Sentiment-Based Bitcoin Volatility Prediction: A Comparative Study of Linear and Non-Linear Models

Zixin Ye

{zy4874@my.utexas.edu}

Department of Economics, University of Texas at Austin, Austin, TX 78701, The United States

Abstract. The paper explores the forecasting ability of social media aggregate sentiment measures, with historic variance, of Bitcoin return volatility. The researchers use a data set consisting of more than 16 million 2014-2019 tweets and the results of linear Ordinary Least Squares (OLS) benchmark predictions are compared with those of nonlinear ensemble methods (Random Forest and XGBoost). As a result of empirical findings, it is clear that although the XGBoost method has the lowest Mean Absolute Error, as it strongly represents the median trend, the linear OLS model has the lowest Mean Squared Error as a result of avoiding colossal over-predictions. On the other hand, random forest model shows a high level of overfitting. Those results indicate that sentiment and structural volatility clustering are valuable predictors, although the selection and regularization of the model are essential to identifying the nature of operations within cryptocurrencies markets without being noise.

Keywords: Bitcoin, Volatility forecasting, Investor sentiment.

1 Introduction

Cryptocurrency market has become a unique asset market which has increased volatility and retail investors take up the market structure. Cryptocurrency prices are highly subjective to the influence of speculative funneling and instant spread of information through the social media as opposed to more traditional equities that are pegged to underlying base of cash flows or dividend yields. Absentee of constant fundamental indicators, previous studies show that the collective market mood tends to dominate the rational valuation schemes and thus it turns into a large scale trigger of excessive price swings [1]. Thus, to achieve a proper prediction of the Bitcoin volatility, the historical price data must be surpassed to properly account for the cumulative impact of these social indicators. The paper will solve this forecasting problem by creating a high-frequency sentiment index on a large dataset of more than 16 million Bitcoin-related tweets between 2014 and 2019 [2].

Although machine learning applications in finance are rapidly developing, the dynamic of relationship between sentiment and volatility has not been thoroughly studied in terms of its structure. There is always confusion on whether the increased predictive power is linear or rather it is more complicated with market-state-dependent behavioral patterns. The given

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365223

paper makes a contribution to the sphere as it proposes a rigorous comparison of a linear OLS benchmark with two specific nonlinear ensemble models, i.e. Random Forest and XGBoost. The study seeks to determine whether social sentiment and historical variance affect the volatility of the market by isolating the differences among such approaches using a symmetric and linear process or through a complex and nonlinear amplification process.

2 Literature reviewed

The studies of the cryptocurrency market dynamics have historically been divided into two general groups, i. e. econometric analysis of the volatility clustering and behavioral analysis of investor sentiment. According to the initial theoretical efforts anchored in the behavioral finance, cryptocurrency markets, with no background in an intrinsic value, are delicate to risks of noise trading. Empirical researches by Mohsin et al. and Bedi and Nashier demonstrate some background that aggregate investor sentiment is meaningful as a price determinant and is usually more effective than the traditional fundamental data in various market environments [1][3]. Expanding the literature to high-frequency social media content, Ebaid and Elshqirat and Mnif et al. affirmed that data of trading volume and sentiment polarity extracted in the social media platforms such as Twitter are leading indicators of intraday volatility spikes [4][5]. This confirms the motivation behind using social signals as the predictive features in the cryptocurrency volatility models.

Regarding the forecasting methodology, recent scholarly literature is moving more toward machine learning architectures than the traditional independently fitting GARCH family modeling. Huang et al. does a detailed benchmarking study and they observed that long short-term memory (LSTM) networks usually outperform heterogeneous autoregressive (HAR) models over an extensive forecasting horizon [6]. On the same note, both Lahmiri and Bekiros, as well as Abakah et al. state that deep learning models are especially suitable to represent the complex chaotic behavior of cryptocurrency returns [7][8]. Nevertheless, there is a common weakness to this area in that there is a trade-off between model predictability and interpretability. Although neural networks tend to achieve better prediction accuracy, due to the black-box nature, it is difficult to understand the structural appearance (linear and nonlinear) of the mechanisms of sentiment-volatility transmission.

Recent evidence also indicates that such a relationship can be state dependent. Investigations by Sila et al. and Zhang et al. show that there is huge asymmetry in Bitcoin volatility: the support of the low sentiment (fear) is imbalanced, and the market has a different response to the positive sentiment (mania) [9][10]. These results imply that linear specifications of the relationship between sentiment and volatility, which are based on consistent values of elasticity, might conceptually constrain the discovery of actual dynamics in cryptocurrency markets. It is these backdrops that this paper is part of the literature to provide a direct comparison of a linear benchmark model (OLS) and a nonlinear ensemble model (random forest). This study, as opposed to Zhang and Li, who investigated the realized volatility by both mixed-frequency data, identifies the incremental nonlinear predictive bias of the daily social sentiment [11]. In this way, it is easier to dissect some kind of behavioral structure of cryptocurrency market volatility.

3 Methodology

The research applies a quantitative paradigm to assess the predictive value of social sentiment on the Bitcoin returns volatility. The empirical approach consists of two phases: 1) A synchronous daily dataset is created through the combination of social media sentiments indicators and Bitcoin market information. Second, predictive accuracy of a linear benchmark model is conducted against nonlinear ensemble learning procedures to determine the presence of a nonlinear dynamical correlation between sentiment and volatility.

3.1 Data collection and preprocessing

The sentiment data are sourced by the Twitter Bitcoin Tweet Dataset where a total of 16 million sentiment polarity [2] tagged tweets is found. In order to measure the daily market sentiment (St), tweet sentiment numbers (positive = +1, neutral = 0, negative = -1) are mapped onto numerical values and daily mean sentiment scores are determined as the result of these values being added together by calendar date.

The market information of Bitcoin (BTC–USD) is accessed through the Yahoo Finance API. Simple percentage returns on a daily basis are the daily returns (Rt):

$$\mathrm{Rt} = \frac{P_t - P_{t-1}}{P_{t-1}}. \quad (1)$$

Pt is the price of day t adjusted close. In line with the tradition in the econometric literature, of consistent-frequency data, the squared returns are used to proxy volatility of the returns:

$$\sigma_t^2 \approx R_t^2. \quad (2)$$

The predictive features are a combination of the sentiment variable and the actual volatility, simply lagged by one trading day to avoid look-ahead bias and fully exploit effects of volatility clustering that have been properly documented. These timestamps are then intersected to form the final analysis data, a continuous time series of observations with feature set $X^t = [S_{t-1}, \sigma_{t-1}^2]$:

$$D = \left\{\left([S_{t-1}, \sigma_{t-1}^2], \sigma_t^2\right)\right\}_{t=1}^{N} \quad (3)$$

3.2 Model specifications

Linear benchmark model (OLS). As a benchmark specification, a standard OLS regression product is being estimated to determine a linear relationship between lagged sentiment and contemporaneous volatility:

$$\sigma_t^2 = \alpha + \beta S_{t-1} + \varepsilon_t \quad (4)$$

where β measures the linear sensitivity of volatility to changes in sentiment, and ε_t denotes an error term assumed to have zero mean.

Nonlinear model (random forest regressor). To capture possible nonlinearities, asymmetries and state-dependent distinctiveness in the relationship between sentiment and volatility, a Random Forest Regressor is used. This is a form of ensemble learning (also known as random forest learning) that integrates the predictions of B separate decision trees with the purpose of minimizing variance and enhancing the performance of generalization to identify which predictions are less varied. This approach (predictor) of the random Forest is:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{5}$$

Where T_b denotes the prediction of the b-th decision tree. Based on preliminary tuning, the Random Forest model is instantiated with 200 estimators and a maximum tree depth of 6 to balance model flexibility with overfitting risk. Because all models are trained on the identical bivariate feature set $(S_{t-1}, \sigma^2_{t-1})$, any differences in predictive performance can be strictly attributed to their underlying architectural structures rather than variations in information content.

XGBoost. The study uses XGBoost as a way of dealing with the possible overfitting problem that is often a characteristic of typical Random Forests when used in noisy financial data. The trees constructed by XGBoost are based on the concept of tree sequence unlike the independent tree generation concept used in the Random Forest, in which the trees constructed sequentially correct the residual errors of the ensemble formed by earlier trees. It uses learning rate (shrinkage) to avoid the dependence of any individual trees. The model is fitted using 100 estimators, a depth of 4 as the maximum and a learning rate of 0.05 so as to provide conservative fitting and sound out of sample generalization.

3.2 Model evaluation

Mean Squared Error (MSE) is mainly used to evaluate model performance. MSE is defined as a mean of the silences of the errors, which is the mean squared discrepancy between the predicted value and actual measured volatility. Within a financial forecasting framework, MSE is a metric by far the most commonly used as it squares the errors to unfairly attribute greater costs to large forecasting variances. It is especially useful in risk management, where a large cost of not forecasting big volatility blasts exists compared to the small cost of missing small changes each day.

A detailed analysis requires additional evaluation indicators like Mean Absolute Error (MAE) and the coefficient of determination (R^2), they are also taken into account. The average absolute deviations between the predicted and actual values are computed in MAE; it provides a linear penalty that renders it more resilient to the numerous extreme outliers that occur in cryptocurrency time series. Meanwhile, measures the fraction of the variation in the observed volatility which can be explained by the characteristics of the model, aide intuitive measure of the overall predictive ability of the model.

To perform empirical analysis, the dataset will be divided in time into a training set (first 80% of sample) and testing set (remaining 20% of sample) in order to make a completely out-of-sample assessment. The time-varying nature of financial data in this chronological division allows one to keep time in proper sequence, thus avoiding the look-ahead bias and information leakage due to future observations.

4 Result

4.1 Predictive performance

The summary of out-of-sample forecasting results of the nonlinear ensemble models (Random Forest and XGBoost) and the linear OLS benchmark is stated in Table 1. The evaluation metrics will be MSE, MAE, as well as the R-squared value, which will be computed on a holdout dataset in the form of the last 20 percent of the time series.

Table 1. Out-of-Sample Performance Comparison

Model Specification	Test MSE(10^{-5})	Test MAE(10^{-3})	R-Squared
Linear Benchmark (OLS)	1.268	1.898	-0.0230
Random Forest (Nonlinear)	1.489	1.998	-0.2016
XGBoost	1.277	1.859	-0.0300

Table 1 demonstrates that the lagged volatility and social sentiment can be used to provide a more subtle view of predictive behavior. Linear benchmark model has the best Mean Squared Error. Since the MSE is more severe in deviations of the extreme forecasting, this means that the linear model is the most conservative and least likely to give wild overpredictions.

XGBoost model on the other hand has the smallest Mean Absolute Error. This finding indicates that XGBoost is highly resistant to the extreme outliers that cryptocurrency markets are characterized by because MAE gives linear penalty, and hence the model represents the median trend of the data more than the other. Random Forest model has the worst out-of-sample performance in all the metrics which imply a high probability of overfitting the training data.

The negative value of the R-squared of all of the specifications indicates the notoriously low signal-to-noise-ratio of financial time series forecasting. Under these conditions, it is very hard to violate a mere average-compatibility benchmark on unobserved data. However, the comparative metrics indicate that although the linear models offer a safer base and rate than two enormous squared slides, the technical forms of gradient boosting tools such as XGBoost has a minor advantage on the natural tracking precision.

4.2 Visualization of fit

Fig.1 illustrates the fit of the three predictive models compared to the observed volatility dynamics during the test period.

On the visual level, the difference in the model behavior is evident. The Random Forest forecast (orange line) has been shown to be violently overreactive to the localized market stress events and results in tremendous spikes that do not happen in the real out of sample data. This aggressive forecasting behavior can be explained by its high penalty under MSE.

On the contrary, Linear Regression (blue dashed) and XGBoost (green) models are both smooth models. XGBoost model is effective in recording the directional changes and clustering of the volatility along with its upper limits when driving noise leads to the sentiment spikes. This ascertains that, although pure sentiment-related extreme events are difficult to identify, nonlinear boosting models execute the structural interdependence between historical variance and sentiment indications in a superior way to decrease the absolute forecasting mistakes.

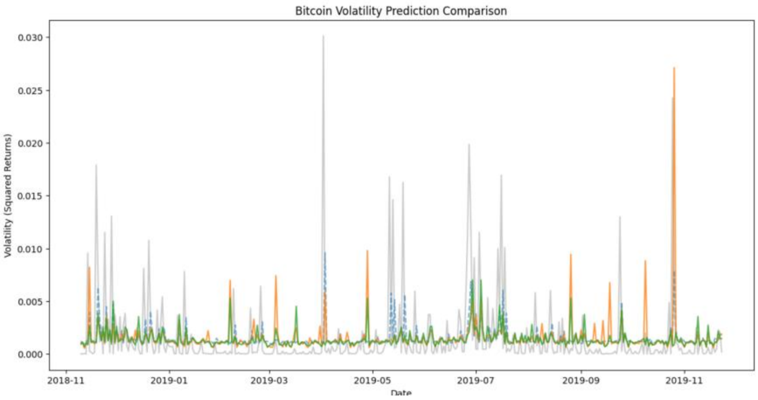

Fig. 1. Bitcoin Volatility Prediction Comparison (Picture credit: Original)

5 Conclusion

This paper uses a rich sample of more than 16 million tweets to examine the structural dependence involving the aggregated social sentiment, history variance and Bitcoin returns volatility. The research directly isolates the predictive dynamics in a high-noise market graphical setting by learning to predict surface values in a linear OLS model and nonlinear ensemble learning structures.

The empirical findings indicate the existence of a subtle truth concerning cryptocurrency market prediction. Random Forest model which does not strongly depend on regularization shows extreme overfitting, becoming highly responsive to noisy sentiment signals and shows the worst tropism to out-of-sample performance measures. On the other hand, the linear OLS benchmark is the best with lowest MSE because a direct, mean conservative linear assumption is operational as a sound defensive protection against colossal over-predictions when markets are at extremities. Nevertheless, XGBoost model produces the lowest MAE, which shows that the median trend followed and the severe outliers present in the cryptocurrency data are the tasks of gradient boosting architectures with the right shrinkage (learning rate).

These results have great theoretical and practical implications. The social sentiment to market volatility mechanism is a complicated set of reactions to capture, even though linear models are a secure place to start, nonlinear models (such as XGBoost) should be regularized carefully to offer a minor benefit in the ultimate accuracy. This affirms that investor sentiment of having actionable predictive information exists, under the condition that the algorithm used to select it can effectively screen out behavioral noise, and it is not overfitting.

These conclusions are supposed to be viewed through certain limitations. To begin with, the squared daily returns serve as an indicator of latent volatility and the measure itself creates a lot of noise. The signal-to-noise ratio can be improved by using the realized volatility measures which are built with high-frequency intraday data (i.e., 5 minutes) in future studies. Second, tree based models have the capability to interact feature interactions, but they do not directly model deterministic temporal deterioration of the variance commonly represented by GARCH types of processes.

To conclude, this paper confirms that the cryptocurrency markets are meaningfully dependent on the behavioral components in terms of which the social media activity can be quantified. To practitioners and risk managers, the integrative approach of properly regularized sentiment-sensitive nonlinear models will offer a solid analysis advantage in the prediction of turbulent markets.

References

[1] Mohsin, M., Naseem, S., Ivașcu, L., Cioca, L. I., Sarfraz, M., Stănică, N. C.: Gauging the effect of investor sentiment on cryptocurrency market: An analysis of Bitcoin currency. Romanian Journal of Economic Forecasting. Vol. 24(4), pp. 87–102 (2021)

[2] Dutta, G.: Bitcoin tweets – Market sentiment. Kaggle Dataset (2021)

[3] Bedi, P., Nashier, T.: The economics of cryptocurrency: A survey. The Review of Austrian Economics. (2024)

[4] Ebaid, A., Elshqirat, K.: Analyzing the interaction between tweet sentiments and price volatility of cryptocurrencies. European Journal of Business and Management Research. Vol. 8(2), pp. 15–19 (2023)

[5] Mnif, E., Jarboui, A., Mouakhar, K.: Cryptocurrency volatility forecasting: The role of social media and search engines. Data. Vol. 10(2), 50 (2020)

[6] Huang, Z., Sangiorgi, I., Urquhart, A.: Forecasting Bitcoin volatility using machine learning techniques. Journal of International Financial Markets, Institutions and Money. Vol. 97, 103265 (2024)

[7] Lahmiri, S., Bekiros, S.: Deep learning in predicting cryptocurrency volatility. Physica A: Statistical Mechanics and its Applications. Vol. 596, 127158 (2022)

[8] Abakah, E. J. A., Gil-Alana, L. A., Madikt, A.: Identification of Bitcoin volatility drivers using statistical and machine learning methods. Applied Soft Computing. Vol. 150, 114384 (2025)

[9] Sila, J., Kocenda, E., Kristoufek, L., Kukacka, J.: Good vs. bad volatility in major cryptocurrencies: The dichotomy and drivers of connectedness. European Journal of Operational Research (2025)

[10] Zhang, C., Ma, H., Arkorful, G. B.: The impacts of futures trading on volatility and volatility asymmetry of Bitcoin returns. International Review of Financial Analysis. Vol. 86, 102497 (2023)

[11] Zhang, Y., Li, X.: Forecasting realized volatility of Bitcoin: The role of social sentiment and trading volume. SSRN Electronic Journal (2023)