

Enhancing Bitcoin Price Forecasting Through Integrated Sentiment Analysis and XGBoost Models [†]

Vasileios Dellopoulos ^{1,*}, Ioannis Antoniadis ¹, Evangellos Saprikis ¹ and George Fragulis ²

¹ Department of Management Science and Technology, University of Western Macedonia, GR50100 Kozani, Greece; iantoniadis@uowm.gr (I.A.); esaprikis@uowm.gr (E.S.)

² Department of Electrical and Computer Engineering, University of Western Macedonia, ZEP Campus, GR50150 Kozani, Greece; gfragulis@uowm.gr

* Correspondence: mpp00217@uowm.gr

[†] Presented at the 8th International Global Conference Series on ICT Integration in Technical Education & Smart Society, Aizuwakamatsu City, Japan, 20–26 January 2026.

Abstract

This study investigates Bitcoin price forecasting using integrated sentiment analysis and gradient boosting within digital financial ecosystems. Two XGBoost models were developed using sentiment scores derived from Bitcoin news (2021–2024) and technical indicators, including GARCH-estimated volatility, Bollinger Bands, MACD, and RSI. The analysis uses 1042 daily Bitcoin observations and 10,025 sentiment records. Two model configurations were evaluated: one using only technical indicators and another incorporating daily aggregated sentiment scores. Model performance was assessed using Diebold–Mariano tests with Newey–West HAC variance estimation and walk-forward validation across 40 rolling windows. Contrary to expectations, sentiment features provided no statistically significant improvement over the technical-only model ($p = 0.4888$). Both models achieved identical test performance ($R^2 = -0.16\%$). Walk-forward validation revealed substantial temporal instability (Mean $R^2 = -126.30\%$, Std = 233.05%), highlighting the challenges of forecasting daily Bitcoin returns. Nevertheless, both XGBoost models significantly outperformed the random walk benchmark (DM statistic = -8.58 , $p < 0.0001$), indicating that technical indicators capture exploitable market structure despite limited predictive accuracy for practical trading. These findings support the efficient market hypothesis and have implications for digital financial ecosystems integrating multimodal information.

Keywords: Bitcoin price forecasting; sentiment analysis; natural language processing; XGBoost; machine learning; cryptocurrency forecasting; digital financial ecosystems

1. Introduction

The emergence of cryptocurrency markets has fundamentally challenged traditional financial analysis paradigms. Unlike equities or fixed-income securities, Bitcoin lacks conventional valuation measures such as earnings, cash flows, and balance sheets [1]. Consequently, sentiment analysis—the computational extraction of opinions and emotions from textual data—has emerged as a promising approach for cryptocurrency price forecasting [1–4].

The continued evolution of digital financial ecosystems increasingly relies on integrating alternative information sources, including news, social media, and blockchain analytics, with quantitative forecasting models [5,6]. Among these sources, news sentiment provides a lower-latency information signal than conventional price discovery mechanisms,



Academic Editors: Debopriyo Roy and Peter Ilic

Published: 2 July 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

particularly in emerging and highly volatile markets [5]. Consequently, numerous studies have reported substantial predictive improvements in cryptocurrency forecasting through sentiment-enhanced models [1,5,6].

Despite these advances, several methodological limitations remain. Many previous studies rely on limited out-of-sample evaluation, omit formal statistical comparisons between competing models, and overlook non-stationarity in financial time series [1,2,6]. Furthermore, the relationship between news sentiment and daily Bitcoin returns remains insufficiently understood under conditions dominated by market noise and rapidly changing trading dynamics [5].

This study addresses these limitations by implementing rigorous statistical evaluation using the Diebold–Mariano test with Newey–West HAC variance estimation [7,8]. It further applies walk-forward validation across multiple temporal windows instead of relying solely on a fixed train–test split. In addition, GARCH modeling captures volatility clustering while maintaining strict temporal alignment throughout the forecasting framework [9].

The proposed framework also incorporates production-grade machine learning pipelines with systematic hyperparameter optimization using seven-fold time-series cross-validation and fifty candidate model configurations [10]. Finally, the study reports negative findings transparently, providing a realistic empirical benchmark for future cryptocurrency forecasting research. The principal finding—that sentiment features provide no additional predictive value beyond technical indicators—contrasts with many previous studies and has important implications for sentiment integration within digital financial ecosystems.

2. Methodology

2.1. Data Sources and Preprocessing

2.1.1. Bitcoin Price Data

Historical Bitcoin price data covering the period from 5 November 2021 to 12 September 2024 were retrieved using the *yfinance* Python library, which provides programmatic access to Yahoo Finance market data. The resulting dataset comprised 1042 daily observations with the following features:

- Open: Opening price recorded at 00:00 UTC.
- High: Intraday maximum price.
- Low: Intraday minimum price.
- Close: Closing price recorded at 23:59 UTC.
- Volume: Daily BTC-USD trading volume in satoshis.

Dataset characteristics: Daily frequency (D1); Observation period: 1042 trading days (2.75 years); Missing observations: None.

2.1.2. Sentiment Data

Sentiment annotations were obtained from the publicly available Kaggle dataset “Sentiment Analysis of Bitcoin News 2021–2024” [11]. The sentiment dataset used in this study is a forked and extended version of *Oliviervha*’s Bitcoin news corpus on Kaggle, originally covering multiple news sources between 12 October 2021 and 19 December 2023. The extended dataset incorporates additional Bitcoin-related articles from *Binance News* and *Yahoo Finance*, expanding the temporal coverage through 12 September 2024. In the original dataset, article sentiment was derived using *TextBlob* and assigned as positive, neutral, or negative; in the extended version, these labels are replaced by scores from a transformer model fine-tuned on financial news (*mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis*), accessed via a text-classification pipeline. Articles with scores above zero are treated as positive, scores equal to zero as neutral, and scores below

zero as negative, this approach provides a more robust, finance-specific sentiment signal suitable for market analysis and cryptocurrency forecasting.

2.1.3. Data Alignment and Forward-Fill Strategy

Temporal alignment between the daily price series and the irregularly sampled sentiment data required missing-value imputation. Forward-fill imputation was applied to missing sentiment observations: $[(t) = (t - 1) t]$. This approach assumes sentiment inertia within a trading day, with market participants operating with information from the most recent sentiment signal available. Forward-filling addressed 843 missing dates (33.8% of the training period), introducing potential look-ahead bias if articles published at market close affected next-day returns. This limitation is acknowledged in Section 4.

2.1.4. Software

Python version 3.14.2 (Python Software Foundation, Wilmington, DE, USA) was used for data preprocessing, feature engineering, model development, statistical analysis, visualization, and performance evaluation. The analytical workflow employed pandas and NumPy for data manipulation, yfinance for financial data retrieval, scikit-learn for preprocessing, machine learning pipelines, model validation, and performance evaluation, XGBoost for predictive modeling, SciPy and statsmodels for statistical analyses, arch for GARCH volatility estimation, ta for technical indicator computation, matplotlib and seaborn for visualization, and joblib for model persistence.

2.2. Feature Engineering

2.2.1. Technical Indicators

A comprehensive set of 21 technical features was engineered to capture multiple dimensions of Bitcoin market behavior, including price dynamics, volatility, momentum, and trading activity. Price-based features comprised normalized Open, High, Low, and Close prices (standardized using z-scores), daily returns, and one- and three-day lagged returns to capture short-term momentum persistence. Volatility features included conditional volatility estimated using a GARCH(1,1) model with Student- t innovations [9], enabling the model to capture volatility clustering and the heavy-tailed return distributions characteristic of cryptocurrency markets. Historical volatility was additionally computed as the 20-day rolling standard deviation of daily returns. Trend-following indicators included short- and medium-term simple moving averages (SMA-5 and SMA-20) together with exponential moving averages (EMA-12 and EMA-26), where the latter assign greater weight to recent observations. Momentum indicators comprised the Moving Average Convergence Divergence (MACD), calculated as the difference between short- and long-term exponential moving averages together with its signal line, and the Relative Strength Index (RSI), a bounded momentum oscillator ranging from 0 to 100. RSI values above 70 indicate overbought conditions, whereas values below 30 indicate oversold conditions. Bollinger Bands were computed using the 20-day moving average and its associated rolling standard deviation, providing a dynamic measure of price deviation from the recent trading range. Volume-related features included raw daily trading volume (in satoshis), trading volume normalized by its 30-day moving average, and a volume-price interaction feature computed as daily trading volume multiplied by the absolute daily return. Finally, Configuration B incorporated an additional sentiment feature, defined as the daily aggregated sentiment score scaled between -1 and $+1$.

2.2.2. Correlation Analysis

Pearson correlation analysis was performed to evaluate linear relationships among the engineered features and to identify potential multicollinearity prior to model development.

2.3. Modeling Approach

2.3.1. Data Partitioning

Data were partitioned into mutually exclusive train and test sets:

- Train set: 5 November 2021–19 December 2023 (738 observations).
- Test set: 20 December 2023–12 September 2024 (266 observations).
- Train/test ratio: 73.5%/26.5%.

This temporal split prevents look-ahead bias and reflects a realistic out-of-sample evaluation framework, in which models are trained on historical observations and evaluated using future unseen data.

2.3.2. Model Specification: XGBoost

XGBoost (Extreme Gradient Boosting) was selected as the primary machine learning algorithm because of its strong theoretical foundation and demonstrated empirical performance on structured tabular data [10]. As a gradient boosting framework, it constructs an ensemble of decision trees sequentially, where each successive tree minimizes the residual errors of its predecessors, thereby enabling the modeling of complex, non-linear relationships. The algorithm incorporates both L1 (lasso) and L2 (ridge) regularization terms within its objective function to mitigate overfitting and enhance generalization. Model interpretability was evaluated using feature importance metrics. XGBoost reduces the impact of multicollinearity through implicit feature selection during tree construction, while its optimized parallel implementation ensures computational scalability for large datasets.

Two model specifications were trained: Configuration A, comprising 21 technical indicators exclusively, and Configuration B, which extended the feature space to 22 variables by incorporating daily aggregated sentiment. The primary objective was to determine whether incorporating the sentiment feature yielded incremental predictive value relative to the technical-only baseline.

2.3.3. Hyperparameter Optimization

Hyperparameter optimization was performed using randomized search combined with 7-fold time-series cross-validation, thereby preserving temporal ordering and preventing look-ahead bias. For each model configuration, 50 randomly sampled hyperparameter combinations were evaluated with the objective of minimizing the Mean Squared Error (MSE). The optimal hyperparameter values for both XGBoost configurations are summarized in Table 1.

Table 1. Optimal hyperparameter values for the two XGBoost configurations.

Hyperparameter	Configuration A (Technical Features)	Configuration B (Technical + Sentiment)
Max depth	3	5
Number of estimators	130	60
Learning rate	0.077	0.157
Subsample	0.639	0.715
Column sample by tree	0.717	0.721
Gamma	0.432	0.335

Source: Author’s own calculations based on the implemented code.

Configuration A selected shallower trees and a larger number of estimators, indicating a more conservative boosting strategy. In contrast, Configuration B favored deeper trees, fewer estimators, and a higher learning rate, reflecting a more aggressive learning strategy after incorporating the sentiment feature.

Although the optimal hyperparameters differed between the two configurations, both models achieved nearly identical out-of-sample performance. This finding suggests that incorporating sentiment altered the optimization landscape without improving predictive accuracy.

2.3.4. Preprocessing Pipeline

A standardized preprocessing pipeline was implemented to ensure methodological rigor and to prevent data leakage. Continuous features were normalized using scikit-learn's StandardScaler, transforming them to zero mean and unit variance based solely on training data statistics. Feature alignment was strictly enforced to maintain identical feature ordering across the training, validation, and test splits. Missing values in sentiment-related variables were addressed by forward-fill imputation, whereas technical features contained no missing observations and therefore required no imputation. The preprocessing workflow was implemented using a scikit-learn Pipeline, guaranteeing that all transformation parameters were fitted exclusively on the training set and subsequently applied to validation and test data, thereby preventing information leakage from future observations.

2.4. Statistical Testing Methodology

2.4.1. Diebold–Mariano Test for Forecast Comparison

The Diebold–Mariano (DM) test, proposed by Diebold and Mariano [7], was employed to compare the predictive accuracy of competing forecasting models. The test evaluates whether the mean loss differential between two forecasts differs significantly from zero. The loss differential was computed using squared forecast errors (Mean Squared Error loss) at each prediction horizon, and the mean of these differences formed the basis of the test statistic.

Because forecast errors in time-series applications often exhibit serial correlation and heteroskedasticity, variance estimation was performed using the Newey–West heteroskedasticity and autocorrelation consistent (HAC) estimator [8]. This estimator adjusts the variance of the mean loss differential by accounting for serial dependence through automatic bandwidth selection.

The null hypothesis states that both forecasting models possess equal predictive accuracy, whereas the alternative hypothesis assumes a statistically significant difference in predictive performance. Statistical significance was assessed using a two-tailed test at the 5% significance level.

2.4.2. Walk-Forward Validation

Static train–test splits may conceal temporal instability in forecasting performance, particularly in financial time series. To address this limitation, walk-forward (rolling-window) validation was performed across 40 sequential temporal windows. The rolling-window design provides a more realistic assessment of forecasting performance than a single fixed train–test split because model parameters are repeatedly re-estimated using only information available at each evaluation period.

For each iteration, the training window consisted of 200 observations (approximately nine months), whereas the test window contained 15 observations (approximately three weeks). The evaluation window then advanced by five observations (approximately one week), and the procedure was repeated for all 40 windows.

Each model was retrained independently within every rolling window, producing 40 out-of-sample estimates of R^2 and RMSE. This evaluation framework assesses model stability over time and determines whether predictive performance varies across different market conditions, including bullish markets, market corrections, and periods of elevated volatility.

3. Results

3.1. Primary Findings: Model Performance

The predictive performance of both XGBoost configurations on the held-out test set (266 observations; December 2023–September 2024) is summarized in Table 2.

Table 2. Predictive performance of the XGBoost models and the random walk benchmark on the held-out test set.

Metric	Technical Only	Technical + Sentiment	Random Walk
MSE	0.000828	0.000828	0.001704
RMSE	0.028777	0.028776	0.041277
MAE	0.020864	0.020864	0.031458
MAPE	106.99%	107.51%	1566.45%
R ²	−0.16%	−0.16%	−106.08%

Source: Author's own calculations based on the implemented code.

The negative R² values indicate that both XGBoost configurations performed slightly worse than a baseline predictor based on the mean of the observed returns. Specifically, R² = −0.16% indicates that the prediction errors marginally exceeded the variance explained by the baseline predictor, a result that is not uncommon in highly efficient financial markets. The MAPE values exceeding 100% reflect the presence of numerous near-zero daily returns, causing percentage-based error measures to become highly sensitive to relatively small prediction errors. Both XGBoost configurations produced virtually identical predictive performance across all evaluation metrics despite employing different feature sets and distinct hyperparameter configurations. Formal statistical comparison of the two models is presented in the subsequent Diebold–Mariano analysis. Although both XGBoost configurations yielded negative R² values, they substantially outperformed the random walk benchmark across all error-based evaluation metrics.

3.2. Diebold–Mariano Test Results

The Diebold–Mariano (DM) test was conducted to determine whether differences in predictive accuracy between the competing forecasting models were statistically significant. The results are summarized in Table 3.

Table 3. Diebold–Mariano test results for forecast comparison.

Model Comparison	DM Statistic	<i>p</i> -Value	Conclusion
With Sentiment vs. Without Sentiment	−0.6922	0.4888	Fail to reject H_0
XGBoost (Technical Only) vs. Random Walk	−8.5781	<0.0001	Reject H_0
XGBoost (Technical + Sentiment) vs. Random Walk	−8.5780	<0.0001	Reject H_0

Source: Author's own calculations based on the implemented code.

For the comparison between the two XGBoost configurations, the DM statistic was −0.6922 with a *p*-value of 0.4888, indicating no statistically significant difference in predictive accuracy. Consequently, the null hypothesis of equal predictive performance cannot be rejected, suggesting that the inclusion of sentiment features provides no incremental predictive value.

In contrast, both XGBoost configurations significantly outperformed the random walk benchmark ($p < 0.0001$). Relative to the benchmark, the Mean Squared Error (MSE) decreased from 0.001704 to 0.000828, corresponding to a 51.4% reduction in forecasting error.

Although both XGBoost models achieved significantly lower prediction errors than the random walk benchmark, this improvement reflects greater numerical forecasting accuracy rather than enhanced directional prediction. The models reduced forecast errors but remained unable to consistently capture the direction of daily Bitcoin price movements, consistent with the weak predictability expected in highly efficient cryptocurrency markets.

3.3. Feature Importance Analysis

Despite identical predictive performance, the feature rankings differed slightly between the two XGBoost configurations. The highest-ranked features identified by each model are summarized in Table 4.

Table 4. Comparison of the highest-ranked features identified by the two XGBoost configurations.

Rank	Technical Only	Technical + Sentiment
1	Open	Open
2	MACD Signal	High
3	Volume–Price Interaction	Volume–Price Interaction
4	Volume-Normalized	Volume-Normalized
5	Lagged Return (3-day)	Lagged Return (3-day)
8	–	Sentiment

Source: Author’s own calculations based on the implemented code.

Although the feature rankings differed slightly between the two configurations, all estimated feature importance values were effectively zero (0.0000). This finding indicates that neither model identified any feature with meaningful predictive influence on daily Bitcoin returns.

These findings are consistent with the poor predictive performance reported previously. Under the selected feature set and modeling framework, daily Bitcoin returns appear to contain limited exploitable predictive information. Possible explanations include the scarcity of persistent market patterns, substantial multicollinearity among the technical indicators, and the regularization strategy employed by XGBoost, which favored simpler tree structures.

3.4. Walk-Forward Validation: Temporal Stability

Walk-forward validation was performed across 40 sequential rolling windows to evaluate the temporal stability of the forecasting models. The aggregated performance statistics are summarized in Table 5.

Table 5. Summary statistics of walk-forward validation performance.

Statistic	Value
Number of windows (n)	40
Mean R^2	−126.30%
Standard deviation (R^2)	233.05%
Minimum R^2	−1332.49% (Window 1)
Maximum R^2	4.10% (Window 28)
Range	1336.59 percentage points
Windows with positive R^2	1 (2.5%)
Windows with $R^2 < -100\%$	22 (55%)
Median R^2	−35.45%
25th percentile	−108.36%
75th percentile	−22.85%

Source: Author’s own calculations based on the implemented code.

Figure 1 illustrates the variation in out-of-sample predictive performance across the 40 walk-forward validation windows.

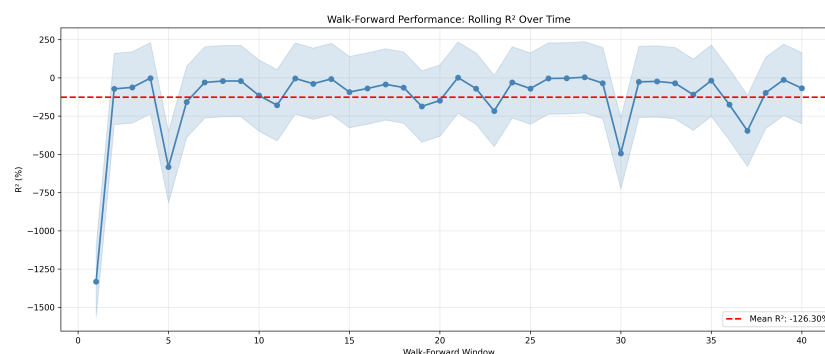


Figure 1. Rolling out-of-sample R^2 values across 40 walk-forward validation windows.

Early validation windows (1–5) exhibited extremely negative R^2 values (ranging from -72.15% to -1332.49%), indicating substantial model instability during the initial evaluation period. Intermediate windows (8–20) showed modest improvement (R^2 ranging from -114.97% to 4.10%), with only one window producing a positive R^2 value. Performance deteriorated again during the final evaluation period (Windows 30–40), where the mean R^2 decreased to -65.68% .

The pronounced temporal variability indicates that the models captured short-lived market-specific patterns that failed to generalize across successive evaluation windows.

Possible explanations include:

- Market regime shifts associated with transitions between bullish and bearish market conditions during 2023–2024.
- Volatility clustering that was not fully captured through static feature engineering.
- Non-stationarity in the relationship between sentiment and Bitcoin returns over time.

3.5. Residual Analysis

Residual analysis was performed to examine the distribution of prediction errors for each forecasting model. Summary statistics are presented in Table 6, while Figure 2 illustrates the corresponding residual distributions.

Table 6. Summary statistics of model residuals.

Statistic	Technical Only	Technical + Sentiment	Random Walk
Mean residual	0.0012	0.0011	0.0018
Standard deviation	0.0237	0.0237	0.0330
Distribution shape	Slight right-skew	Slight right-skew	Slight right-skew

Source: Author's own calculations based on the implemented code.

Both XGBoost configurations produced nearly identical residual distributions, further confirming that the inclusion of sentiment features did not improve predictive performance. As shown in Figure 2, the residuals are centered close to zero and exhibit slight positive skewness, with occasional larger positive prediction errors. Compared with the XGBoost models, the random walk benchmark displays a noticeably wider residual distribution, consistent with its higher residual variance and poorer forecasting performance.

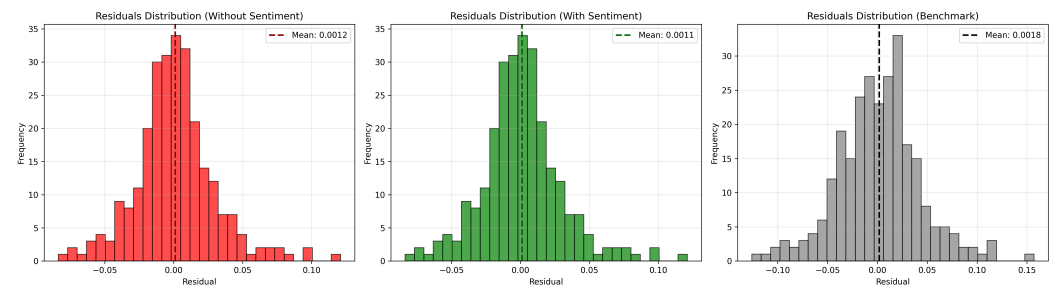


Figure 2. Residual distributions for the Technical Only model, the Technical + Sentiment model, and the random walk benchmark. Dashed vertical lines indicate the mean residual for each model.

3.6. Correlation Findings

Feature correlation analysis revealed substantial multicollinearity among the technical indicators. Price-based variables (Open, High, Low, and Close) exhibited near-perfect pairwise correlations ($r = 0.98\text{--}1.00$), indicating considerable redundancy. Similarly, moving average indicators (SMA-5, SMA-20, EMA-12, and EMA-26) showed correlations exceeding 0.99 both with the underlying price series and among themselves, reflecting their construction from overlapping price windows. Bollinger Bands likewise demonstrated strong dependence on price movements, with correlations ranging from 0.97 to 0.99.

The correlation analysis revealed only modest linear associations between sentiment and the technical indicators, as summarized in Table 7. The strongest observed linear association was with historical volatility ($r = 0.34$), whereas all remaining correlations were below 0.35.

Table 7. Correlation between sentiment scores and technical indicators.

Technical Indicator	Correlation with Sentiment
Close Price	0.30
RSI	−0.15
MACD	−0.13
MACD Signal	−0.15
Volatility	0.34
Upper Bollinger Band	0.32
Volume	0.08
GARCH Volatility	0.12

Source: Author's own calculations based on the implemented code.

Despite these relatively weak linear associations, the inclusion of sentiment did not improve predictive performance in either XGBoost configuration. This finding demonstrates that weak linear associations with technical indicators do not necessarily imply incremental predictive value. Several explanations may account for this result:

1. Bidirectional causality: Market participants may incorporate news into prices rapidly, leaving little exploitable delay between information arrival and price adjustment.
2. Non-linear relationships: Any remaining relationships between sentiment and returns may be highly nonlinear or regime-dependent.
3. Confounding variables: Unobserved market factors, including derivatives activity and institutional positioning, may jointly influence both sentiment and price dynamics.

4. Discussion

The absence of incremental predictive value from sentiment analysis has important implications for digital financial ecosystems and smart societies that increasingly integrate heterogeneous data sources into financial decision-making.

4.1. Market Efficiency and Sentiment

The empirical findings are consistent with the semi-strong form of market efficiency [12], whereby publicly available information, including news-based sentiment, is rapidly incorporated into Bitcoin prices. At the daily forecasting horizon, Bitcoin appears to exhibit a level of informational efficiency that limits the predictive value of publicly available sentiment information.

Continuous 24/7 trading eliminates market closures, allowing new information to be incorporated into prices without interruption. In addition, broad global participation and the strong presence of retail investors may accelerate the dissemination and processing of news sentiment across cryptocurrency markets. The widespread adoption of algorithmic and high-frequency trading systems further reduces the time available to exploit publicly available information.

Consequently, any predictive advantage derived from daily aggregated news sentiment may disappear before the next trading period. These findings provide a plausible explanation for the absence of statistically significant improvements after incorporating sentiment into the XGBoost forecasting models.

4.2. Smart Ecosystem Architecture

The findings have important implications for financial institutions developing smart ecosystems that integrate alternative data sources, including news sentiment, social media activity, and blockchain analytics. Although sentiment did not improve daily Bitcoin price forecasting in this study, its practical value should not be dismissed entirely.

Sentiment information may remain useful at shorter forecasting horizons, such as hourly or 30-min intervals, where market microstructure effects are more pronounced. Likewise, the absence of predictive value for Bitcoin does not necessarily extend to other cryptocurrencies. Less mature and less efficient altcoins may exhibit stronger sentiment-driven price dynamics.

In addition, persistent divergences between market sentiment and price movements may serve as anomaly detection signals deserving further investigation. Sentiment indices may also complement portfolio risk management by supporting hedging strategies, portfolio diversification, and market monitoring, even when they do not directly improve return prediction.

4.3. Methodological Contributions

The methodological rigor of this study strengthens confidence in the reported null findings. The application of the Newey–West heteroskedasticity and autocorrelation consistent (HAC) estimator [8] accounted for potential serial correlation in forecast errors, thereby supporting robust statistical inference. Walk-forward validation reduced the risk of data-snooping bias associated with fixed train–test splits and provided a more reliable assessment of out-of-sample performance.

The incorporation of GARCH-based volatility estimates [9] captured the volatility clustering and heavy-tailed characteristics commonly observed in cryptocurrency returns. In addition, randomized hyperparameter optimization using 50 candidate configurations [10] and 7-fold time-series cross-validation reduced the likelihood of suboptimal model specification. Finally, the transparent reporting of negative findings contributes to reducing publication bias and encourages more balanced evidence within the cryptocurrency forecasting literature.

4.4. Limitations and Caveats

Several limitations should be considered when interpreting the findings of this study. First, the sentiment dataset presents important data-related constraints. The publicly available Kaggle dataset does not document the underlying natural language processing methodology used to generate the sentiment scores. Consequently, the scores may have been produced using lexicon-based approaches, such as VADER or TextBlob [3,4], rather than domain-specific transformer models such as FinBERT [13]. This limitation may partially explain the weak predictive signal observed in the present analysis. In addition, potential article-selection bias cannot be excluded if the dataset disproportionately represents major news events while underrepresenting routine market coverage. Furthermore, averaging sentiment at the daily level may obscure meaningful intraday dynamics, thereby weakening short-term relationships between sentiment and Bitcoin returns.

Methodological limitations should also be acknowledged. The analysis was restricted to XGBoost models, leaving open the possibility that alternative deep learning architectures, including LSTM and Transformer-based models, may better capture nonlinear and temporal dependencies between sentiment and cryptocurrency returns. The study also focused exclusively on daily BTC–USD returns, excluding potentially informative multivariate variables such as trading volume, derivatives market activity, and additional macro-financial indicators.

Finally, the temporal and asset coverage of the study limits the generalizability of the findings. The 2021–2024 sample captures only a limited number of major market regime transitions, which may reduce the ability to evaluate sentiment effects under a wider range of market conditions. In addition, Bitcoin represents one of the most mature and informationally efficient cryptocurrency markets. Consequently, the reported findings may not directly generalize to younger or less efficient cryptocurrencies, where sentiment-driven price dynamics could be considerably stronger.

4.5. Reconciliation with Prior Literature

The findings of this study differ from several previous investigations reporting improved cryptocurrency forecasting performance after incorporating sentiment information [1,5].

These differences may be explained by methodological and structural variations across studies. Previous research frequently employed different datasets, prediction horizons, validation procedures, and target cryptocurrencies, making direct comparison challenging. For example, Bhatt et al. combined news sentiment with on-chain information, whereas Long et al. focused on TikTok-based sentiment and speculative cryptocurrencies such as Dogecoin rather than Bitcoin.

The present findings suggest that the predictive value of sentiment depends on the forecasting context rather than representing a universally applicable source of information. Differences in market efficiency, asset characteristics, data quality, and evaluation methodology may substantially influence the observed contribution of sentiment features.

Moreover, the rigorous evaluation framework adopted in this study, including walk-forward validation and the Diebold–Mariano test with Newey–West HAC variance estimation, reduces the likelihood that the reported results are driven by overfitting or favorable sample selection. Consequently, the absence of statistically significant improvements should not be interpreted as evidence that sentiment lacks economic relevance. Rather, it indicates that, under the examined daily forecasting horizon and modeling framework, aggregated news sentiment did not provide incremental predictive value beyond traditional technical indicators.

Finally, publication bias may also contribute to discrepancies within the literature. Positive findings are more likely to be published than null results, potentially overstating

the effectiveness of sentiment-based forecasting models. Transparent reporting of statistically supported negative findings is therefore essential for developing a more balanced and reproducible body of evidence in cryptocurrency forecasting.

4.6. Future Research Directions

The limitations identified in this study provide several opportunities for future research. One important direction involves evaluating alternative natural language processing architectures. Comparing domain-specific transformer models, such as FinBERT [13], BERT [14], and RoBERTa [15] and recent fine-tuned financial language models [16], with traditional lexicon-based sentiment approaches may improve sentiment quality and reveal model-specific differences in cryptocurrency forecasting.

Another promising direction concerns the prediction horizon. Rather than focusing exclusively on daily returns, future studies should investigate intraday, multiday, weekly, and monthly forecasting horizons. Such analyses may determine whether sentiment influences cryptocurrency prices immediately or accumulates over longer periods.

Future work should also examine more advanced forecasting algorithms. Deep learning architectures, including Long Short-Term Memory (LSTM) networks and Transformer-based models [6,14,15], may better capture nonlinear relationships and long-term temporal dependencies between sentiment signals and cryptocurrency returns.

Expanding the analysis beyond Bitcoin represents another important research direction. Applying the proposed framework to alternative cryptocurrencies may reveal stronger sentiment effects in markets characterized by lower efficiency and reduced liquidity.

The practical applicability of sentiment-based forecasting should also be evaluated in realistic trading environments. Live or simulated trading experiments would enable researchers to account for transaction costs, latency, slippage, and other operational constraints that influence the economic value of forecasting models.

Future studies may further benefit from integrating multiple information sources. Combining news sentiment with social media data from platforms such as Twitter, Reddit, and Discord could provide a more comprehensive representation of market sentiment and improve forecasting performance.

Finally, causal inference methods should be explored to better understand the relationship between sentiment and cryptocurrency prices. Techniques such as Granger causality testing [17] and vector autoregression (VAR) [18] may help determine whether sentiment drives market movements or whether price changes influence subsequent media sentiment.

5. Conclusions

This study presented a rigorous empirical evaluation of whether news sentiment provides incremental predictive value for daily Bitcoin forecasting within an XGBoost framework. Using 1042 daily observations, rigorous statistical evaluation, and comprehensive out-of-sample validation, the results demonstrate that aggregated news sentiment does not significantly improve predictive performance beyond traditional technical features.

Five principal findings emerged from the analysis. First, XGBoost models with and without sentiment features achieved virtually identical predictive performance on the held-out test set ($R^2 = -0.16\%$). Second, the Diebold–Mariano test ($p = 0.4888$) confirmed that the difference between the two forecasting models was not statistically significant. Third, walk-forward validation revealed substantial temporal instability (mean $R^2 = -126.30\%$, standard deviation = 233.05%), highlighting the inherent difficulty of forecasting daily Bitcoin returns. Fourth, despite their limited predictive accuracy, both XGBoost configurations significantly outperformed the random walk benchmark ($p < 0.0001$). Finally, feature importance

analysis indicated that none of the examined variables contributed meaningfully to model predictions under the adopted modeling framework.

These findings contribute to the cryptocurrency forecasting literature by providing robust statistical evidence that challenges the assumption that sentiment information consistently improves daily Bitcoin price prediction. The results also demonstrate the importance of rigorous validation procedures, including walk-forward evaluation and formal statistical testing, when assessing predictive models in financial markets.

Although aggregated news sentiment did not improve daily Bitcoin forecasting, this finding should not be interpreted as evidence that sentiment lacks practical value. Sentiment information may remain useful at shorter forecasting horizons, for less efficient cryptocurrency markets, or in applications beyond return prediction, including anomaly detection, portfolio risk management, and market monitoring.

Overall, the findings are consistent with the semi-strong form of the Efficient Market Hypothesis [12], suggesting that publicly available news sentiment is rapidly incorporated into Bitcoin prices. Future research should investigate alternative sentiment representations, advanced deep learning and large language models, multimodal data integration, and different prediction horizons to further clarify the role of sentiment within digital financial ecosystems.

Author Contributions: Conceptualization, V.D.; methodology, V.D.; software, V.D.; validation, V.D.; formal analysis, V.D.; investigation, V.D.; resources, V.D.; data curation, V.D.; writing—original draft preparation, V.D.; writing—review and editing, V.D., I.A., E.S. and G.F.; visualization, V.D.; supervision, I.A., E.S. and G.F.; project administration, I.A., E.S. and G.F. 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 source code, processed datasets, and supporting materials generated during this study are publicly available at: <https://github.com/Vassilis-Dell/Bitcoin-price-prediction-with-Sentiment-Analysis-using-XGBoost> (accessed on 26 February 2026). The sentiment dataset analyzed in this study was derived from the publicly available Kaggle dataset “Sentiment Analysis of Bitcoin News 2021–2024” by Imad Allal, which is cited in the References.

Acknowledgments: This work was conducted within the framework of the Master’s Programme in e-Business and Digital Marketing at the University of Western Macedonia, Department of Management Science and Technology. The authors gratefully acknowledge the University of Western Macedonia for providing the academic environment that supported this research. The authors would like to express their sincere gratitude to Ioannis Antoniadis, supervisor of the underlying master’s thesis, for his encouragement to extend this work into a conference paper, as well as for his continuous guidance, valuable comments, and editorial suggestions throughout the preparation of this manuscript. During the preparation of this manuscript, Python version 3.14.2 was used to develop the forecasting models, perform statistical analyses, generate tables and figures, conduct walk-forward validation, produce correlation heatmaps, perform Diebold–Mariano tests, compare predictive performance metrics, and evaluate feature importance with and without sentiment. ChatGPT-5 (OpenAI) and Perplexity AI (<https://www.perplexity.ai>) were used to assist with code review, debugging, optimization suggestions, language editing, text polishing, and improving the clarity of the manuscript. The authors reviewed, verified, and edited all AI-assisted output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

BTC	Bitcoin
DM	Diebold–Mariano
EMA	Exponential Moving Average
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
HAC	Heteroskedasticity and Autocorrelation Consistent
LSTM	Long Short-Term Memory
MACD	Moving Average Convergence Divergence
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
NLP	Natural Language Processing
RMSE	Root Mean Squared Error
RSI	Relative Strength Index
SMA	Simple Moving Average
VAR	Vector Autoregression
R ²	Coefficient of Determination
XGBoost	Extreme Gradient Boosting

References

1. Bhatt, S.; Ghazanfar, M.A.; Amirhosseini, M. Sentiment-driven cryptocurrency price prediction: A machine learning approach utilizing historical data and social media sentiment analysis. *Mach. Learn. Appl. Int. J.* **2023**, *10*, 1–15. [CrossRef]
2. Zamani, N.A.M.; Liew, J.A.S.; Ahmad, S. XLNet-GRU sentiment regression model for cryptocurrency news in English and Malay. In *Proceedings of the 4th Financial Narrative Processing Workshop, Marseille, France, 24 June 2022*; European Language Resources Association: Paris, France, 2022; pp. 35–47.
3. Mishev, K.; Gjorgjevikj, A.; Vodenska, I.; Chitkushev, L.T.; Trajanov, D. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access* **2020**, *8*, 131662–131682. [CrossRef]
4. Malo, P.; Sinha, A.; Korhonen, P.; Wallenius, J.; Takala, P. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.* **2014**, *65*, 782–796. [CrossRef]
5. Long, S.; Xie, Y.; Zhou, Z.; Lucey, B.; Urquhart, A. From whales to waves: Social media sentiment, volatility, and whales in cryptocurrency markets. *Br. Account. Rev.* **2025**, 101682. [CrossRef]
6. Li, S.; Wang, R.; Wan, Y.; Hu, J. Cryptocurrency price prediction using LSTM and FEDformer enhanced by sentiment analysis. *IEEE Access* **2025**, *13*, 213306–213322. [CrossRef]
7. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
8. Newey, W.K.; West, K.D. Hypothesis testing with efficient method of moments estimation. *Int. Econ. Rev.* **1987**, *28*, 777–787. [CrossRef]
9. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
10. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
11. Allal, I. Sentiment Analysis of Bitcoin News 2021–2024. Kaggle. 2024. Available online: <https://www.kaggle.com/datasets/imadallal/sentiment-analysis-of-bitcoin-news-2021-2024> (accessed on 2 December 2025).
12. Fama, E.F.; Malkiel, B.G. Efficient capital markets: A review of theory and empirical work. *J. Financ.* **1970**, *25*, 383–417. [CrossRef]
13. Araci, D. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv* **2019**, arXiv:1908.10063.
14. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
15. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
16. Nasiopoulos, D.K.; Roumeliotis, K.I.; Sakas, D.P.; Toudas, K.; Reklitis, P. Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models. *Int. J. Financ. Stud.* **2025**, *13*, 75. [CrossRef]

17. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438. [[CrossRef](#)]
18. Sims, C.A. Macroeconomics and reality. *Econometrica* **1980**, *48*, 1–48. [[CrossRef](#)]

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.