Cryptocurrency Price Prediction Using Machine Learning Model with Sentiment

[1]Leenapackyasri S, [2]Priya Dharshini A, [3]Mrs.P.Sabeena Burvin, [4]Dr.J.Hemalatha, [5]Dr.Senthil Pandian
[1,2] UG Student, Department of Computer Science and Engineering, AAA College of Engg & Tech, Amathur, Sivakasi.
[3]Assistant Professor, Department of Computer Science and Engineering, AAA College of Engg & Tech, Amathur, Sivakasi.
[4]Professor &Head, Department of Computer Science and Engineering, AAA College of Engg & Tech, Amathur, Sivakasi.
[5]Associate Professor, Department of Computer Science and Engineering, AAA College of Engg & Tech, Amathur, Sivakasi.

Abstract: **Cryptocurrency price prediction remains a challenging task due to extreme market volatility and the influence of public sentiment. In this study, we present a hybrid machine learning framework that can integrate real-time and historical OHLC price data with sentiment analysis from Reddit and Wikipedia. Inspired by prior research on weighted sentiment influence, our model incorporates sentiment scores derived using the DistilBERT NLP pipeline and technical indicators such as moving averages and price change ratios. Unlike traditional LSTM-based systems, our final implementation utilizes XGBoost for its effectiveness in handling structured tabular data. The datasets are continuously collected and stored in MongoDB, enabling scalable real- time updates. Feature engineering techniques include rolling averages of sentiment, sentiment-based features, and lag- based technical signals. Experimental evaluation using RMSE and MAE metrics demonstrates the efficacy of our model in forecasting the future prices of Bitcoin, showing improved accuracy over standalone approaches. The results highlight the importance of integrating sentiment analysis, real-time market signals, and machine learning for robust financial forecasting in the cryptocurrency domain.**

Index Terms: **Cryptocurrency, Feature Engineering, Hybrid Model, Price Prediction, Real-time Data, Sentiment Analysis, XGBoost.**

I. INTRODUCTION

Cryptocurrencies such as Bitcoin and Ethereum have fundamentally reshaped digital finance, enabling fast, decentralized transactions and offering new opportunities for investment and innovation. Operating on blockchain networks, these digital assets facilitate borderless transfers, reduce transaction fees, and ensure transparency and security. Since Bitcoin's inception in 2009, the market has expanded to include thousands of cryptocurrencies, including Ethereum, Solana, and Binance Coin, each providing unique utilities beyond just digital payments. Despite this rapid growth, the cryptocurrency market remains highly volatile, with price fluctuations driven by a wide range of factors, including regulatory changes, global economic events, and, notably, public sentiment. The influence of social media, online communities, and news platforms on investor perception introduces a significant non-technical variable that traditional time-series forecasting models fail to capture [12],[13].

Consequently, reliable cryptocurrency price prediction remains a major challenge, requiring models that can adapt to both historical patterns and the dynamic market mood. In this study, we propose a machine learning framework for cryptocurrency price prediction that integrates real-time and historical price data with sentiment analysis using natural language processing (NLP) techniques. Data is continuously collected and stored in a MongoDB-based system, including OHLC price data from sources such as Yahoo Finance and CoinGecko, as well as sentiment scores 1 Data Collection extracted from Reddit and Wikipedia using a DistilBERT-based pipeline [4],[5],[14].

This approach enables real-time monitoring and ensures that the model adapts to the latest market signals. Several models were evaluated, including Long Short-Term Memory (LSTM) networks and hybrid combinations of LSTM with XGBoost.

However, through comprehensive experimentation, the XGBoost model was selected as the final architecture due to its superior performance on structured tabular data, lower training complexity, and more consistent results across various evaluation metrics. Feature engineering played a vital role in enhancing the model's predictive capabilities by incorporating technical indicators and sentiment rolling averages. This research demonstrates the advantages of combining real-time financial data, public sentiment, and machine learning for building a practical and accurate cryptocurrency forecasting model. By focusing on the interpretability and efficiency of XGBoost, our framework aims to support investors and analysts in making informed decisions in a rapidly evolving financial environment.

II. METHODOLOGY

This section outlines the methodology adopted for developing a machine learning model to forecast cryptocurrency prices, specifically for Bitcoin. The approach integrates historical and real-time price data with sentiment analysis. While previous experiments explored LSTM and hybrid models combining LSTM and XGBoost, the final implementation is based solely on XGBoost due to its superior performance in handling structured financial and sentiment data. The methodology is organized into the stages of data collection, preprocessing, feature engineering, model development, and evaluation.

Data Collection

To ensure a comprehensive learning base, the model was trained using both historical and real- time data. Historical price data was gathered using the yFinance API, which provided Open, High, Low, and Close (OHLC) prices of Bitcoin from its inception. Additionally, real-time price updates were periodically fetched from the CoinGecko API at ten-minute intervals, from which daily OHLC values were computed.

This allowed the dataset to represent a wide range of market behaviors, including bull and bear cycles, sudden surges, and corrections.

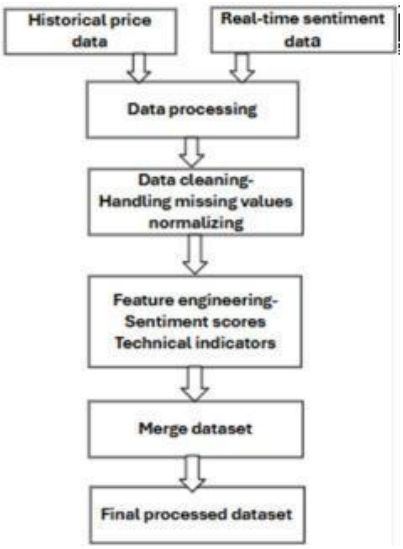

FIGURE 1. Data collection and processing

Sentiment data was obtained from two distinct sources to reflect public opinion. First, Reddit posts mentioning "Bitcoin" were collected on an hourly basis using the PRAW API from the CryptoCurrency subreddit [7],[9]. Second, Wikipedia edit history on the Bitcoin article was utilized to infer historical sentiment based on user edit comments [14],[15]. A sentiment classification pipeline was applied using the DistilBERT transformer model fine-tuned for binary sentiment analysis, classifying content as positive or negative and assigning sentiment scores [4],[5].

Data Preprocessing

To improve data consistency and quality, several preprocessing techniques were applied. All timestamps were converted to a standardized datetime format and sorted chronologically. Missing values in historical and real-time price data were handled using forward-fill and backward-fill strategies to preserve the continuity required for time-series modeling.

Sentiment gaps were smoothed using moving averages over rolling windows. To reduce the impact of anomalies, outliers in the price data were addressed through statistical techniques such as the Z-score and interquartile range (IQR) method.

Both the price and sentiment datasets were merged by aligning the closest timestamp entries, allowing each price point to be associated with its corresponding sentiment context. This combined dataset was used for both training and testing.

The final dataset was exported in CSV format for further analysis and model input.

Feature Engineering

Multiple features were engineered to enhance the predictive capabilities of the model. Technical indicators included moving averages of the closing price over 7-day and 21-day windows, which helped identify short- and medium-term trends. A price_change feature was calculated as the daily percentage change in the closing price. From the sentiment dataset, the raw sentiment score was used along with its 3-day and 7-day rolling averages to smooth out short-term emotional noise and emphasize consistent public sentiment trends. These features were normalized using Min-Max scaling to bring all input variables to the same scale.

The target variable was the next day's closing price, transforming the task into a supervised regression problem. The complete set of features and target values were then split into training and testing subsets for model development.

III. MODEL DEVELOPMENT

The final model implemented in this study is based on the Extreme Gradient Boosting (XGBoost) algorithm, a widely recognized ensemble learning technique known for its effectiveness in structured data prediction tasks. XGBoost builds a series of decision trees sequentially, where each tree attempts to correct the errors of the previous one [12]. The model's efficiency, robustness to overfitting, and ability to handle multicollinearity and missing data make it particularly well suited for financial time series forecasting.

The dataset used to train the model combined historical and real-time OHLC (Open, High, Low, Close) price data along with sentiment scores from Reddit and Wikipedia. Prior to training, feature engineering was performed to construct meaningful predictors such as moving averages (ma_7, ma_21), daily percentage price change (price_change), raw sentiment scores, and rolling averages of sentiment over 3 and 7 days. These features allowed the model to consider both technical market trends and psychological market signals in making predictions.

For the supervised learning task, the target variable was defined as the closing price of the next trading day. The feature-target matrix was then split into training and testing subsets. The training set spanned the period from January 1, 2023, to December 31, 2024, and the testing set included data from January 1, 2025, to March 31, 2025. This chronological split was maintained to preserve temporal consistency and avoid data leakage.

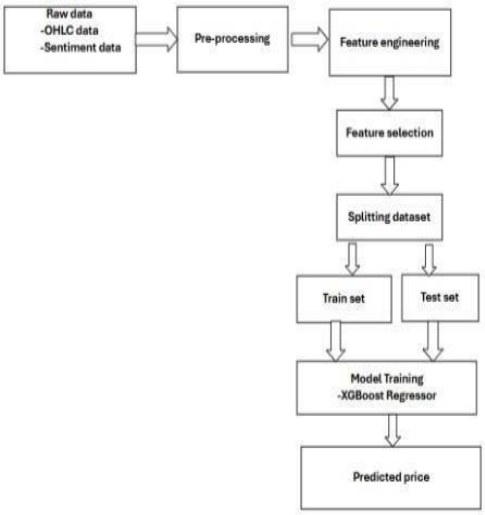

FIGURE 2. Modelling and prediction phase

The XGBoost regressor was configured with the following hyperparameters after preliminary tuning: n_estimators = 100, learning_rate = 0.1, and max_depth = 5. These settings were selected to balance training time with generalization ability.

The model was trained using the full feature set on the training data and was then evaluated on the test set to assess its predictive accuracy.

Upon completion of training, the model was serialized for reuse using joblib, and a prediction function was created to allow for future price forecasting using the most recent available data. This function takes a feature vector derived from the latest OHLC and sentiment information and outputs the predicted closing price for the next day,Providing a ready-to-use interface for deployment or integration into future applications.

IV. EVALUATION

The performance of the XGBoost model was quantitatively evaluated using standard regression metrics tailored for financial time- series applications. These metrics provide insights into both the average prediction accuracy and the sensitivity to larger errors.

Mean Absolute Error (MAE) was used to measure the average magnitude of the model's prediction errors. This metric is

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

intuitive and easy to interpret, offering a direct sense of how much the predicted closing prices deviate from the actual values on average.

Root Mean Squared Error (RMSE) was also calculated to provide a more penalty weighted view of prediction performance. RMSE emphasizes larger errors more than MAE, making it a critical metric in financial forecasting, where significant deviations in predictions can result in major investment losses.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

On the test set, the model produced the following outcomes:

- MAE ≈ 1347.99
- RMSE ≈ 1900.47

These values indicate a strong level of prediction accuracy, particularly considering the highly volatile nature of the cryptocurrency market. The relativel low RMSE demonstrates themodel's ability to minimize large deviations, making it a reliable forecasting tool.

Additionally, a line plot of actual vs. predicted closing prices was generated for the test period (January to March 2025). The plot shows that the predicted prices closely follow the trend of the actual prices, with minimal lag or divergence, further validating the model's effectiveness.

V. MODEL COMPARISION

To validate the choice of the XGBoost model, a comparative analysis was conducted using two other architectures: a standalone LSTM model and a hybrid LSTM + XGBoost model. Each model was trained and evaluated using the same dataset and feature set, allowing for a fair comparison.

The LSTM (Long Short-Term Memory) model was trained on 30-day sliding sequences of historical price data. As a recurrent neural network, LSTM is known for its ability to model long- term temporal dependencies. However, it required significantly more time and computational resources to train. Moreover, while LSTM captured the overall price direction reasonably well, it struggled with sharp fluctuations and lagged behind real-time changes, particularly in volatile periods [10],[11].

The hybrid model attempted to combine the strengths of both LSTM and XGBoost. Predictions from the LSTM network were fed into XGBoost along with other features [4],[13]. Although this approach slightly improved accuracy compared to the LSTM only model, it added significant architectural complexity and yielded only marginal performance gains over XGBoost alone.

TABLE 1. Performance comparision of LSTM, Hybrid (LSTM + XGBoost), XGBoost

MODEL	RMSE	MAE
LSTM	3589.91	2753.60
XGBoost	1925.57	1382.41
Hybrid (LSTM + XGBoost)	2174.84	1644.92

In contrast, the XGBoost model offered the best balance of performance, interpretability, and resource efficiency. It consistently delivered lower MAE and RMSE scores, and its training process was considerably faster and less computationally intensive. Additionally, feature importance scores provided by the model helped explain which technical or sentiment indicators contributed most to price prediction, offering useful insights for further model refinement or decision making applications. Based on these findings, the standalone XGBoost model was selected as the final implementation due to its superior accuracy, simplicity, and practical deployment advantages in the context of cryptocurrency price forecasting.

VI. RESULTS

To evaluate the performance of the developed cryptocurrency price prediction models, three architectures—LSTM, Hybrid (LSTM + XGBoost), and XGBoost—were implemented and tested on the same dataset. Each model was trained using data from January 2023 to December 2024 and evaluated on unseen data from January to March 2025. The performance was assessed using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), which are widely used metrics in financial forecasting.

The LSTM model, despite its strength in handling sequential dependencies, recorded the highest prediction errors with an RMSE of 3589.91 and an MAE of 2753.60. It was observed that the LSTM architecture struggled to adapt to sharp and rapid fluctuations in the price data.

The Hybrid model, which combined LSTM-generated features with XGBoost, improved upon the

LSTM's performance. It achieved an RMSE of 2174.84 and an MAE of 1644.92, indicating that the ensemble of temporal and structured features provided some benefits. However, the added complexity and training overhead did not justify its marginal improvements in accuracy.

The XGBoost model delivered the most accurate and stable results among all models. It achieved an RMSE of 1925.57 and an MAE of 1382.41, demonstrating its effectiveness inlearning from structured data and sentiment- driven features.

In addition to quantitative performance, XGBoost offered faster training, easier interpretability, and greater scalability. These outcomes confirm that XGBoost was chosen as the best prediction model. The performance comparison is summarized in Table I (see previous section).

VII. CONCLUSION

This paper proposed a machine learning framework for cryptocurrency price prediction using structured market data and sentiment analysis. While initial experiments included LSTM and hybrid models, XGBoost emerged as the most effective solution, offering superior accuracy with reduced complexity. In addition to historical and real-time OHLC data, the model leveraged sentiment features from Reddit and Wikipedia using a transformer-based sentiment pipeline.With an RMSE of 1925.57 and MAE of 1382.41, the XGBoost model outperformed both LSTM and hybrid approaches. These results confirm the viability of using structured learning with sentiment-driven features for accurate financial forecasting in volatile markets.

Future enhancements could include support for multiple cryptocurrencies, integration of news sentiment, and deployment within a real-time decision support system.

REFERENCES

[1] Greaves and B. Au, "Using Text Mining and Sentiment Analysis for Online Forums: A Financial Application," in Proc. IEEE Conf. on Intelligent Systems, pp. 118–123, 2018. Internet of Things Journal, vol. 9, no. 4.

[2] F. Rustam, A. Mehmood, G. Muhammad, M. R. A. Memon, and A. de Albuquerque, "Forecasting of Cryptocurrency Using Machine Learning Techniques: A Survey," IEEE Access, vol. 8, pp. 181543 181574, 2020.

[3] Z. Shahbazi and Y.-C. Byun, "Improving the Cryptocurrency Price Prediction Performance Based on Reinforcement Learning," IEEE Access, vol. 9, pp. 162651– 162659, 2021.

[4] N. Patel, H. Shah, and T. Potdar, "Time Series Based Bitcoin Price Prediction Using XGBoost," in Proc. Int. Conf. on Smart Electronics and Communication (ICOSEC), 2021, pp. 622 627.

[5] H. Liu, Y. Wang, and H. Zhang, "Predicting Cryptocurrency Prices with Deep Learning," in Proc. Int.Conf. on Intelligent Computation and Applications, 2021, pp. 101–112.

[6] R. Parekh, P. Modi, J. Ameta, and M. Nene, "DL-GuesS: Deep Learning and Sentiment Analysis-Based Cryptocurrency PricePrediction," IEEE Access, vol. 10, pp. 35398– 35409, 2022.

[7] N. Aslam, F. Rustam, E. Lee, P. B. Washington, and I. Ashraf, "Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model," IEEE Access, vol. 10, pp. 39313–39324, 2022.

[8] Y. Qin et al., "A Dual-Attention Based LSTM Framework for Cryptocurrency Price Forecasting," IEEE Trans. on Computational Social Systems, vol. 9, no. 4, pp. 1031–1042, Aug. 2022.

[9] Li, Y. Li, and Y. Chen, "Multi source Sentiment Cryptocurrency Analysis Forecasting," for IEEE Internet of Things Journal, vol. 9, no. 4, pp. 2687–2695, Feb. 2022.

[10] F. Feizian and B. Amiri, "Cryptocurrency Price Prediction Model Based on Sentiment Analysis and Social Influence," IEEE Access, vol. 11, pp. 142177–142195, 2023.

[11] S. Girsang and Stanley, "Hybrid LSTM and GRU for Cryptocurrency Price Forecasting Based on Social Network Sentiment Analysis Using FinBERT," IEEE Access, vol. 11, pp. 120530–120540, 2023.

[12] Park and Y.-S. Seo, "Twitter Sentiment Analysis-Based Adjustment of Cryptocurrency Action Recommendation Model for Profit Maximization," IEEE Access, vol. 11, pp. 44828-44841, 2023.

[13] M. Zubair et al., "An Improved Machine Learning-Driven Framework for Cryptocurrencies Price Prediction With

Sentimental Cautioning," IEEE Access, vol. 12, pp. 51395–51418, 2024.

[14] K.-H. Ho, Y. Hou, M. Georgiades, and K. C. K. Fong, "Exploring Key Properties and Predicting Price Movements of Cryptocurrency Market Using Social Network Analysis," IEEE Access, vol. 12, pp. 65058–65077, 2024.

[15] S. Chavan, J. Gundakaram, S. Dyuti Vaishnavi, S. Prasad, and K. Deepa, "Real- Time Data Extraction and Prediction of Cryptocurrency," IEEE Access, vol. 12, pp. 186703–186709, 2024.