features were hypothesised to impact price dynamics significantly. Twitter sentiment scores, derived from the textual data of Bitcoin-related tweets, were aggregated to provide a sentiment index reflecting market mood. Historical prices were preprocessed to ensure temporal alignment with the sentiment scores, and all features were normalised using MinMax scaling for optimal model performance.

The architecture consisted of three LSTM layers, each containing 50 neurons. Outputs from each LSTM layer were sequentially fed into the next, ensuring that the model captured long-term temporal dependencies across the multivariate input data. The input sequence length was set to 48 time steps, corresponding to a two-day window (assuming hourly intervals), which provided sufficient historical context for the model to learn meaningful patterns. The final dense layer, containing a single neuron, provided the price prediction.

The choice of 50 neurons per LSTM layer balanced the need to model complexity with the risk of overfitting. Fewer neurons might have limited the model's ability to capture intricate patterns, while more neurons could have increased computational demands and the risk of overfitting. The input sequence length of 48 captured two days' worth of data, aligning with the hypothesis that recent historical trends were more predictive. The Adam optimiser was selected for its ability to handle sparse gradients efficiently, crucial for training deep architectures like LSTM. Using three layers allowed the model to progressively extract complex patterns while maintaining computational efficiency.

The Adam optimiser and MSE were employed to guide the training process. The training spanned 50 epochs with a batch size of 25. A validation split of 20% ensured a robust evaluation of the model's generalisation ability. Figure 2 depicts the training and validation error trends for the Multi-LSTM-Sentiment model across MSE, MAE, and MAPE metrics, showing that the model effectively learned the training data.

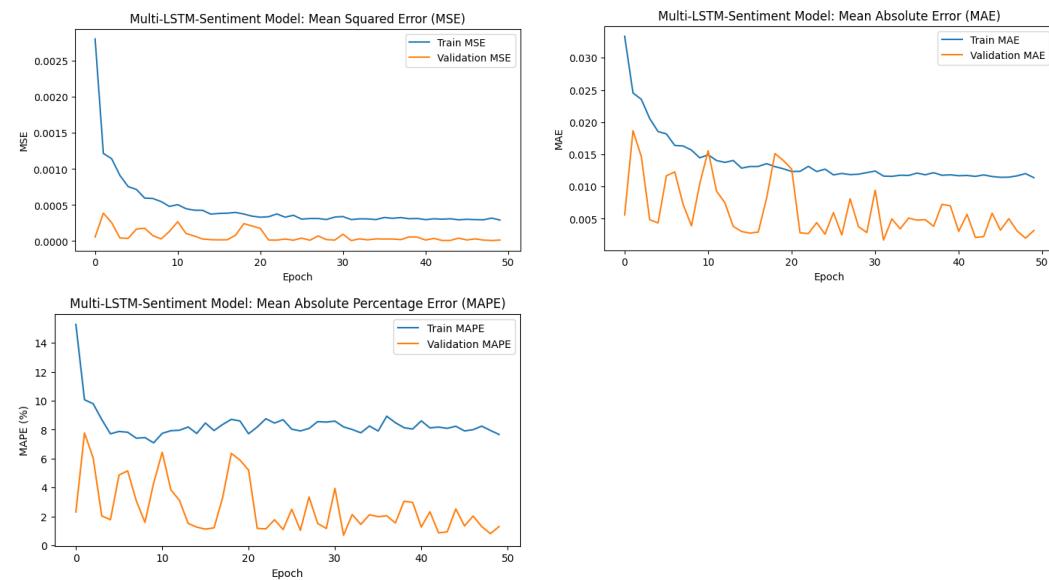


Figure 2. Training and validation error trends for the Multi-LSTM-Sentiment model across metrics: MSE, MAE, and MAPE. The decreasing trends indicate effective learning and generalisation.

The metrics dropped significantly over the epochs, showing that the model effectively learned the training data. By the 50th epoch, the MAE was very low, indicating minimal prediction errors on the training set. The low MAPE showed the predictions were accurate across different scales of the target variable.

Similarly, the validation metrics also decreased over time. By the 50th epoch, both MAE and MAPE were very low, reflecting the model's effective prediction accuracy on unseen data.

Overall, the model performed well on both training and validation datasets. The low MAE and MAPE values, along with consistent metrics across training and validation, indicated that the model was not overfitting and had generalised effectively.

3.3.2. Tuned Multi-LSTM-Sentiment

Hyperparameters such as the number of neurons, batch size, epochs, learning rate, and dropout rate were optimised for the Multi-LSTM-Sentiment model to improve its performance. Initially, the model exhibited slight overfitting, prompting adjustments such as reducing the number of neurons to 30 per LSTM layer and increasing the dropout rate to 30%. These changes helped improve generalisation while mitigating overfitting.

Reducing the neuron count mitigates the risk of overfitting by simplifying the model's architecture, while a dropout rate of 30% ensured regularisation by randomly disabling neurons during training. This dropout rate was chosen based on prior experiments showing its effectiveness in reducing overfitting without overly compromising model learning. Retaining three LSTM layers ensured that sufficient depth existed to capture intricate temporal dependencies across the multivariate dataset.

The tuned Multi-LSTM-Sentiment model was then trained using the same input features (price and sentiment) over 40 epochs, with a batch size of 30. The 20% validation split was retained to ensure consistency in model evaluation.

We visualised the training and validation trends for the MSE, MAE, and MAPE in Figure 3. The validation metrics confirmed that the tuning process improved the model's prediction accuracy while maintaining low error rates across all metrics.

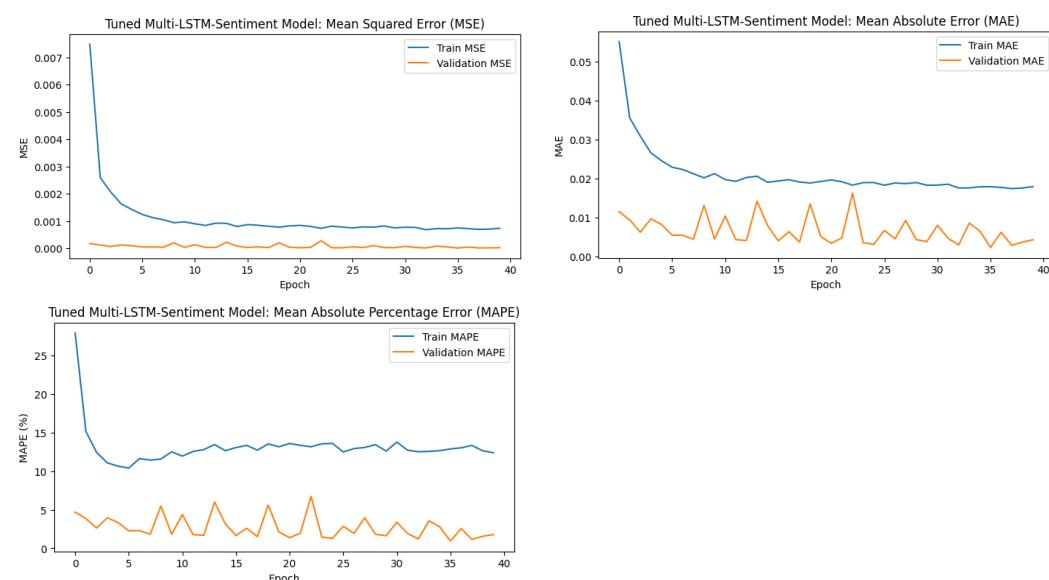


Figure 3. Training and validation error trends for the tuned Multi-LSTM-Sentiment model across metrics: MSE, MAE, and MAPE. Tuning resulted in a consistent reduction in errors and improved generalisation.

3.3.3. Bidirectional LSTM-Sentiment

To further explore temporal dependencies, a bidirectional LSTM model, referred to as Bidirectional LSTM-Sentiment, was developed using the same input features: Bitcoin price and sentiment scores. The bidirectional LSTM architecture incorporated three bidirectional LSTM layers, each containing 50 neurons.

The bidirectional setup leverages both past and future context, which is critical for datasets where patterns may depend on both preceding and succeeding data points. This architecture enhances the model's ability to learn temporal dependencies, making it particularly suitable for financial time-series data. Employing three layers provided adequate depth for capturing the bidirectional temporal relationships without introducing excessive computational complexity.

This design allowed the model to capture both forward and backward temporal relationships in the data, providing a more comprehensive understanding of price trends influenced by historical patterns, volume spikes, and sentiment fluctuations.

The training process mirrored that of the sequential LSTM (described in Section 3.3.1), with the Adam optimiser, MSE loss function, 50 epochs, and a batch size of 25. Similar to the earlier models, a 20% validation split was employed. Figure 4 visualises the training and validation error trends for MSE, MAE, and MAPE, demonstrating the model's ability to generalise effectively.

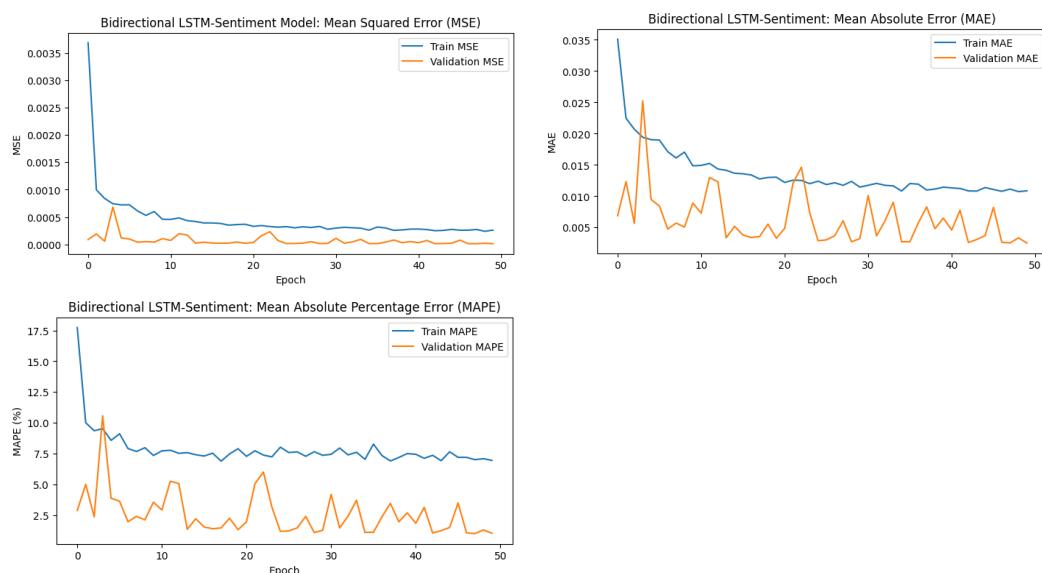


Figure 4. Training and validation error trends for the Bidirectional LSTM-Sentiment model across metrics: MSE, MAE, and MAPE. The decreasing trends indicate effective learning and generalisation.

3.3.4. Hybrid LSTM-Volume

Our final model, which we referred to as Hybrid LSTM-Volume, integrated the benefits of the Multi-LSTM-Sentiment architecture (from Section 3.3.1) while optimising performance through refined training and hyperparameter adjustments. This model's training differed as it used a multivariate input feature set, including historical Bitcoin price as well as the total volume of tweets.

The architecture consisted of three LSTM layers, each containing 50 neurons. A dropout rate of 30% was applied after each LSTM layer to improve generalisation. The input sequence length remained at 48 time steps, maintaining a two-day temporal context for learning meaningful relationships among features. A dense output layer with a single neuron was employed to produce the final Bitcoin price prediction.

By combining tweet volume with historical price, the model exploited a broader set of features, potentially improving prediction accuracy. The choice of neuron count and increased dropout rate reflected an emphasis on preventing overfitting while maintaining sufficient model capacity for capturing complex relationships. Retaining three layers allowed us to balance model depth and computational efficiency.

As before, the Adam optimiser was used for gradient-based learning, and the MSE was used to measure the model's performance. The model was trained for 50 epochs, with

a batch size of 25. The 20% validation split was retained, allowing the evaluation of the model's ability to generalise to unseen data throughout training. Figure 5 presents the training and validation error trends for the Hybrid LSTM-Volume model across MSE, MAE, and MAPE metrics.

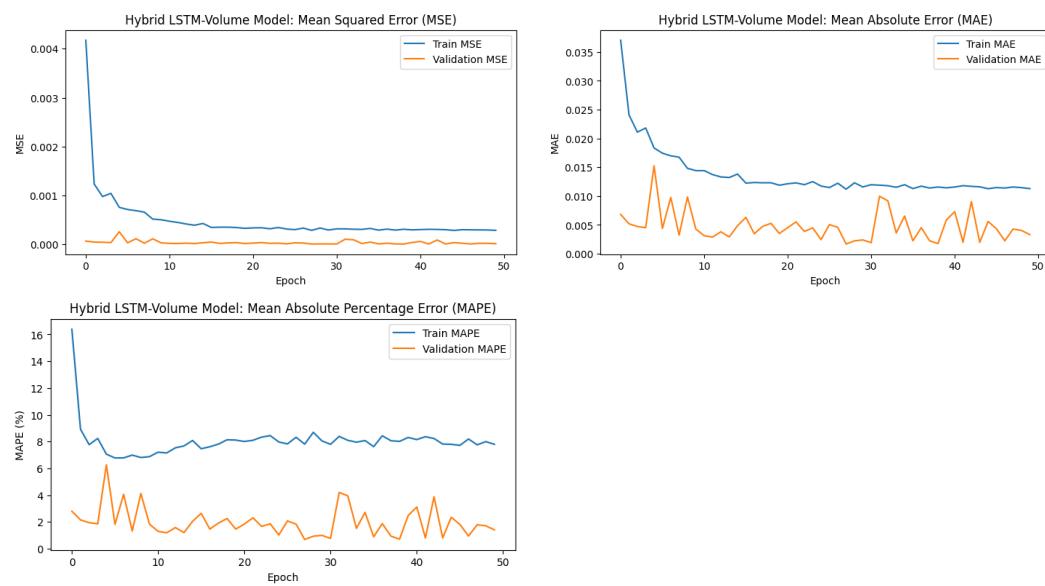


Figure 5. Training and validation error trends for the Hybrid LSTM-Volume model across metrics: MSE, MAE, and MAPE. The decreasing trends indicate effective learning and generalisation.

3.3.5. Tuned Hybrid LSTM-Volume

Hyperparameters, including the number of neurons, batch size, number of epochs, learning rate, and dropout rate, were fine-tuned for the Hybrid LSTM-Volume from Section 3.3.4. These adjustments aimed to reduce model complexity, shorten training time, and enhance overall performance and accuracy.

Fine-tuning hyperparameters ensured the model struck a balance between training efficiency and predictive accuracy. By iteratively adjusting parameters, the model's capacity to generalise was enhanced, mitigating issues like overfitting or underfitting. Using three LSTM layers ensured the model maintained sufficient depth to capture complex temporal dependencies.

After fine-tuning the Hybrid LSTM-Volume model, the training loss (MSE) decreased consistently from an initial value of 0.0045 in the first epoch to 0.0003 by the final epoch, indicating the model effectively minimised errors. Validation loss (MSE) closely tracked training loss throughout the process, starting at a slightly lower value and stabilising by the final epoch. The absence of significant fluctuations in the validation loss indicated strong generalisation without overfitting.

The final validation loss achieved was 0.0001, reflecting minimal error between predicted and actual price values on the validation set. These results confirmed that the Hybrid LSTM-Volume Model successfully leveraged multivariate input features to capture complex temporal relationships influencing Bitcoin price trends. Figure 6 displays the training and validation error trends for MSE, MAE, and MAPE.

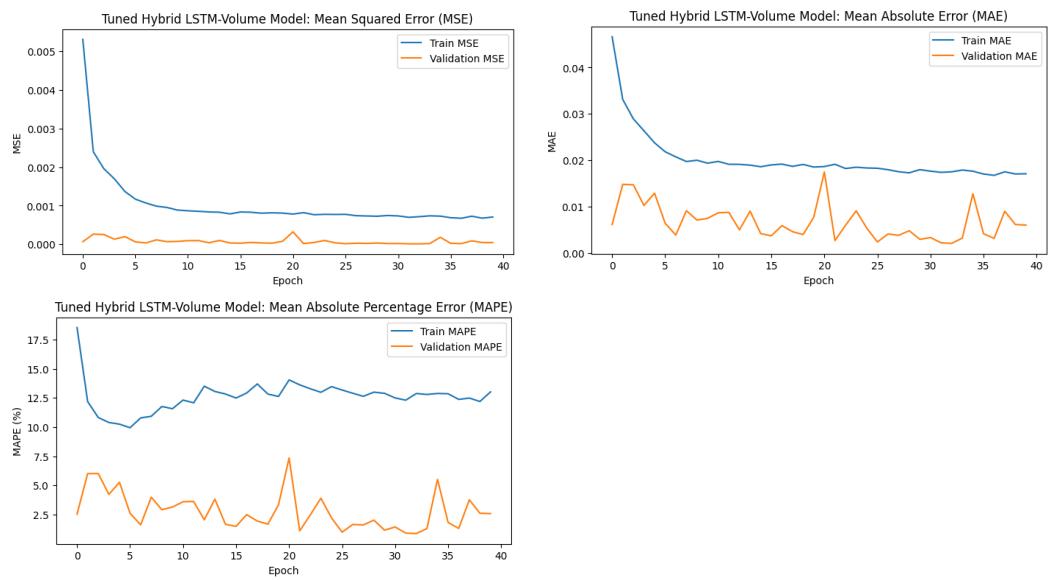


Figure 6. Training and validation error trends for the tuned Hybrid LSTM-Volume model across metrics: MSE, MAE, and MAPE. The decreasing trends indicate effective learning and generalisation.

4. Results

4.1. Performance of the Multi-LSTM-Sentiment Models

The performance of the Multi-LSTM-Sentiment model (from Section 3.3.1) was evaluated using error metrics, including mean absolute error (MAE) and root-mean-square error (RMSE), to measure its predictive accuracy on unseen test data. The model achieved an MAE of 0.0093 and an RMSE of 0.0099. These low values indicate that the model generally performed well, with minimal deviations from the actual Bitcoin prices during testing.

A graph comparing the actual Bitcoin prices and the predicted prices over the test period is shown in Figure 7. The Multi-LSTM-Sentiment model successfully captured the overall trend of Bitcoin price fluctuations. However, the model struggled with sudden price spikes or drops, highlighting areas for improvement in handling high volatility.

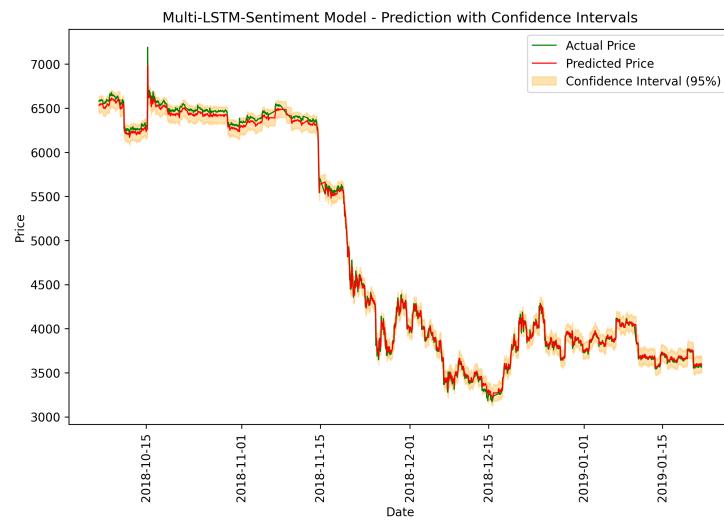


Figure 7. Multi-LSTM-Sentiment: actual vs. predicted prices.

The differences between the actual and predicted values, called residuals, were used to measure the model's errors. The standard deviation of these errors helped estimate how much the predictions might vary. Confidence intervals were calculated by multiplying this standard deviation by a specific value (for example, 1.96 for 95% confidence). These

intervals are shown as a shaded area around the predicted values on the plot, giving a visual sense of how uncertain the model's predictions were. All figures in this section now have a shaded area demonstrating the confidence score, enhancing the interpretability of the visualized trends and the reliability of the predictions.

The tuned Multi-LSTM-Sentiment model, which, as described in Section 3.3.2, was derived from fine-tuning the hyperparameters of the Multi-LSTM-Sentiment model, exhibited superior performance. The MAE and RMSE values decreased significantly to 0.00196 and 0.00304, respectively. This highlights the effectiveness of hyperparameter tuning in improving model accuracy.

Figure 8 compares the actual and predicted prices for the tuned Multi-LSTM-Sentiment model. The predictions aligned closely with the actual prices, demonstrating the model's ability to generalise well. Minor deviations were observed, particularly during periods of rapid price changes, but overall, the tuned model performed exceptionally well.

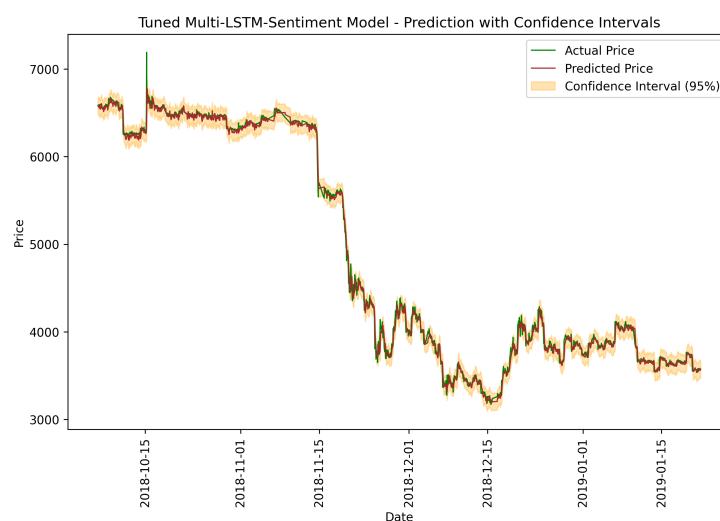


Figure 8. Tuned Multi-LSTM-Sentiment: actual vs. predicted prices.

A comparative analysis of the performance of the Multi-LSTM-Sentiment and the tuned Multi-LSTM-Sentiment models on test data is presented in Table 2, along with results from the literature. The tuned Multi-LSTM-Sentiment model's improvements in MAE and RMSE confirmed it as the most accurate and reliable.

Table 2. Performance metrics comparison for Multi-LSTM models.

Model	MSE	RMSE	MAE
Multi-LSTM-Sentiment	9.80247	0.00990	0.00934
Tuned Multi-LSTM-Sentiment	9.29112	0.00304	0.00196
Univariate RNN LSTM [14]	340,685,452.13732	18,446.6821	13,094.2243

4.2. Performance of the Bidirectional LSTM-Sentiment Model

The Bidirectional LSTM-Sentiment model, implemented with a bidirectional LSTM architecture as discussed in Section 3.3.3, achieved competitive performance on the test dataset. The Bidirectional LSTM produced an MAE of 0.00253 and an RMSE of 0.00350. Although slightly higher than the tuned Multi-LSTM-Sentiment, these values indicate robust predictive performance.

A visual comparison of the actual and predicted prices is presented in Figure 9. The Bidirectional LSTM model captured the general trend of Bitcoin price fluctuations

effectively. However, it occasionally struggled to accurately track sudden price spikes or drops, mirroring challenges faced by the other models.

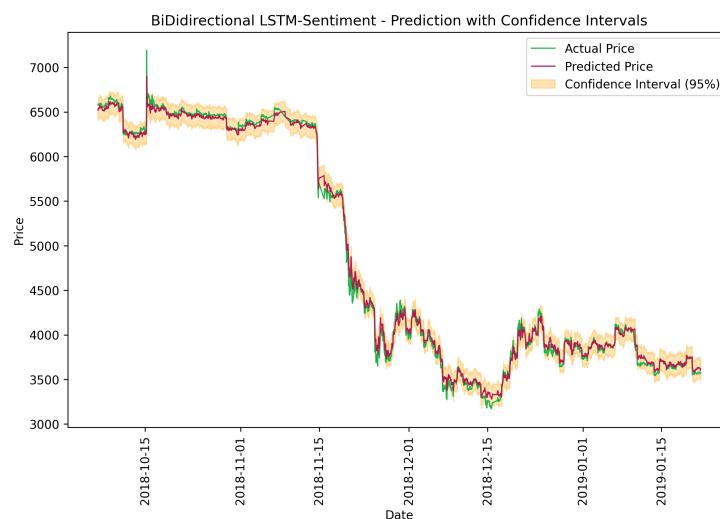


Figure 9. Bidirectional LSTM-Sentiment: actual vs. predicted prices.

A performance comparison between the tuned Multi-LSTM-Sentiment model and the Bidirectional LSTM-Sentiment model is shown in Table 3. While the Bidirectional LSTM-Sentiment demonstrated excellent results, the tuned Multi-LSTM-Sentiment outperformed it slightly, reinforcing its suitability for this dataset.

Table 3. Performance comparison for Bidirectional LSTM.

Model	MSE	RMSE	MAE
Tuned Multi-LSTM-Sentiment	9.29112	0.00304	0.00196
Bidirectional LSTM-Sentiment	9.20827	0.00350	0.00253

4.3. Performance of the Hybrid LSTM-Volume Models

The Hybrid LSTM-Volume model achieved an MAE of 0.00499 and an RMSE of 0.00544, indicating robust performance on the test dataset. Figure 10 presents a graphical comparison of actual versus predicted Bitcoin prices. The low values of these metrics demonstrate that the model effectively learned the relationship between the input features (price and tweet volume) and Bitcoin price movements. The predictions aligned closely with the actual prices, capturing the general trend and fluctuations. However, slight deviations were observed during periods of rapid price movement, suggesting room for further refinements.

The tuned Hybrid LSTM-Volume model outperformed its base counterpart, achieving an MAE of 0.00420 and an RMSE of 0.00545. Figure 11 illustrates the model's predictions compared to the actual prices, demonstrating significant accuracy improvements. The reduced MAE and RMSE values highlight the improvements gained through hyperparameter tuning.

A summary of the performance metrics for both the Hybrid LSTM-Volume model and its tuned version is provided in Table 4. The improvements in MAE and RMSE confirm the effectiveness of the hyperparameter tuning process.

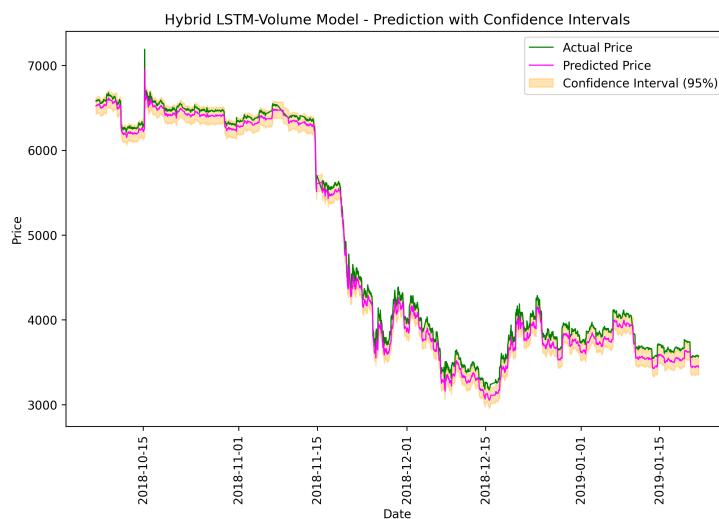


Figure 10. Hybrid LSTM-Volume model: actual vs. predicted prices.

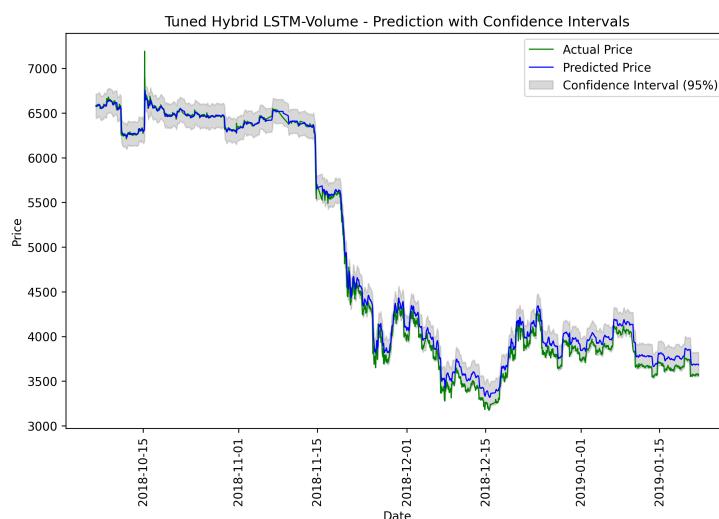


Figure 11. Tuned Hybrid LSTM-Volume model: actual vs. predicted prices.

Table 4. Performance comparison for Hybrid LSTM-Volume models.

Model	MSE	RMSE	MAE
Hybrid LSTM-Volume	2.96226	0.00544	0.00499
Tuned Hybrid LSTM-Volume	2.97542	0.00545	0.00420

5. Discussion

5.1. Robustness Evaluation Through Data Splits

This study systematically evaluated and extended the univariate approach of Zuvela et al. [14] by developing five LSTM-based models to predict Bitcoin prices, using sentiment and tweet volume data as primary features. Among these, the tuned Multi-LSTM-Sentiment model (see Section 3.3.2) outperformed the others, achieving the lowest MAE of 0.00196 and RMSE of 0.00304, as presented in Table 2. These metrics indicate its robustness and reliability in capturing price fluctuations effectively, making it the most suitable model for the dataset.

The Multi-LSTM-Sentiment model was further evaluated using an alternative data partitioning strategy of 70%–15%–15%, as discussed in Section 3.2. The results are depicted in Figures 12 and 13. The model demonstrated strong robustness, as reflected in its low and

stable validation MSE, rapid convergence, and consistent performance between training and validation datasets. While minor fluctuations in validation MAE and MAPE occurred, the overall trends indicated effective learning, strong generalisation, and high predictive accuracy. These characteristics suggest the model is well suited for handling unseen data, though further tuning could improve its stability and performance.

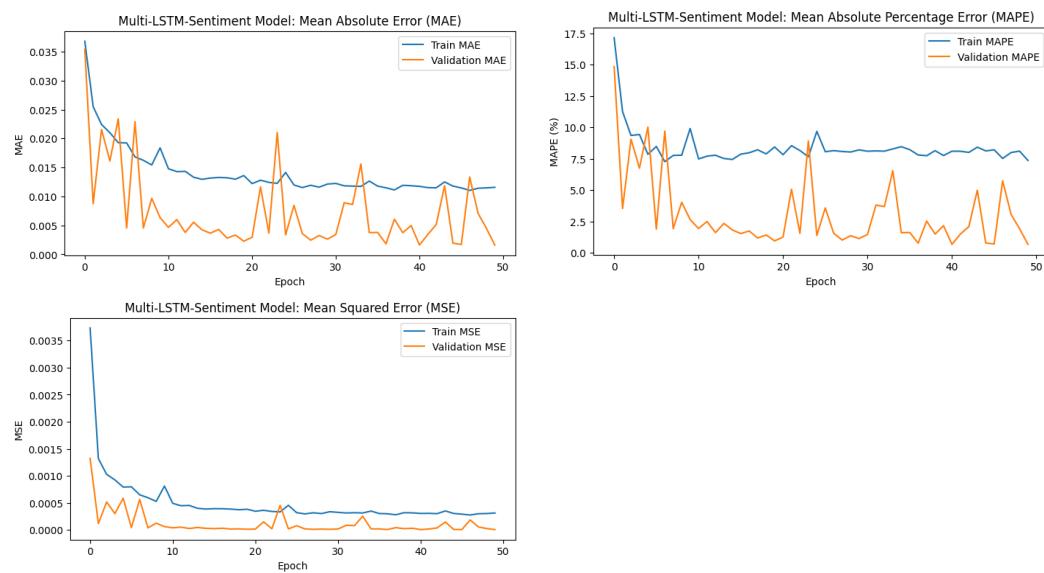


Figure 12. Training and validation error trends for Multi-LSTM-Sentiment under 70%–15%–15% split across metrics: MAE, MAPE, and MSE.



Figure 13. Multi-LSTM-Sentiment: Actual vs. Predicted Prices under 70%–15%–15% split.

The tuned Multi-LSTM-Sentiment model was also evaluated under the 70%–15%–15% split. The results, shown in Figures 14 and 15, indicate a significant improvement in prediction accuracy. The model exhibited rapid convergence and low error metrics across all evaluation criteria, confirming the effectiveness of hyperparameter tuning under different training conditions.

These findings underscore the flexibility and robustness of the Multi-LSTM-Sentiment and its tuned counterpart across different data splits. However, the study identified limitations, as discussed below, which highlight areas for future research.

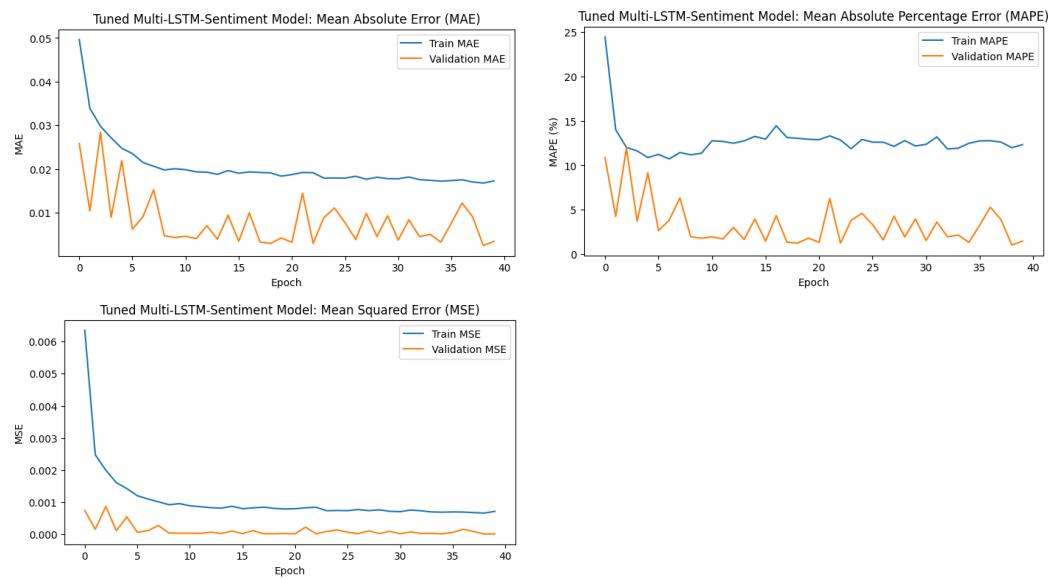


Figure 14. Training and validation error trends for Tuned Multi-LSTM-Sentiment under 70%–15%–15% split across metrics: MAE, MAPE, and MSE.

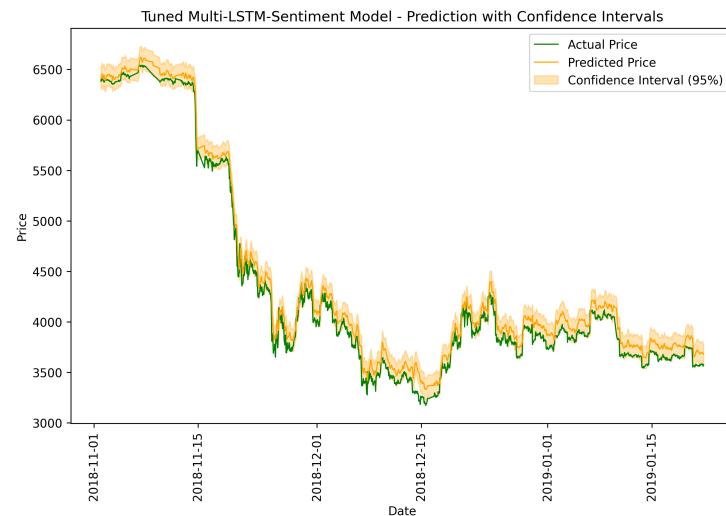


Figure 15. Tuned Multi-LSTM-Sentiment: actual vs. predicted prices under 70%–15%–15% split.

5.2. Ablation Study

To further test the robustness of the Multi-LSTM-Sentiment model, an ablation experiment was conducted. The feature sequence length was adjusted from 48 to 72 data points, representing a 3-day period, while the number of neurons in the LSTM layers was reduced to 32. These modifications allowed for an evaluation of the model's sensitivity to architectural and feature sequence changes.

The results indicate that while the model retained generalisation ability, as evidenced by low validation MSE, MAE, and MAPE, a slight increase in these metrics during the final epoch suggested challenges with unseen data. This was reflected in the fluctuations observed in MAE and MAPE during training. The training and validation error trends are visualised in Figures 16 and 17.

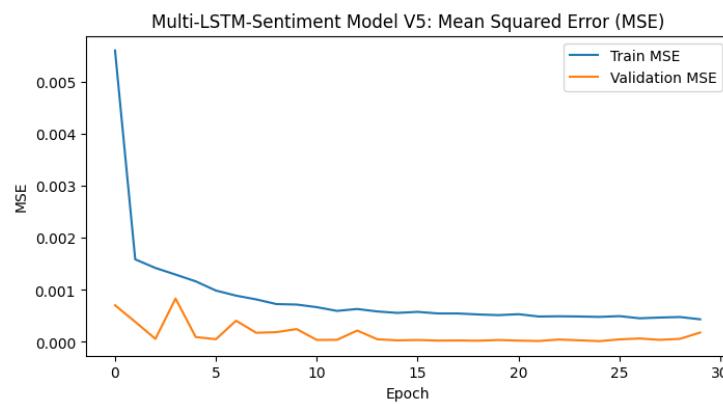


Figure 16. Training error trends for feature sequence and model ablation.

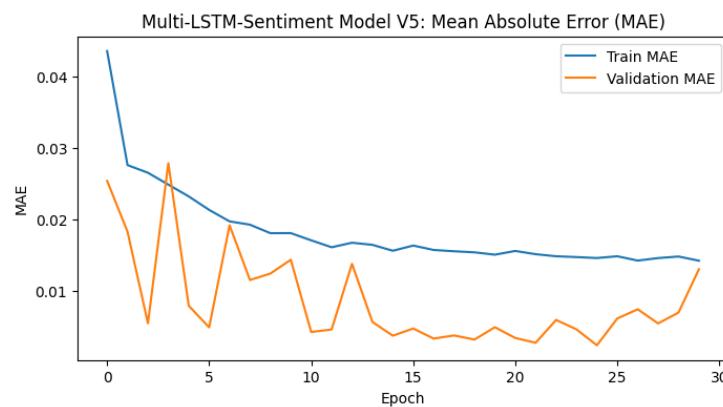


Figure 17. Validation error trends for feature sequence and model ablation.

The test performance of the model is summarised in Table 5. The ablation adjustments negatively impacted the model's ability to capture price trends, with significant deviations from actual prices in the testing phase, as shown in Figure 18.

Table 5. Performance metrics for feature sequence and model ablation.

Model	MSE	RMSE	MAE	MAPE
Multi-LSTM V5	0.00015	0.01235	0.01198	13.33914

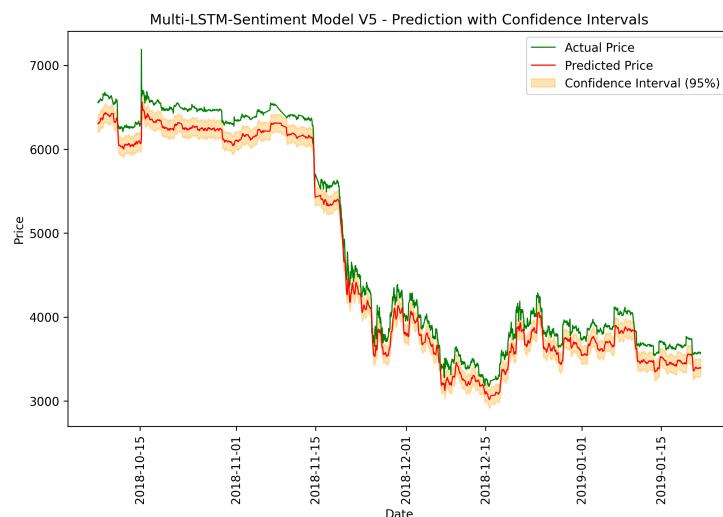


Figure 18. Multi-LSTM V5: actual vs. predicted prices under ablation experiment.

Furthermore, a univariate LSTM model trained using only the “Close” price feature was compared against the Tuned Multi-LSTM-Sentiment model (see Section 3.3.2). The results, shown in Table 6, demonstrate that the multivariate approach significantly outperformed the univariate model in all evaluation metrics, underscoring the importance of incorporating sentiment data. The actual vs. predicted prices for the univariate model are visualised in Figure 19.

Table 6. Test performance metrics for univariate and multivariate models.

Model	MSE	RMSE	MAE	MAPE
Univariate LSTM	0.00011	0.01072	0.01013	13.34481
Tuned Multi-LSTM-Sentiment	9.29013	0.00304	0.00196	2.20874

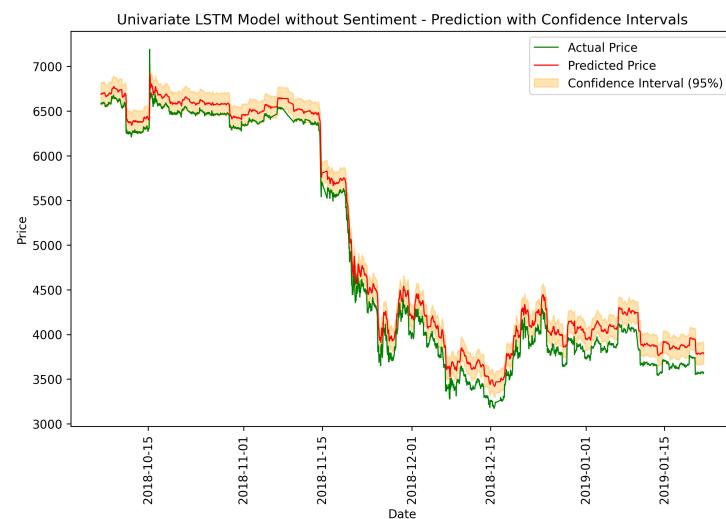


Figure 19. Univariate LSTM: actual vs. predicted prices.

These findings highlight the robustness of the Multi-LSTM-Sentiment model and the critical role of its architecture and feature set. Nevertheless, the model’s sensitivity to architectural changes underscores the need for careful optimization when adapting it to different datasets or conditions.

5.3. Limitations

Several limitations constrain the generalisability and practical application of this study’s findings. The dataset spanned from 1 August 2017 to 21 January 2019, limiting the models’ applicability to more recent market conditions or other cryptocurrencies. Additionally, the sentiment scores and tweet volume data were preprocessed and provided as-is, raising concerns about their reliability and the lack of transparency in the sentiment analysis methodology.

The robustness of the models was evaluated during the training, validation, and testing phases. Some models performed well on training and validation, but less convincingly on test data (e.g., the Tuned Hybrid LSTM-Volume model). The study focused on a limited number of features, such as sentiment scores and tweet volume, without incorporating additional indicators such as external economic data. These factors could provide a more comprehensive view of market dynamics. Computational constraints also posed challenges, particularly in training the Bidirectional LSTM model, which required significant resources for processing and optimisation.

Furthermore, the models were trained on historical data, with no attempt made to predict future prices due to dataset limitations. This restricts the study’s applicability to

forward-looking financial strategies. A comparison of performance metrics for all models is presented in Table 7.

Table 7. Performance metrics comparison for all models.

Model	MSE	RMSE	MAE
Multi-LSTM-Sentiment	9.80247	0.00990	0.00934
Tuned Multi-LSTM-Sentiment	9.29112	0.00304	0.00196
Bidirectional LSTM-Sentiment	9.20827	0.00350	0.00253
Hybrid LSTM-Volume	2.96226	0.00544	0.00499
Tuned Hybrid LSTM-Volume	2.97542	0.00545	0.00420

6. Conclusions and Future Work

This study systematically extended the univariate model of Zuvela et al. [14] by developing multivariate LSTM architectures to evaluate the impact of integrating sentiment data on Bitcoin price prediction. We developed five LSTM-based models, including the Multi-LSTM-Sentiment (Section 3.3.1) and Bidirectional LSTM-Sentiment (Section 3.3.3) models, and evaluated their efficiency in predicting Bitcoin prices. By leveraging historical price and social media sentiment datasets, the tuned Multi-LSTM-Sentiment model (Section 3.3.2) emerged as the most reliable model, achieving the lowest error metrics (MAE of 0.00196 and RMSE of 0.00304).

As future research, we recommend expanding the dataset to include more recent data or leveraging APIs like the X API to experiment with live data and assess the robustness of predictive models in real-time trading environments. It would also be interesting to integrate additional external features, such as economic indicators and news sentiment scores, within the dataset to enhance the predictive capabilities of hybrid models. Future work will also focus on implementing and benchmarking state-of-the-art (SOTA) models, such as the GRU and LSTM models from [19] or Transformer-based architectures mentioned in [11], to provide a more comprehensive evaluation of predictive performance.

As future work, we aim to explore advanced regularization techniques, adaptive learning rate strategies, or multitask learning paradigms to enhance model stability under varied data splits and architectural changes. We would additionally like to compare the performance of our models with state-of-the-art models like the Transformer-based architectures mentioned in [11] to provide a more comprehensive evaluation of predictive performance. The high computational resources required for real-time sentiment analysis and training deep learning models pose a barrier, which future works can mitigate by exploring distributed computing or cloud-based solutions.

Despite the above-mentioned limitations, we believe that the findings reaffirm the potential of LSTM-based architectures for cryptocurrency price prediction. This research contributes valuable insights into the role of sentiment-driven models in addressing the inherent volatility of the cryptocurrency market.

Author Contributions: Conceptualization, H.S.H., M.G. and S.S.; methodology, H.S.H.; software, H.S.H.; validation, H.S.H., M.G. and S.S.; formal analysis, H.S.H.; investigation, H.S.H.; resources, M.G.; data curation, H.S.H.; writing—original draft preparation, H.S.H.; writing—journal paper preparation, M.G.; writing—review and editing, S.S.; visualization, H.S.H.; supervision, M.G.; project administration, S.S.; guidance and submission planning, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available under an open license and can be accessed at <https://www.dataandsons.com/categories/markets/bitcoin-tweets-dataset-2017-to-2019> 7 August 2024. Additional information is available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BTC	Bitcoin
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MSE	Mean square error
RMSE	Root-mean-square error

References

1. Harvey, C.R. Bitcoin Myths and Facts. *SSRN Electron. J.* **2014**. [CrossRef]
2. Edwards, J.; Mansa, J.; Kvålhaug, S. Bitcoin’s Price History. *Investopedia*, 2023. Available online: <https://www.investopedia.com/articles/forex/121815/bitcoins-price-history.asp> (accessed on 1 December 2024).
3. Popper, N. *Digital Gold: The Untold Story of Bitcoin*; Penguin Books UK: City of Westminster, UK, 2015.
4. Hochreiter, S. Long Short-Term Memory. In *Neural Computation*; MIT Press: Cambridge, MA, USA, 1997.
5. Olusegun, R.; Oladunni, T.; Audu, H.; Houkpati, Y.; Bengesi, S. Text mining and emotion classification on monkeypox Twitter dataset: A deep learning-natural language processing (NLP) approach. *IEEE Access* **2023**, *11*, 49882–49894. [CrossRef]
6. Brownlee, J. Multivariate Time Series Forecasting with LSTMs in Keras. Available online: <https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/> (accessed on 1 December 2024).
7. Saleem, T.; Yaqub, U.; Zaman, S. Twitter Sentiment Analysis and Bitcoin Price Forecasting: Implications for Financial Risk Management. *J. Risk Financ.* **2024**, *25*, 407–421. [CrossRef]
8. Philippas, D.; Rjiba, H.; Guesmi, K.; Goutte, S. Media Attention and Bitcoin Prices. *Financ. Res. Lett.* **2019**, *30*, 37–43. [CrossRef]
9. Abraham, J.; Higdon, D.; Nelson, J.; Ibarra, J. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Sci. Rev.* **2018**, *1*, 1.
10. Gu, W.J.; Zhong, Y.H.; Li, S.Z.; Wei, C.S.; Dong, L.T.; Wang, Z.Y.; Yan, C. Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis. In Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing, Oxford, UK, 15–17 August 2024; pp. 67–72. [CrossRef]
11. Kehinde, T.O.; Khan, W.A.; Chung, S.H. Financial Market Forecasting Using RNN, LSTM, BiLSTM, GRU and Transformer-Based Deep Learning Algorithms. In Proceedings of the IEOM International Conference on Smart Mobility and Vehicle Electrification, Detroit, MI, USA, 10–12 October 2023.
12. Sangwan, V.; Kumar, V.; Christopher, V.B. Contrasting the Efficiency of Stock Price Prediction Models Using Various Types of LSTM Models Aided with Sentiment Analysis. In *AIP Conference Proceedings*; AIP Publishing: Long Island, NY, USA, 2024; Volume 3075.
13. Mardjo, A.; Choksuchat, C. HyBiLSTM: Multivariate Bitcoin Price Forecasting Using Hybrid Time-Series Models With Bidirectional LSTM. *IEEE Access* **2024**, *12*, 50792–50808. [CrossRef]
14. Zuvela, T.; Lazarevic, S.; Djordjevic, S.; Arsenovic, M.; Sladojevic, S. Cryptocurrency Price Prediction Using Deep Learning. In Proceedings of the 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 25–28 May 2022; pp. 89–94. [CrossRef]
15. Pattanayak, S. *Pro Deep Learning with TensorFlow 2.0: A Mathematical Approach to Advanced Artificial Intelligence in Python*, 2nd ed.; Apress L.P.: New York, NY, USA, 2023.
16. IBM. What Is a Recurrent Neural Network (RNN)? Available online: <https://www.ibm.com/think/topics/recurrent-neural-networks> (accessed on 1 December 2024).
17. Mudassir, M.; Bennbaia, S.; Unal, D.; Hammoudeh, M. Time-Series Forecasting of Bitcoin Prices Using High-Dimensional Features: A Machine Learning Approach. In *Neural Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2020. [CrossRef]
18. Critien, J.V.; Gatt, A.; Ellul, J. Bitcoin Price Change and Trend Prediction Through Twitter Sentiment and Data Volume. *Financ. Innov.* **2022**, *8*, 45. [CrossRef]

19. Seabe, P.L.; Moutsinga, C.R.B.; Pindza, E. Forecasting Cryptocurrency Prices Using LSTM, GRU, and Bi-Directional LSTM: A Deep Learning Approach. *Fractal Fract.* **2023**, *7*, 203. [[CrossRef](#)]
20. Janssens, J. *Data Science at the Command Line*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2021.
21. Explorei; Data and Sons Bitcoin Tweets and Price Dataset. Available online: <https://www.dataandsons.com/categories/markets/bitcoin-tweets-dataset-2017-to-2019> (accessed on 1 December 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.