Zero-Shot LLM Sentiment Meets Market Microstructure: Parsimonious Feature Selection for High-Frequency Bitcoin Forecastings

Luís A. Moraes[1], Daniel Marques[1], Lucas Nojiri[1], Elisa T. Albergaria[1], Diego R. C. Dias[2(✉)], Vinícius F. S. Mota[2], Marcelo P. Guimarães[3], and Leonardo C. D. Rocha[1]

[1] Universidade Federal de São João del Rei (UFSJ), São João del Rei, Brazil
{luisassismoraes,daniellmarquesfacul,lucas_nojiri}@aluno.ufsj.edu.br, {etuler,lcrocha}@ufsj.edu.br

[2] Universidade Federal do Espírito Santo (UFES), Vitória, Brazil
{diego.dias,vinicius.f.mota}@ufes.br

[3] Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil
marcelo.paiva@unifesp.br

Abstract. High-frequency cryptocurrency forecasting is a difficult task, given the numerous regime changes, microstructure noise, and feature redundancy between technical and social signals. Recent literature has explored large language models (LLMs) for sentiment extraction, but problems such as the increasing number of features leading to greater multicollinearity and overfitting negatively impact economic performance when considering transaction costs. In this paper, we present a multimodal pipeline for Bitcoin trend forecasting, integrating market microstructure, technical and macroeconomic indicators, and contextual Twitter sentiment inferred through zero-shot sampling with Llama-2 7B. To control dimensionality, we applied recursive feature elimination (RFE) with a random forest estimator, selecting a reduced subset of 17 variables from a set of 30 features. We train deep recurrent models (stacked long short-term memory (LSTM) and gated recurrent unit (GRU)) and evaluate them under an equity-based backtesting protocol with long/short execution and 0.1% transaction costs. Results indicate that feature curation matters more than architectural complexity for trading stability: at the 12-hour horizon, the GRU model with RFE achieves a Sharpe Ratio of 5.28 and a Total Return of 48.32%, delivering performance comparable to a Buy & Hold benchmark with lower dispersion. Performance degrades as the forecasting horizon increases to 5 days, suggesting that exogenous signals are more relevant at shorter horizons. Since the evaluation period coincides with a bullish regime in Bitcoin, these results are regime-conditional. Our findings emphasize that parsimonious feature selection is a structural component for stabilizing multimodal forecasting pipelines in high-frequency crypto markets.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 569–583, 2027.
https://doi.org/10.1007/978-3-032-30488-9_35

Keywords: Bitcoin · Multimodal Learning · Zero-Shot Sentiment · Feature Selection · Recursive Feature Elimination

1 Introduction

Predictive modeling of cryptoassets combines continuous trading, retail participation, and abrupt volatility regimes. These factors make predicting high-frequency returns difficult to overcome using classical statistical models and trading strategies in realistic environments. Bitcoin is dynamic in terms of jumps, liquidity, and microstructures. These characteristics increase noise in short-term scenarios, hindering the generalization of the model. [19].

Social signals often appear before market movements, especially on Twitter/X, where mood, attention, and sentiment capture reactions to information before prices change [14,17,20]. The inclusion of large volumes of textual and derived features seems to contribute to the emergence of multicollinearity among some features, which tends to degrade the model's stability in real-world environments. Traditional lexicon-based approaches have limited ability to handle irony, slang, and context shifts common in the crypto ecosystem [17,21]. Recent work employs neural representations and multimodal models to predict extreme events or trends. These approaches still report dependence on feature engineering choices, forecasting horizons, and decision thresholds [23].

Large language models (LLMs) expand the ability to infer contextual sentiment without extensive supervised labeling via zero-shot and instruction prompting [4,13]. Labeling massive social streams remains computationally expensive. Combining these labels with many technical and macroeconomic indicators can worsen the signal-to-noise ratio. We address the following problem: how to integrate contextual social signals extracted by LLMs into high-frequency financial data in a parsimonious way, avoiding multicollinearity and overfitting, while maximizing economic performance under transaction costs?

We propose and evaluate a hybrid architecture that combines market microstructure data, technical indicators, macroeconomic variables, and contextual sentiment. We extract this sentiment via Llama-2 7B in a zero-shot regime, aggregating tweets into discrete 4-hour windows. We adopt the parsimony hypothesis rather than prioritizing feature volume. We apply recursive feature elimination (RFE) guided by a random forest to select an optimal subset of variables, which reduces dimensionality and mitigates multicollinearity and noise [8,9,22]. On this reduced feature set, we train deep recurrent networks (stacked long short-term memory (LSTM) and gated recurrent unit (GRU)) to forecast future log returns. We leverage evidence that GRUs converge on noisy sequences with fewer parameters than LSTMs [5,6,11].

We evaluate the models using an equity-based protocol with long/short execution and 0.1% transaction costs. This approach prioritizes financial metrics (Total Return, Sharpe, and Maximum Drawdown) over statistical errors, following recommendations for economic validation of algorithmic strategies [2,7]. Results indicate that feature selection plays a role in stabilizing trading performance. At the 12-hour horizon, the GRU + RFE model achieved a Sharpe Ratio

of 5.28 and a Total Return of 48.32%. This provides performance similar to the Buy & Hold benchmark over the same period. The evaluation period we used refers to a Bitcoin price increase, making these results conditional on the regime. However, we can see that performance worsens over longer periods, up to 5 days, suggesting that the signals lose relevance as the forecast window increases.

In summary, this work presents the following contributions: a multimodal pipeline aimed at predicting high-frequency Bitcoin trends, integrating microstructure, macroeconomics, and zero-shot contextual sentiment through an LLM; empirical evidence that curation (RFE) mitigates the issue related to model complexity by reducing the number of features; and evaluation performed over larger windows, highlighting the limits of generalization over longer periods.

2 Related Work

Research on cryptocurrency forecasting has advanced on three main fronts: first, using artificial intelligence to deal with the price rollercoaster of this market; second, attempting to combine prices with external information, such as what people say on social media; and finally, seeking more realistic results by creating tests that consider brokerage fees and prevent the model from simply "memorizing" the past instead of learning to predict the future.

Because the cryptocurrency market is volatile and full of unpredictable variations, traditional mathematical models are losing ground to more modern AI architectures, such as LSTM and GRU [5,6,11], which handle this complexity better. Recent studies show that the secret lies not only in choosing the most advanced AI, but also in how the signals are prepared and how the tests are conducted. Authors like Omole and Enke [16] emphasize that it is not enough for the model to be accurate; it needs to be profitable in practice, combining statistical results with real financial returns. Furthermore, new techniques seek to combine different types of data—such as price indicators and market sentiment—using intelligent filters to prevent information overload from hindering more than helping [10].

Several studies indicate that what happens on social media—both the volume of posts and the mood of investors—often provides clues about what will happen to the price, volume, and volatility of Bitcoin shortly afterward [15,20]. Studies show that counting the number of tweets or analyzing the sentiment of the messages helps predict the behavior of the currency in the very short term. In practice, these social signals can be transformed into rules for investment robots, generating above-average profits at certain times. Going beyond Bitcoin, studies with various cryptocurrencies show that Twitter especially influences the larger assets, although this effect is not an absolute rule: it varies greatly depending on the currency and the market situation, which reinforces that public sentiment is a powerful, but often unstable, indicator [14].

A key difference when predicting the market using sentiment is the tool you choose to read what people write. Many people still use simpler and cheaper methods, based on word lists (like VADER [12]), which serve as a common

starting point. However, more modern AI models, such as Transformers, understand context much better. A famous example is FinBERT [1], a version of BERT trained specifically for the world of finance. Recent research shows that these contextual models work better for predicting cryptocurrencies than older methods [23]. Following this trend, the use of large language models (LLMs) has emerged as a way to obtain very high-quality analyses without having to train the AI from scratch, although this is more expensive [4,13]. That's why we made a strategic decision in this article: we use the power of LLMs to extract sentiment accurately, but we control the costs and noise of the data by selecting only the information that really makes a difference to the model.

Adding technical indicators, economic data, and social media sentiment to prices gives the model a more complete view, but it also creates a problem: too much information can generate confusion and bias the AI into patterns that don't exist (the famous curse of dimensionality). To solve this, we use methods like RFE, which act as an intelligent filter, eliminating weak information step by step to keep only what really matters [8,9]. Recent research in the cryptocurrency world shows that cleaning up the data can completely change the game, directly impacting the final profit [10,16]. This goes straight to the central point of our study: instead of believing that "the more data, the better," we tested whether simplicity (or parsimony) is, in fact, the main factor in making the model truly profitable in practice.

A common mistake in many studies is focusing solely on mathematical metrics (such as the model's average error) without proving that it actually generate positive returns net of transaction costs. To avoid this overly optimistic approach, modern evaluations use simulations that include transaction costs and real buying and selling rules [7]. In addition, there is a risk that the model will simply memorize the past due to excessive testing, which creates a false profit on the stocks that will not be repeated in the future. To address this, we use tools such as the Deflated Sharpe Ratio (DSR) [3], which adjusts the results by taking into account how many attempts were made to arrive at that strategy, delivering a much more realistic view than the standard index. Ultimately, how you test the model is just as important as its intelligence. That is why our protocol was rigorous: we divided the data chronologically, repeated the tests multiple times to ensure it was not luck, and simulated a complete operation with real costs.

Against this background, our contribution is to empirically evaluate a *cost-aware* multimodal pipeline for short-horizon Bitcoin forecasting that combines (i) market/macro/technical indicators, (ii) zero-shot LLM-derived sentiment, and (iii) RFE-driven parsimony. Unlike approaches that emphasize ever-larger enrichment, we focus on the practical question of whether selective curation of multimodal features is the primary driver of economically robust performance under realistic backtesting constraints.

3 Methodology

In this section, we detail the proposed and implemented experimental pipeline to evaluate whether the use of multimodal data (market, technical, macroeconomic,

and social signals) and feature selection can improve the performance of Bitcoin market predictions.

Our pipeline is organized into three stages that integrate data collection through economic validation. First, we performed acquisition and temporal alignment of microstructure variables (OHLCV), technical indicators, macroeconomic proxies, and social signals extracted from Twitter. Next, we moved on to feature engineering and selection, where a set of 30 variables was reached via Recursive Feature Elimination (RFE) with a Random Forest estimator to isolate a subset of 17 predictors; finally, we trained and validated two stacked recurrent models (LSTM and GRU).

We constructed a multimodal time series with a fixed granularity of 4-hour intervals, which served as the baseline frequency for both prediction and trading simulation. In this context, we let t represent each 4-hour time step. To compose the central market variables, we collected OHLCV candles of the BTC/USDT pair on the KuCoin exchange via the `ccxt` library, thus defining the Baseline (OHLCV) scenario composed of the set $\{\texttt{BTC_Open}, \texttt{BTC_High}, \texttt{BTC_Low}, \texttt{BTC_Close}, \texttt{BTC_Volume}\}$.

To capture broader risk-on/risk-off dynamics and cross-market comovements, we enrich the dataset with macro proxies obtained via `yfinance` (e.g., DXY, Gold, Oil WTI, Nasdaq). Since these series are not observed continuously at 4-hour intervals, we reindexed them to the study's time grid and applied the forward fill method, carrying over the last available value during weekends and holidays; this procedure allowed us to preserve causal availability, ensuring that no future information was used for interpolation. Additionally, we calculated standard trend and momentum indicators via the pandas-ta library, including moving averages (SMA and EMA in different windows), RSI, MACD components (line, signal, and histogram), realized volatility, and logarithmic returns.

We used a corpus of approximately 1M tweets containing the cashtag "$BTC" from the Bitcoin Tweets Dataset. Because full-scale labeling is computationally expensive, we label a 10% strategic sample using Llama-2 7B Chat in a zero-shot setting. Concretely, we run `meta-llama/Llama-2-7b-chat-hf` quantized to 4-bit (NF4) with `bitsandbytes` on Google Colab (T4 GPU). Each tweet is processed with a fixed instruction prompt to output:

- a sentiment score in $[-1, 1]$, and
- auxiliary intensity signals (market/emotional intensity).

We used a low temperature ($\tau = 0.1$) to reduce output variance and improve reproducibility of the inferred labels.

Tweets arrive asynchronously. For each 4-hour interval t, we aggregate tweet-level outputs into four interval-level features:

$$\texttt{Sentiment_Mean}_t,\ \texttt{Market_Intensity}_t,\ \texttt{Emotional_Intensity}_t,\ \texttt{Tweet_Vol}_t.$$

Intervals with no observed tweets are assigned neutral values (sentiment/intensity set to mid-point) and $\texttt{Tweet_Vol} = 0$, ensuring the time index remains continuous and avoiding leakage due to missingness patterns.

The prediction target is the next-step log return:

$$y_{t+1} = \log(\texttt{BTC_Close}_{t+1}) - \log(\texttt{BTC_Close}_t).$$

We evaluated multiple horizons by forecasting log returns y_{t+h} with $h \in \{3, 7, 14, 30\}$, corresponding to 12 h, 28 h, 56 h, and 5 days at 4-h resolution. This multiple horizons enables assessment of the hypothesis that social signals are short-lived and decay as the horizon increases.

Before feature selection and model training, we quantify linear dependencies among inputs to characterize redundancy and multicollinearity in the enriched feature set. We compute pairwise Pearson correlation coefficients:

$$r(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

and inspect the resulting correlation matrix. This step is used *diagnostically* (not as a feature filter) to motivate RFE and to anticipate potential instability from highly correlated indicators (e.g., overlapping moving-average windows).

To enforce parsimony and reduce redundancy, we applied RFE using a random forest regressor as the external estimator. Random forests are well-suited as a selection backbone because they can capture nonlinear interactions and produce feature importance scores without assuming linearity [8,9].

For regression forests, feature importance is computed as the mean decrease in impurity (MDI), where impurity is typically measured by variance/MSE reduction. For feature X_j, importance is:

$$\mathrm{Imp}(X_j) = \frac{1}{N_{\mathrm{trees}}} \sum_{t=1}^{N_{\mathrm{trees}}} \sum_{\nu \in S_{X_j}(t)} \Delta I(\nu, t),$$

where $S_{X_j}(t)$ denotes the set of split nodes on X_j in tree t and $\Delta I(\nu, t)$ is the impurity decrease at node ν.

RFE iteratively:

1. fits the Random Forest on the training subset;
2. ranks features by importance;
3. removes the least important features; and
4. evaluates validation error (RMSE) to identify the best trade-off between performance and complexity.

We selected k using the elbow/plateau point in the RMSE-versus-#features curve (Fig. 2), yielding a parsimonious set of 17 features from the 30-feature enriched pool. Importantly, this selection is performed using only training/validation data to prevent test contamination.

To isolate the impact of enrichment and parsimony, we evaluate four data scenarios:

1. **Baseline (OHLCV):** only the five market variables.

2. **Macro/Tech Only:** technical indicators and macro proxies added (excluding social features).
3. **Full Enriched:** all 30 variables, including LLM-derived sentiment features.
4. **RFE Optimized:** the 17-variable subset selected by RFE.

For each scenario, we train two deep recurrent architectures: stacked LSTM and stacked GRU. Both are configured to forecast continuous log returns and are trained under identical optimization settings to ensure fairness.

Each model uses:

- **Recurrent layer 1:** 256 hidden units with `return_sequences=True` and tanh activation;
- **Dropout:** 0.2 after the first recurrent layer;
- **Recurrent layer 2:** 128 hidden units with tanh activation;
- **Output:** a linear dense layer producing $\hat{y}_{t+h}$.

We trained the models with Adam ($\eta = 10^{-4}$) and the MSE loss, using early stopping with a patience of 10 epochs. To account for stochasticity in neural training, each configuration is trained 10 times with different random seeds, and we report the mean $\pm$ standard deviation for financial metrics.

All input variables are standardized via StandardScaler (zero mean, unit variance) fit on the training set and applied to validation/test sets, ensuring no leakage from future information [18].

Following equity-based evaluation principles [7], we evaluate models as trading agents over the test set, using a deterministic long/short rule driven by the predicted return:

$$\mathrm{Pos}_{t+1} = \begin{cases} 1 & \text{if } \hat{y}_{t+h} > \epsilon, \\ -1 & \text{if } \hat{y}_{t+h} < -\epsilon, \\ 0 & \text{otherwise,} \end{cases} \qquad \epsilon = 5 \times 10^{-5}.$$

Each position change incurs a transaction cost of 0.1% per trade. This design explicitly tests whether predictive improvements hold up under realistic frictions rather than just improving error metrics.

We report three standard financial metrics [13]:

1. **Total Return (TR):**
$$TR = \prod_{t=1}^{T}(1 + r_t) - 1,$$
where r_t is the strategy return at time t after costs.
2. **Sharpe Ratio (SR):**
$$SR = \frac{\mu(R_p)}{\sigma(R_p)}\sqrt{K},$$
where R_p is the strategy return series and K is the annualization factor for 4-hour intervals.

3. **Maximum Drawdown (MDD):**

$$MDD = \min_t \left(\frac{V_t - \max_{\tau \leq t} V_\tau}{\max_{\tau \leq t} V_\tau} \right),$$

where V_t is the equity curve.

In the next section, we present the results in three stages: (i) exploratory statistics of the social characteristics derived from the LLM and their sparsity/dispersion in 4-hour intervals; (ii) correlation structure and RFE selection dynamics, including the relationship between RMSE and the number of characteristics and the final ranking of importance; and (iii) a comparative financial assessment between horizons and scenarios, highlighting when to employ new characteristics, as well as when to reduce them.

4 Experimental Results

This section reports empirical evidence for the claims raised in the Introduction: (i) multimodal enrichment can help, but naive feature expansion may amplify redundancy and overfitting; (ii) parsimonious feature curation is a first-order driver of *tradable* performance; and (iii) the predictive value of social sentiment is short-lived and degrades as the forecasting horizon increases. Results are presented in three steps aligned with our pipeline (Sect. 3): (1) exploratory analysis of LLM-derived social features, (2) correlation structure and the RFE selection dynamics, and (3) trading-realistic performance under transaction costs across multiple horizons.

4.1 Exploratory Analysis of LLM-Derived Social Features

We first validate whether the LLM-derived social signals exhibit meaningful variability at the 4-hour resolution and whether they reflect the expected *burstiness* of social attention. Table 1 summarizes descriptive statistics for the four aggregated social features over 1,184 time windows.

Table 1. Descriptive Statistics of Social (LLM) Features (4-h windows)

Statistic	Sentiment	Market Int.	Emotional Int.	Tweet Vol.
Mean	0.5142	0.5036	0.5297	86.05
Std. Dev.	0.0277	0.0142	0.0457	191.99
Min	0.0231	0.4176	0.5000	0.00
Median (50%)	0.5000	0.5000	0.5000	0.00
Max	0.5986	0.7000	0.8000	1389.00

Two observations are relevant to our hypotheses. First, `Tweet Vol.` exhibits a strong long-tail distribution: the median is zero while the maximum reaches

1,389 interactions per window, consistent with attention spikes around market-moving events. Second, the sentiment and intensity variables are relatively concentrated around neutral values (median ≈ 0.5), as expected when aggregating heterogeneous social content into coarse time bins. This combination of (i) sparse attention bursts and (ii) weakly varying sentiment levels motivates two design choices discussed earlier: (a) using models capable of learning nonlinear interactions among sparse exogenous signals and market dynamics, and (b) controlling redundancy and noise via parsimonious feature selection rather than unquestioningly expanding the feature set.

4.2 Correlation Structure and RFE Selection Dynamics

Redundancy and Multicollinearity. Before evaluating trading performance, we inspect the correlation structure of the enriched dataset. Figure 1 shows the Pearson correlation heatmap. Consistent with our motivation, many technical indicators (especially overlapping moving-average windows and related momentum measures) exhibit strong mutual correlation, whereas their correlations with the return target are generally weak.

Selecting a Parsimonious Subset (RFE). To test feature reduction, we applied RFE with a Random Forest estimator and tracked validation RMSE with respect to the number of features retained. Figure 2 illustrates the relationship between complexity and validation error. We observed a stagnation point, beyond which adding features produces marginal (or negative) gains, indicating that additional variables tend to generate noise and redundancy rather than a useful signal. Based on the results of the experiments, we selected $k = 17$ features.

Which Variables Survive RFE? Figure 3 presents the ranking in order of relevance of each variable in the selected subset. The resulting set highlights the variables that describe the market microstructure (variables related to price/volume) and by momentum/volatility descriptors, while macroeconomic variables play a secondary role. Sentiment and intensity variables derived from the LLM are excluded by the RFE, while tweet activity (volume) remains. This indicates that, within the evaluated configuration, simple social attention variables carry a more measurable signal than the contextual polarity measures extracted by the LLM.

4.3 Trading-Realistic Performance Across Horizons

We now evaluate whether the predictive models translate into *tradable* performance under realistic frictions. Tables 2, 3, 4 and 5 report the equity-based backtest results for four horizons: 12 h (3 periods), 28 h (7 periods), 56 h (14 periods), and 5 days (30 periods). Each neural configuration is trained 10 times with different seeds; for Total Return, we report the mean $\pm$ standard deviation.

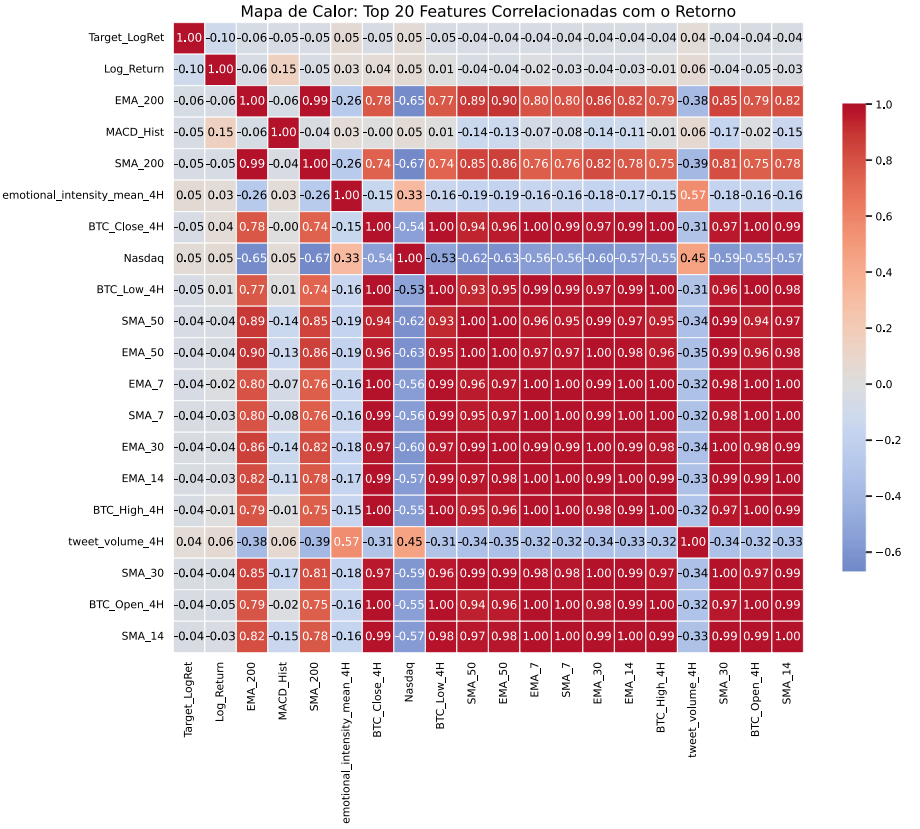

Fig. 1. Pearson correlation heatmap. Strong inter-feature correlations (particularly among technical indicators) indicate multicollinearity.

Table 2. Financial Results – Horizon of 3 periods (12 h)

Model	Data Scenario	Total Return (%)	Sharpe Ratio	MDD (%)
Benchmark	Buy & Hold	47.93%	5.27	−10.83%
LSTM	Baseline (OHLCV)	27.24% ± 14.45	3.34	−14.56%
LSTM	Macro/Tech Only	20.42% ± 34.30	2.05	−20.05%
LSTM	Full Enriched	8.65% ± 31.91	0.81	−22.29%
LSTM	Feature Selection (RFE)	35.61% ± 9.83	4.17	−12.71%
GRU	Baseline (OHLCV)	25.80% ± 28.54	2.88	−15.56%
GRU	Macro/Tech Only	25.03% ± 30.48	2.68	−16.85%
GRU	Full Enriched	22.75% ± 35.96	2.29	−19.93%
GRU	**Feature Selection (RFE)**	**48.32% ± 10.81**	**5.28**	**-11.21%**

Table 3. Financial Results – Horizon of 7 periods (28 h)

Model	Data Scenario	Total Return (%)	Sharpe Ratio	MDD (%)
Benchmark	Buy & Hold	47.93%	5.27	−10.83%
LSTM	Baseline (OHLCV)	−1.28% ± 15.54	0.05	−22.35%
LSTM	Macro/Tech Only	41.21% ± 20.69	4.56	−13.04%
LSTM	Full Enriched	16.38% ± 28.68	1.85	−19.30%
LSTM	Feature Selection (RFE)	47.36% ± 3.54	5.23	−10.91%
GRU	Baseline (OHLCV)	4.96% ± 25.49	0.60	−22.62%
GRU	Macro/Tech Only	28.03% ± 29.94	3.06	−17.47%
GRU	Full Enriched	24.37% ± 29.31	2.70	−15.94%
GRU	Feature Selection (RFE)	47.09% ± 18.84	5.12	−11.91%

Table 4. Financial Results – Horizon of 14 periods (56 h)

Model	Data Scenario	Total Return (%)	Sharpe Ratio	MDD (%)
Benchmark	**Buy & Hold**	**47.93%**	**5.27**	**-10.83%**
LSTM	Baseline (OHLCV)	−7.08% ± 25.81	−1.01	−25.90%
LSTM	Macro/Tech Only	24.98% ± 34.12	2.58	−18.59%
LSTM	Full Enriched	28.19% ± 33.67	2.94	−16.45%
LSTM	Feature Selection (RFE)	40.66% ± 21.20	4.44	−12.97%
GRU	Baseline (OHLCV)	−2.32% ± 14.92	−0.08	−24.55%
GRU	Macro/Tech Only	18.53% ± 25.88	2.15	−18.74%
GRU	Full Enriched	29.44% ± 30.45	3.22	−16.09%
GRU	Feature Selection (RFE)	35.40% ± 20.52	3.98	−13.80%

Table 5. Financial Results – Horizon of 30 periods (5 days)

Model	Data Scenario	Total Return (%)	Sharpe Ratio	MDD (%)
Benchmark	**Buy & Hold**	**47.93%**	**5.27**	**-10.83%**
LSTM	Baseline (OHLCV)	−19.15% ± 13.38	−2.49	−30.13%
LSTM	Macro/Tech Only	7.06% ± 34.02	0.56	−24.95%
LSTM	Full Enriched	5.22% ± 33.70	0.40	−24.77%
LSTM	Feature Selection (RFE)	32.30% ± 21.72	3.68	−15.01%
GRU	Baseline (OHLCV)	−16.28% ± 10.47	−1.98	−30.41%
GRU	Macro/Tech Only	19.53% ± 31.26	2.14	−20.48%
GRU	Full Enriched	3.92% ± 21.14	0.60	−23.65%
GRU	Feature Selection (RFE)	18.65% ± 30.90	2.05	−19.55%

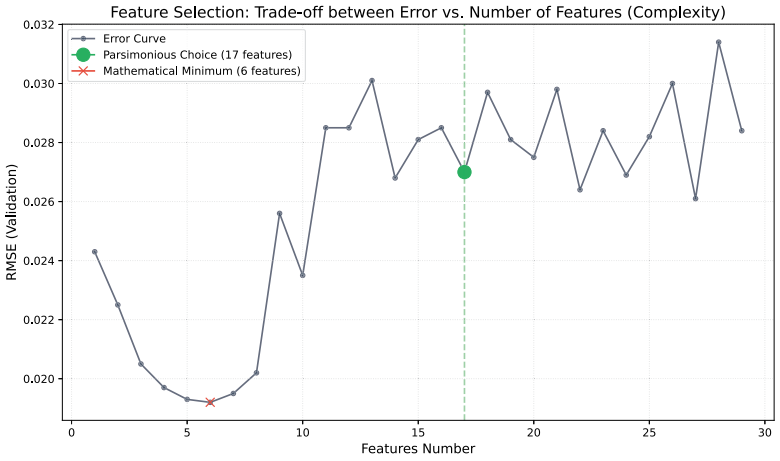

Fig. 2. RFE optimization curve. The chosen operating point ($k = 17$) balances predictive performance and parsimony.

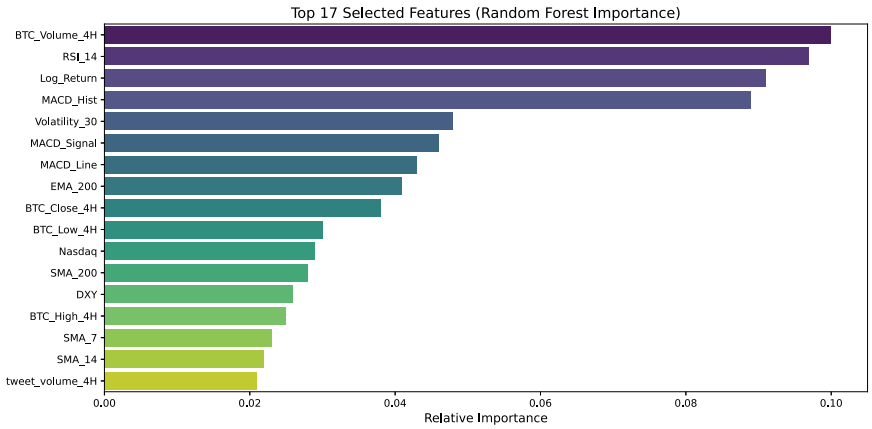

Fig. 3. Top-17 features selected by RFE.

4.4 Integrated Discussion and Key Takeaways

At different time intervals, the RFE-optimized scenario shows improvements compared to the scenario with extracted variables, demonstrating that a simpler dataset outperforms a more complex one. The result that allows us to affirm this can be seen in the window presented in Table 3: the GRU drops from 48.32% (RFE) to 22.75% (with all variables), despite using the same architecture and training protocol. This pattern indicates that naive multimodal expansion introduces redundancy and noise (consistent with the correlation analysis in Fig. 3), leading to unstable adjustments and weaker realized returns after applying costs.

Therefore, selecting a compact subset results in more stable performance (lower variability).

Over a 12-hour horizon, the GRU + RFE strategy achieved the best overall performance, reaching a Sharpe ratio of 5.28 and a total return of 48.32%, with an MDD of -11.21%, matching the Buy & Hold benchmark under transaction costs. This is the only horizon in which the proposed strategy remains competitive with the passive strategy. Within the evaluated period (characterized by a strong bullish regime), this suggests that carefully crafted short-term setups can remain economically viable when microstructure effects and short-term attention dynamics are exploited under controlled dimensionality.

As we move to longer timeframes, such as 56 h or 5 days (Tables 4 and 5), the performance of AI-based strategies begins to decline, falling below the simple Buy & Hold strategy. Even our best combination (LSTM + RFE) failed to outperform the market on the 5-day horizon. This confirms that factors such as social media behavior and details of market microstructure have a fleeting effect, being quickly absorbed by the price. Since the selection filter (RFE) ended up discarding sentiment variables from the LLM, it is clear that the model's short-term advantage comes much more from analyzing market dynamics and attention volume than from the content of news or external opinions.

Although GRU tends to be more competitive at very short horizons (e.g., 12 h) and LSTM can match performance at intermediate horizons under RFE (e.g., 28 h), these differences are smaller than the gains produced by the feature selection itself. This reinforces the article's main contribution: filtering noise and redundancy in the inputs is more determinant of performance than the sophistication of the algorithm or the volume of external variables.

5 Conclusion and Future Work

In this work, we explore whether a parsimonious multimodal pipeline is capable of delivering viable performance in high-frequency cryptoasset markets, especially when considering transaction costs. As discussed in Sects. 1 and 2, while the use of multimodal variables increases the model's representational capacity, it also introduces redundancy and multicollinearity issues. Without proper control, these factors amplify overfitting and destroy real profitability as soon as operational fees are applied. To address this challenge, we combine zero-shot sentiment inference (via Llama-2 7B) with a Random Forest-based RFE selection strategy. Finally, we evaluate the performance of stacked GRU and LSTM predictors under a realistic backtesting protocol, with long/short execution and explicit cost recognition.

Our results demonstrate that the model's effectiveness lies more in the rigorous curation of input data than in the sophistication of its architecture, highlighting that an excess of variables introduces noise that hinders generalization. By selecting only the 17 most relevant predictors via RFE, the GRU model achieved superior performance, reaching a Sharpe Ratio of 5.28 and returns of 48.32% over 12-hour horizons, remaining competitive against the Buy & Hold

strategy. However, the degradation of performance over longer periods reveals that predictability is highly sensitive to sudden market changes, limiting the strategy's competitive advantage to short-term operations, where the signal still outweighs the structural drift of the data.

The RFE procedure did not retain contextual LLM-derived sentiment or polarity variables. It selected primarily endogenous market descriptors and tweet volume. Attention intensity proxies carried more measurable signal than contextual semantic polarity within the evaluated configuration. The contribution of large-scale language-derived sentiment was limited in this setup. We cannot attribute the observed short-horizon performance to contextual LLM enrichment.

This study suggests that profitability in high-frequency crypto modeling under realistic frictions depends less on complex architectures or multimodal expansion. It relies more on parsimonious signal selection, regime awareness, and cost-aware validation.

Acknowledgment. The authors acknowledge partial funding from the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001; the National Council for Scientific and Technological Development - Brazil (CNPq) - Funding Codes 442344/2023-6, 444602/2024-0, and 407466/2025-8; the Espírito Santo Research and Innovation Support Foundation (FAPES) – Call No. 02/2026 (Participation in Technical-Scientific Events) – Grant No. 428/2026 and project #2024-9G1WP BPC; and the Minas Gerais Research Funding Foundation (FAPEMIG)

References

1. Araci, D.: FinBERT: financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 (2019). https://arxiv.org/abs/1908.10063
2. Bailey, D.H., Borwein, J.M., López de Prado, M., Zhu, Q.J.: The probability of backtest overfitting. J. Comput. Finance **20**(4), 39–69 (2017). https://doi.org/10.21314/JCF.2016.322
3. Bailey, D.H., López de Prado, M.: The deflated sharpe ratio: correcting for selection bias, backtest overfitting and non-normality. J. Portfolio Manag. **40**(5), 94–107 (2014). https://doi.org/10.2139/ssrn.2460551
4. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
5. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734 (2014)
6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling arXiv preprint arXiv:1412.3555 (2014)
7. De Prado, M.L.: Advances in Financial Machine Learning. John Wiley & Sons (2018)
8. Genuer, R., Poggi, J.M., Tuleau-Malot, C.: Variable selection using random forests. Patt. Recogn. Lett. **31**(14), 2225–2236 (2010)
9. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)

10. Han, P., Chen, H., Rasool, A., Jiang, Q., Yang, M.: MFB: a generalized multimodal fusion approach for Bitcoin price prediction using time-lagged sentiment and indicator features. Expert Syst. Appl. **261**, 125515 (2025). https://doi.org/10.1016/j.eswa.2024.125515
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. Hutto, C., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 8, pp. 216–225 (2014). https://doi.org/10.1609/icwsm.v8i1.14550
13. Jung, H.S., Lee, H.: Explainable zero-shot trading using multi-agent LLM architecture: a backtested approach for Bitcoin price. Inf. Process. Manag. **63**(2), 104466 (2026)
14. Kraaijeveld, O., De Smedt, J.: The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. J. Int. Finan. Markets. Inst. Money **65**, 101188 (2020). https://doi.org/10.1016/j.intfin.2020.101188
15. Mai, F., Shan, Z., Bai, Q., Wang, X., Chiang, R.H.L.: How does social media impact bitcoin value? a test of the silent majority hypothesis. J. Manag. Inf. Syst. **35**(1), 19–52 (2018). https://doi.org/10.1080/07421222.2018.1440774
16. Omole, O., Enke, D.: Deep learning for Bitcoin price direction prediction: models and trading strategies empirically compared. Finan. Innov. **10**(1), 117 (2024). https://doi.org/10.1186/s40854-024-00643-1
17. Pano, T., Kashef, R.: A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. Big Data Cogn. Comput. **4**(4), 33 (2020)
18. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
19. Scaillet, O., Treccani, A., Trevisan, C.: High-frequency jump analysis of the Bitcoin market. J. Financ. Economet. **18**(2), 209–232 (2020). https://doi.org/10.1093/jjfinec/nby013
20. Shen, D., Urquhart, A., Wang, P.: Does Twitter predict Bitcoin? Econ. Lett. **174**, 118–122 (2019). https://doi.org/10.1016/j.econlet.2018.11.007
21. Suardi, S., Rasel, A.R., Liu, B.: On the predictive power of tweet sentiments and attention on bitcoin. Int. Rev. Econom. Finance **79**, 289–301 (2022). https://doi.org/10.1016/j.iref.2022.02.017
22. Youssefi, A., Hessane, A., Zeroual, I., Farhaoui, Y.: Optimizing forecast accuracy in cryptocurrency markets: evaluating feature selection techniques for technical indicators. Comput. Mater. Continua **83**(2), 3411 (2025). https://doi.org/10.32604/cmc.2025.063218
23. Zou, Y., Herremans, D.: PreBit–a multimodal model with Twitter FinBERT embeddings for extreme price movement prediction of Bitcoin. Expert Syst. Appl. **233**, 120838 (2023). https://doi.org/10.1016/j.eswa.2023.120838